



بايثون والذكاء الاصطناعي للتشخيص الطبي

منهاج تدريبي مكثف لعشرة أيام موجه للكوادر الصحية
من البيانات المخبرية إلى نماذج التنبؤ التشخيصي — من دون برمجة تقليدية

بيانات مخبرية واقعية

صفر حلقات — معالجة موجبة 100%

90 دقيقة يومياً

10 أيام

د. حسام صلاح مهدي
رئيس قسم الذكاء الاصطناعي — جامعة المصطفى
Dr. Husam Salah Mahdi — Head of AI Department, Al-Mustafa University

تمهيد

يشهد قطاع الرعاية الصحية تحولاً تاريخياً: من ممارسة سريرية تقودها الخبرة الفردية والحدس، إلى رعاية تسترشد بالبيانات والأدلة الكمية. فالسجلات الصحية الإلكترونية، وأنظمة معلومات المختبرات، ومنصات RT-PCR، وأجهزة مراقبة المرضى، تولّد جميعها تياراً متواصلاً من المعلومات — غير أن معظم هذا الكنز يبقى غير مستثمر، لأن أقرب المهنيين إلى البيانات لم يُمنحوا يوماً الأدوات اللازمة لاستنطاقها. صُمّم هذا المنهاج لكوادر مديرية التشخيص في وزارة الصحة لسيّ هذه الفجوة تحديداً. في عشر جلسات مدة كل منها تسعون دقيقة، ينتقل الكادر الطبي والمختبري من أول سطر بايثون إلى بناء نماذج تعلّم الآلة التشخيصية وتقييمها نقدياً — بالمقاييس ذاتها التي يثقون بها أصلاً: الحساسية، والنوعية، والقيم التنبؤية، ومنحنيات ROC، والمساحة تحت المنحنى AUC. يلتزم المنهاج مبدأً تربوياً لا مساومة فيه: لا برمجة تقليدية بجمل التحكم. فلا حلقات for، ولا حلقات while، ولا سلاسل if/elif/else في هذه الدورة إطلاقاً. كل عملية على سجلات المرضى — تصفية، ووسماً، وتجميعاً، وتقييماً — تُنفَّذ بالأدوات الموجهة في مكتبي Numpy وPandas: الأقنعة المنطقية، وnp.where، وmap، وcut، وgroupby. وليس هذا تبسيطاً للمادة، بل هو المعيار المهني لعلم البيانات الحديث، وهو ما يتيح للكادر السريري تكريس جهده الذهني لتفسير البيانات بدلاً من إنشاء الشيفرات. كل مجموعة بيانات في هذا الكتاب تحاكي الواقع اليومي للمختبر التشخيصي: صورة الدم الكاملة، ولوحات الكيمياء السريرية، وقيم عتبة الدورة Ct لفحوصات RT-PCR، وسجلات الأمراض المزمنة. وكل تمرين ينتهي حيث يبدأ عمل المديرية — بقرار.

د. حسام صلاح مهدي

مدير الدورة — مديرية التشخيص، وزارة الصحة

المحتويات

i

تمهيد

1	مخطط الدورة وفلسفتها	0
1	لماذا بايثون لمديرية التشخيص؟	0.1
2	الفلسفة التربوية: لا حلقات تكرارية ولا جمل شرطية	0.2
2	خارطة طريق الأيام العشرة	0.3
3	تشرح الجلسة ذات التسعين دقيقة	0.4
4	بيئة التعلم وإدارة الاعتماديات	0.5
4	مجموعات بيانات الدورة — قاموس البيانات	0.6
5	كيف تقرأ هذا الكتاب	0.7
6	اليوم 1 — الأسس والبيانات الصحية	1
7	السياق السريري والإطار النظري (20 دقيقة)	1.1
10	التطبيق الموجّه (40 دقيقة)	1.2
14	التحدي السريري (30 دقيقة)	1.3
16	الواجب المنزلي والتقييم الذاتي	1.4
18	اليوم 2 — استيراد السجلات وأول تحليل	2
19	السياق السريري والإطار النظري (20 دقيقة)	2.1
19	لماذا يجب أن ينتهي عصر جداول البيانات	2.1.1
20	تشرح إطار البيانات	2.1.2
20	أنواع البيانات بيانات وصفية تشخيصية	2.1.3
21	بروتوكول النظرة الأولى	2.1.4
22	التطبيق الموجّه (40 دقيقة)	2.2
27	التحدي السريري (30 دقيقة)	2.3
30	الواجب المنزلي والتقييم الذاتي	2.4
32	اليوم 3 — تنظيف البيانات الطبية	3
33	السياق السريري والإطار النظري (20 دقيقة)	3.1
36	التطبيق الموجّه (40 دقيقة)	3.2
42	التحدي السريري (30 دقيقة)	3.3
44	الواجب المنزلي والتقييم الذاتي	3.4
46	اليوم 4 — الفلتر المنطقية وتحليلات العمليات	4
47	السياق السريري والإطار النظري (20 دقيقة)	4.1
50	التطبيق الموجّه (40 دقيقة)	4.2
55	التحدي السريري (30 دقيقة)	4.3
57	الواجب المنزلي والتقييم الذاتي	4.4

59	اليوم 5 — التصوير البصري للبيانات	5
60	5.1 السياق السريري والإطار النظري (20 دقيقة)	
63	5.2 التطبيق الموجّه (40 دقيقة)	
68	5.3 التحدي السريري (30 دقيقة)	
70	5.4 الواجب المنزلي والتقييم الذاتي	
72	اليوم 6 — هندسة الميزات	6
73	6.1 السياق السريري والإطار النظري (20 دقيقة)	
75	6.2 التطبيق الموجّه (40 دقيقة)	
80	6.3 التحدي السريري (30 دقيقة)	
83	6.4 الواجب المنزلي والتقييم الذاتي	
85	اليوم 7 — أسس تعلّم الآلة	7
86	7.1 السياق السريري والإطار النظري (20 دقيقة)	
89	7.2 التطبيق الموجّه (40 دقيقة)	
93	7.3 التحدي السريري (30 دقيقة)	
95	7.4 الواجب المنزلي والتقييم الذاتي	
98	اليوم 8 — نماذج التصنيف التشخيصية	8
99	8.1 السياق السريري والإطار النظري (20 دقيقة)	
102	8.2 التطبيق الموجّه (40 دقيقة)	
106	8.3 التحدي السريري (30 دقيقة)	
109	8.4 الواجب المنزلي والتقييم الذاتي	
111	اليوم 9 — تقييم النماذج التشخيصية	9
112	9.1 السياق السريري والإطار النظري (20 دقيقة)	
115	9.2 التطبيق الموجّه (40 دقيقة)	
122	9.3 التحدي السريري (30 دقيقة)	
125	9.4 الواجب المنزلي والتقييم الذاتي	
127	اليوم 10 — مسارات العمل والأخلاقيات ومشروع التخرج	10
128	10.1 السياق السريري والإطار النظري (20 دقيقة)	
132	10.2 التطبيق الموجّه (40 دقيقة)	
135	10.3 التحدي السريري (30 دقيقة)	
140	10.4 الواجب المنزلي والتقييم الذاتي	
140	10.5 التخرج: عشرة أيام وعشر مهارات	
142	مرجع سريع لمكتبي NumPy وPandas	A
142	A.1 تحميل البيانات وفحصها	
143	A.2 الاختيار والفلتر	
143	A.3 التنظيف	
144	A.4 التجميع والدمج	
144	A.5 هندسة الميزات	
145	مرجع سريع لمكتبة scikit-learn	B
146	B.1 سكربت مصغّر متكامل من البداية إلى النهاية	
148	دليل المقاييس التشخيصية	C
148	C.1 التعاريف والصيغ	
149	C.2 مثال محلول بالكامل	
149	C.3 المجالات المرجعية المختبرية للبالغين المستخدمة في هذه الدورة	

151	D	مسرد المصطلحات
154	E	إعداد البيئة ومجموعات البيانات
154	E.1	البداية السريعة مع Colab Google (الخيار الموصى به)
155	E.2	التثبيت دون اتصال (محطات عمل المديرية)
155	E.3	إعادة توليد مجموعات البيانات

د. حسام صلاح مهدي
 رئيس قسم الذكاء الاصطناعي — جامعة المصطفى
 Dr. Husam Salah Mahdi — Head of AI Department, Al-Mustafa University

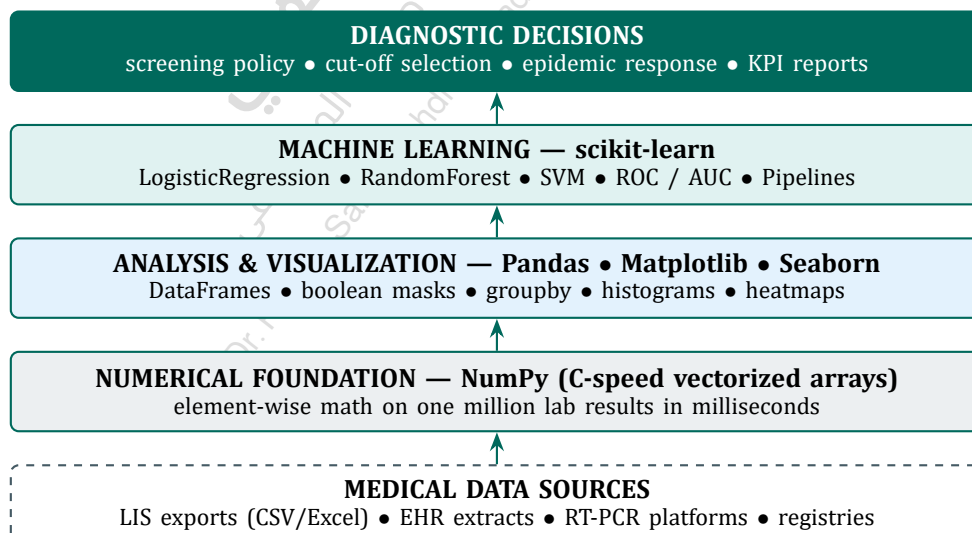
باب 0

مخطط الدورة وفلسفتها

0.1 لماذا بايثون لمديرية التشخيص؟

يتحول قطاع الصحة العالمي تحولاً حاسماً من الممارسة القائمة على الحدس إلى الرعاية الموجهة بالبيانات. فالسجلات الصحية الإلكترونية (EHR)، وأنظمة معلومات المختبرات (LIS)، ومستودعات التجارب السريرية، والمنصات الجزيئية مثل RT-PCR تولّد أحجاماً من البيانات لا يمكن لأي جدول بيانات تقليدي أن يستوعبها واقعياً. وتُعدّ بايثون اليوم اللغة المهيمنة عالمياً في علم البيانات والبحث الطبي لأربعة أسباب عملية:

- تكلفة صفرية ومصدر مفتوح. لا أعباء ترخيص على المؤسسات العامة — فكل أداة في هذه الدورة مجانية.
- منظومة برمجية بمستوى طبي. إلى جانب الحزمة العالمية (NumPy, Pandas, Matplotlib, scikit-learn)، توفر بايثون مكتبات صحية متخصصة مثل PyHealth وMedPy وlifelines (تحليل البقاء) وpydicom (التصوير الطبي).
- صياغة مقروءة. تُقرأ بايثون بصيغة قريبة من اللغة الإنكليزية الطبيعية — وهي مثالية للأطباء وعلماء المختبرات وأخصائيي الوبائيات ممن هم خبراء في الطب لا في هندسة البرمجيات.
- قابلية التوسع. المهارات نفسها تمتد من ملف إكسل واحد مُصدّر إلى تحليلات صحة سكانية على المستوى الوطني.



شكل 0.1: حزمة التحليلات الصحية بايثون التي تُدرّس في هذه الدورة: كل طبقة تُبنى مباشرة على الطبقة التي تحتها، وكل طبقة تنتهي بقرار تشخيصي.

0.2 الفلسفة التربوية: لا حلقات تكرارية ولا جمل شرطية

إن أكثر أنماط الفشل شيوعاً عند تعليم البرمجة للممارسين السريين هو إنفاق ساعات ثمينة على بُنى صُمِّمت أساساً لمهندسي البرمجيات: الحلقات التكرارية `for` و `while` وسلاسل الجمل الشرطية `if/elif/else`. فالتكرار سجلاً بعد سجل عبر آلاف سجلات المرضى بطيء بالنسبة إلى الحاسوب، والأهم من ذلك أنه بطيء بالنسبة إلى المتعلم، لأنه يجبره على التفكير في الآليات البرمجية بدلاً من الطب. لذلك يستبعد هذا المنهج التحكم الانسيابي الكلاسيكي استبعاداً كاملاً، ويستبدل به أدوات المعالجة الموجهة (Vectorization) في مكتبي NumPy و Pandas، التي تعمل دواخلها بشفرة C مُجمَّعة:

المفهوم الكلاسيكي (المستبعد)	البديل الموجه	التطبيق الطبي
حلقات <code>while / for</code> على السجلات	العمليات الحسابية الموجهة على الأعمدة	حساب مؤشر كتلة الجسم لكل مريض بتعبير واحد: <code>df["wt"] / df["ht"]**2</code>
فحص السجلات بجمل <code>if</code> <code>condition:</code>	الأقنعة المنطقية (Boolean Masks) <code>(df[df["glucose"] > 126])</code>	عزل كل مريض يستوفي معياراً تشخيصياً في لحظة واحدة
الحلقات المتداخلة لمقارنة المجموعات	التجميع بدالة <code>groupby()</code>	متوسط مدة الرقود حسب القسم والجنس والفئة العمرية في سطر واحد
سلاسل التصنيف <code>if/elif/else</code>	<code>np.where()</code> و <code>map()</code> و <code>pd.cut()</code>	تصنيف السكر التراكمي HbA1c إلى طبيعي / ما قبل السكري / سكري عبر السجل الكامل دفعة واحدة

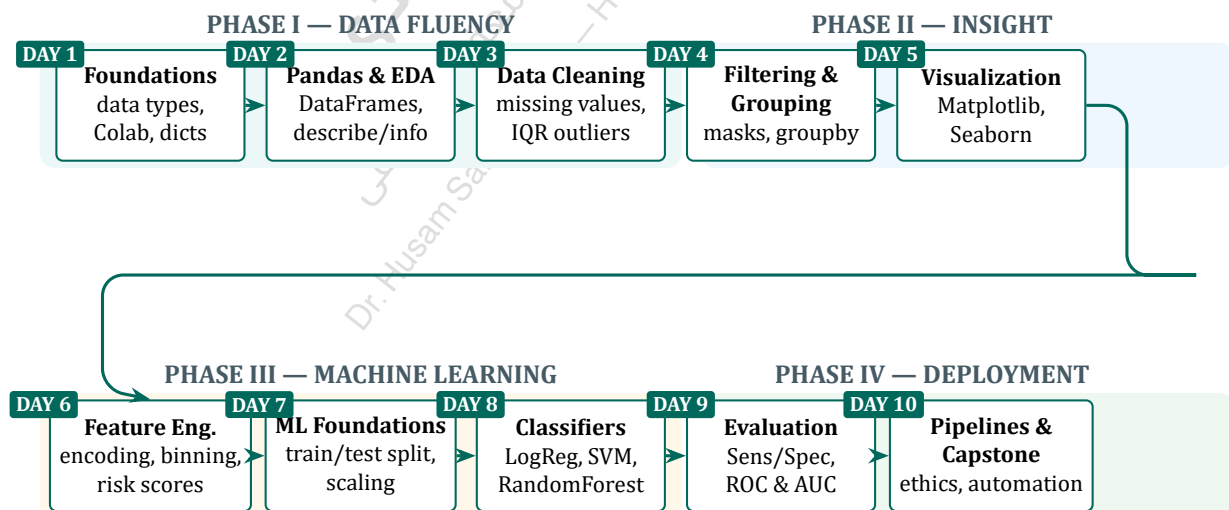
جدول 0.1: الاستبدال الاستراتيجي الكامن في قلب هذا المنهج.

القيمة لمديرية التشخيص — العبء المعرفي

كل ساعة لا تُنفق على آليات الحلقات هي ساعة تُنفق على المجالات المرجعية وعتبات القرار التشخيصية وتفسير النتائج. وهكذا يبقى النموذج الذهني للمتدرب سرياً من الدقيقة الأولى إلى الدقيقة الأخيرة.

0.3 خارطة طريق الأيام العشرة

الدورة قوس واحد متصل: من ملفات التصدير المختبرية الخام إلى نموذج ذكاء اصطناعي تشخيصي خاضع لتقييم نقدي.



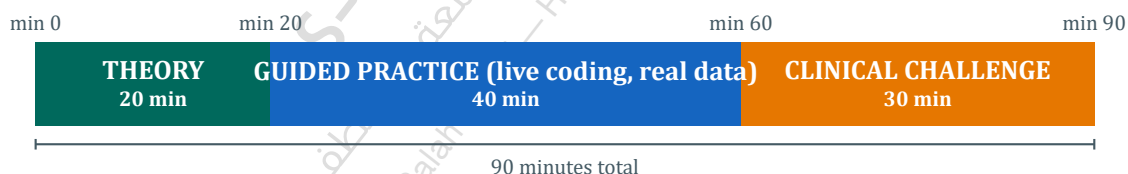
شكل 0.2: قوس التعلم ذو الأيام العشرة: كل يوم يستثمر مهارات اليوم السابق ويغذي اليوم التالي.

اليوم	العنوان	التطبيق التشخيصي الجوهرى
1	الأسس التقنية وهياكل البيانات الصحية	سجل المريض بوصفه قاموساً في بايثون؛ و Colab مختبراً للتدريب
2	استيراد السجلات الطبية والتحليل الاستكشافي الأول	ملف تعريف إحصائي فوري لسجل مرضى السكري بـ describe()
3	تنظيف البيانات الطبية	القيم المفقودة (Missing Values) في صورة الدم الكاملة (CBC)؛ علامات حيوية مستحيلة؛ فرز القيم الشاذة (Outliers) بالمدى الربيعي (IQR) في اللوحات المختبرية
4	الفترة المنطقية المتقدمة وتحليلات العمليات السريرية	مؤشرات الأداء الرئيسية (KPI) لقسم الطوارئ؛ تحليل إعادة الإدخال ومدة الرقود
5	عرض البيانات المرئي وقصة العلامات الحيوية	مدرجات تكرارية للمجالات المرجعية؛ توزيعات قيمة عتبة الدورة (Ct) لفحوصات RT-PCR؛ خرائط حرارية للارتباط
6	هندسة الميزات (Feature Engineering) للطب التنبؤي	أعمدة مؤشر كتلة الجسم، والترميز (Encoding) الفئوي، والتقسيم الطبقي (Binning) للأعمار، ودرجة الخطورة القلبية
7	الأسس الخوارزمية لتعلم الآلة	train_test_split والمعايرة القياسية (StandardScaler) وخطر فرط التخصيص (Overfitting)
8	بناء نماذج التصنيف التشخيصية	احتمالات الانحدار اللوجستي (Logistic Regression)؛ الغابة العشوائية (Random Forest)؛ دراسة حالة آلة المتجهات الداعمة (SVM)
9	التقييم النقدي للأداء التشخيصي	مصفوفة الارتباك (Confusion Matrix)، والحساسية (Sensitivity)، والنوعية (Specificity)، و PPV/NPV، ومنحنى خاصية تشغيل المستقبل (ROC) والمساحة تحت المنحنى (AUC)
10	مسارات العمل والأتمتة والأخلاقيات والمشروع الختامي	مسار عمل (Pipeline) تشخيصي قابل لإعادة الاستخدام؛ التحيز والخصوصية والمشروع الختامي

جدول 0.2: جدول الدورة في لمحة واحدة.

0.4 تشريح الجلسة ذات التسعين دقيقة

تتبع كل جلسة الإيقاع نفسه المسند بالأدلة، بحيث يصبح الشكل ذاته غير مرئي بحلول اليوم الثالث وينصرف الانتباه كله إلى المحتوى:



شكل 0.3: إيقاع الجلسة الثابت 20-40-30 المعتمد في الأيام العشرة كلها.

- النظرية (20 دقيقة). المفهوم الطبي والإحصائي، مُرسَّخاً في سيناريو تشخيصي قبل ظهور أي شيفرة.
- التطبيق الموجه (40 دقيقة). برمجة حية يقودها المدرب على بيانات مختبرية واقعية (اصطناعية وخاضعة لإخفاء الهوية Anonymization)؛ وينفذ المتدربون كل خلية في دفاترهم الخاصة.
- التحدي السريري (30 دقيقة). مسألة مفتوحة تُحل في أزواج، يعقبها نقاش للنتيجة بوصفها معطى تشخيصياً لا مجرد مخرجات شيفرة.

0.5 بيئة التعلم وإدارة الاعتماديات

➤ حقبة أدوات الذكاء الاصطناعي والاعتماديات — الحزمة البرمجية المعتمدة في الدورة

تُنفَّذ الدورة بالكامل في **Google Colab** (<https://colab.research.google.com>) — وهي بيئة Jupyter مجانية تعمل عبر المتصفح مع كل مكتبة مطلوبة مثبتة مسبقاً. لا حاجة إلى تثبيت أي شيء على حواسيب الوزارة، ولا تلزم صلاحيات مسؤول النظام، وبيئة كل متدرب مطابقة تماماً لبيئات زملائه.

وللتثبيت دون اتصال بالإنترنت (كما في محطات عمل المديرية غير المتصلة بالشبكة)، ثَبَّت **Anaconda** أو استخدم `pip`:

```
pip install numpy pandas matplotlib seaborn scikit-learn openpyxl
```

النسخ المثبتة المستخدمة في إعداد هذه الدورة (يتوفر ملف `requirements.txt` في الملحق E):

<code>numpy ≥ 2.0</code>	<code>pandas ≥ 2.2</code>	<code>matplotlib ≥ 3.9</code>	<code>seaborn ≥ 0.13</code>
<code>scikit-learn ≥ 1.5</code>	<code>scipy ≥ 1.13</code>	<code>openpyxl ≥ 3.1</code>	<code>jupyter</code>

0.6 مجموعات بيانات الدورة — قاموس البيانات

جميع مجموعات البيانات اصطناعية وخاضعة لإخفاء الهوية بالكامل، وقد وُلِدَت ببذرة عشوائية ثابتة بحيث يرى كل متدرب الأرقام نفسها. وهي مُنمَّدة إحصائياً على توزيعات واقعية لمجتمع مستشفى عراقي، ومقدمة بصيغة ملفات CSV داخل المجلد `datasets/`.

المحتوى	الصفوف	الملف
,admission_date,department,governorate,sex,age,patient_id readmitted_30d,smoker,length_of_stay_days	1,200	patients_demographics.csv
,neutrophils_pct,plt,hct,hgb,rbc,wbc,sample_date: صورة الدم الكاملة: (patient_id lymphocytes_pct (مفهرسة بالحقل	1,200	cbc_results.csv
,alt_ul,urea_mgdl,creatinine_mgdl,hba1c_pct,glucose_mgdl ,ldl_mgdl,hdl_mgdl,triglycerides_mgdl,total_chol_mgdl,ast_ul potassium_mmol,sodium_mmol	1,200	chemistry_panel.csv
,ct_value,target_gene,collection_date,patient_id,sample_id lab_name,(Positive / Negative / Inconclusive) result	800	rtPCR_results.csv
ملف تصدير مختبري مُفسد عمداً: قيم مفقودة، وقيم شاذة مستحيلة، وسجلات مكررة، وترميز غير متسق — وهو ميدان التدريب لليوم الثالث	600	messy_lab_data.csv
,blood_pressure,glucose,pregnancies: Pima سجل سكري على نمط outcome,age,diabetes_pedigree,bmi,insulin,skin_thickness	768	diabetes_dataset.csv
,cholesterol,resting_bp,chest_pain_type,sex,age: سجل الخطورة القلبية: target,st_depression,exercise_angina,max_hr,fasting_bs	900	heart_risk.csv

جدول 0.3: مجموعات بيانات الدورة السبع. وتظهر المجالات المرجعية المعتمدة في الكتاب كله في الملحق C.

✖ خطأ شائع — حوكمة البيانات

رغم أن كل مجموعة بيانات في هذه الدورة اصطناعية، يجب على المتدربين التعامل معها بالانضباط نفسه المتبع مع السجلات الحقيقية: لا محاولات لإعادة تحديد الهوية، ولا رفع لأي ملفات تصدير حقيقية لبيانات المرضى إلى خدمات سحابية عامة، مع الالتزام الصارم بسياسة حماية البيانات في الوزارة وبالمعايير الدولية (HIPAA, GDPR) التي تُناقش في اليوم العاشر.

0.7 كيف تقرأ هذا الكتاب

يستخدم كل فصل يومي لغة بصرية متسقة:

✓ الأهداف التعليمية

ما ستكون قادراً على فعله — مصاعاً في هيئة نواتج قابلة للقياس في بداية كل يوم.

📖 المفهوم الأساسي

الفكرة الإحصائية أو الحاسوبية لليوم، مشروحة بلغة سريرية.

★ لؤلؤة سريرية — السياق

أين يلامس هذا المفهوم الممارسة التشخيصية الفعلية — المجالات المرجعية وعتبات القرار والإرشادات السريرية.

■ عرض تطبيقي مباشر — التطبيق الموجّه

مادة البرمجة الحية لكتلة التطبيق ذات الأربعين دقيقة، مع المخرجات المتوقعة معروضة في لوحات داكنة.

🏆 التحدي السريري — التحدي

التحدي السريري ذو الثلاثين دقيقة — يُجرى دائماً على مجموعة بيانات من الدورة، وينتهي دائماً بقرار.

✓ حل المدرب — ملاحظات المدرب

حلول كاملة مفصلة ونقاط نقاش موجهة إلى المدرب.

✖ خطأ شائع — الأخطاء الشائعة

مزالق رُصدت في أفواج تدريب حقيقية — اقرأها قبل الشروع في المختبر.

♣ القيمة لمديرية التشخيص — القيمة المؤسسية

العائد الملموس لمديرية التشخيص — كيف تتحول كل مهارة إلى تقارير أسرع، أو ترصد وبائي أفضل، أو قرارات تشخيصية أكثر أماناً.

باب 1

اليوم 1 — الأسس التقنية وهياكل البيانات الصحية

كل وردية عمل في مستشفى عراقي تُنتج سيلاً من المعلومات التشخيصية: صور الدم الكاملة (CBC) من محلل أمراض الدم، ولوحات الكيمياء السريرية من خط الكيمياء الحيوية، ونتائج RT-PCR من مختبر الأحياء الجزيئية، والبيانات الديموغرافية المدونة في سجلات الإدخال. واليوم ما تزال معظم هذه المعلومات تنتقل على الورق أو في جداول إلكترونية معزولة، ما يعني أنها لا تُقرأ إلا من شخص واحد في كل مرة ولا يكاد يحللها أحد. الغاية من هذه الدورة هي تغيير ذلك: على مدى عشرة أيام ستتعلم تحميل البيانات التشخيصية وتنظيفها وتحليلها وعرضها بيانياً ثم نمذجتها أخيراً بلغة بايثون (Python)، وصولاً إلى أدوات تعلم آلي تدعم الحكم السريري — ولا تحل محله أبداً.

يضع اليوم الأول حجر الأساس. ستتعرف على بيئة العمل (Google Colab، دفتر ملاحظات مجاني يعمل من المتصفح)، وستفهم ما المكتبة البرمجية ولماذا لا نثبت شيئاً بأيدينا تقريباً، وستتعلم كيف تمثل بايثون الحقائق السريرية التي تتعامل معها يومياً: عمر، وكراتينين، وحالة تدخين، وسجل مريض كامل — وصولاً في الختام إلى ملف مريض رقمي متكامل مبني من المتغيرات والقوائم والقواميس وحدها.

الدقائق	الجزء	المحتوى والأنشطة
00-05	الترحيب وخارطة طريق الدورة	لماذا تستثمر المديرية في مهارات البيانات؛ مسار الأيام العشرة من الجداول الإلكترونية إلى الذكاء الاصطناعي
05-10	التحول المستند إلى البيانات	من السجلات الورقية إلى سجلات قابلة للاستعلام؛ حلقة التغذية الراجعة التشخيصية
10-15	لماذا بايثون	المصدر المفتوح، والصياغة المقروءة، والمنظومة العلمية، والمجتمع العالمي للبيانات الصحية
15-20	جولة في Colab والاعتماديات	الدفاتر والخلايا وبيئات التشغيل والمشاركة؛ المكتبات و <code>pip</code> و <code>requirements.txt</code>
20-30	التطبيق الموجّه الأول	الخلايا الأولى: المتغيرات والأنواع؛ سطر التقرير المختبري بسلسلة <code>f</code>
30-45	التطبيق الموجّه الثاني	قائمة قراءات الغلوكوز: <code>len</code> و <code>min</code> و <code>max</code> و <code>sum</code> و <code>sorted</code> ؛ آخر حلقة تكرارية سترها
45-60	التطبيق الموجّه الثالث	قاموس المريض؛ سجل CBC المتداخل؛ الردهة كقائمة قواميس؛ التشويق نحو إطار البيانات (DataFrame)
60-85	التحدي السريري	بناء ملف مريض رقمي كامل؛ مؤشر كتلة الجسم؛ ملخص سريري متعدد الأسطر
85-90	الاستخلاص والواجب المنزلي	عرض الحل خطوة بخطوة، واختبار التحقق الذاتي، ولمحة عن اليوم الثاني

✓ الأهداف التعليمية

- بنهاية هذه الجلسة ستكون قادراً على:
1. أن تشرح، في جملتين تصليحان لاجتماع في المديرية، لماذا بايثون مفتوحة المصدر هي الأداة القياسية لتحليل البيانات الصحية وما الذي يقدمه Google Colab مجاناً.
 2. أن تنشئ خلايا شيفرة في دفتر Colab وتشغلها وتشاركها.
 3. أن تطابق الأنواع الأولية الأربعة (`int` و `float` و `str` و `bool`) مع عناصر البيانات السريرية وتختار النوع الصحيح لحقل معين.

4. أن تخزنَ حقائق المريض في متغيرات حسنة التسمية وتجمعها في سطر تقرير بسلسلة نصية منسقة (f-string) مع ضبط المنازل العشرية.
5. أن تلخص سلسلة قراءات مختبرية محفوظة في قائمة باستخدام len وmin وmax وsum وsorted — دون كتابة أي حلقة.
6. أن تمثل مريضاً واحداً بقاموس، ولوحة مختبرية بقاموس متداخل، مع قراءة القيم وتحديثها بالوصول عبر الأقواس المربعة.
7. أن تجمع ردهة كاملة كقائمة من القواميس وتبين لماذا هذا الهيكل هو السلف المباشر لإطار البيانات الذي نلقاه غداً.

➔ حقبة أدوات الذكاء الاصطناعي والاعتماديات — مكتبات اليوم

اليوم الأول يوم بايثون خالصة عن قصد: كل شيء فيه ما عدا اللمحتين الختاميتين يستخدم اللغة نفسها فقط، دون أي استيراد على الإطلاق. أما المكتبتان العلميتان أدناه فتظهران لوهلة قصيرة في النهاية ثم تصبحان رفيقتين يوميتين من اليوم الثاني فصاعداً.

```
1 import numpy as np      # numerical arrays; vectorized statistics
2 import pandas as pd     # DataFrames (tables); previewed today only
```

شيفرة 1.1: الاستيرادات المستخدمة اليوم

على جهاز شخصي تُثبَّت هاتان المكتبتان مرة واحدة من سطر الأوامر: أما على Google Colab فهما حاضرتان مسبقاً، ولذلك يُعرض الأمر أدناه فقط لكي تتعرف عليه عندما تصادفه في التوثيق:

```
pip install numpy pandas
```

توفر NumPy مصفوفات عددية سريعة وإحصاءات بسطر واحد مثل np.mean: فمتوسط ضبط جودة عبر 100,000 نتيجة محلل يكلف سطرًا واحدًا وجزءاً من الثانية. وتوفر pandas إطار البيانات، البنية الجدولية التي ستحمل كل مجموعة بيانات في هذه الدورة — الجسر بين ملفات Excel التي تنتجها مختبراتك أصلاً والأدوات الإحصائية وأدوات الذكاء الاصطناعي القادمة. ويوزع أي مشروع قائمة مكتباته الدقيقة في ملف صغير اسمه requirements.txt، بحيث يستطيع أي زميل أو خادم وزاري إعادة إنشاء البيئة نفسها بأمر واحد.

1.1 السياق السريري والإطار النظري (20 دقيقة)

من السجلات الورقية إلى حلقة تغذية راجعة تشخيصية

تأمل كيف تعيش قيمة الهيموغلوبين اليوم في كثير من مستشفياتنا: يطبعها المحلل، وتُلتصق في سجل، وتُنسخ يدوياً إلى ملف المريض، وربما يُعاد إدخالها في ملخص Excel شهري. كل نسخة تفتح الباب لخطأ نقل، ولا يمكن الاستعلام عن أي منها: فلا أحد يستطيع أن يعرف، في ثوان، كم مريضاً مصاباً بفقر الدم استقبلت الطوارئ في الشهر الماضي، أو ما إذا كان المحلل الثاني ينحرف نحو قيم منخفضة أيام الأحد. القيمة على الورق تجيب عن سؤال واحد لطبيب واحد؛ والقيمة نفسها في سجل مهيكل تجيب عن آلاف الأسئلة للمديرة كلها، وإلى الأبد.

حلقة التشخيص المستندة إلى البيانات

تدير الخدمات التشخيصية الحديثة حلقة متصلة: نقيس (المحللات ومنصات PCR) ← نسجل (سجلات رقمية مهيكلية، لا نصاً حراً) ← نحلل (إحصاءات ولوحات متابعة ونماذج) ← نقرر (رايات تنبيه وضبط جودة وتوزيع موارد) ← ثم عودة إلى القياس. الورق يكسر الحلقة عند الخطوة الثانية. وهذه الدورة تقوّي الخطوتين الثانية والثالثة حتى تستند الخطوة الرابعة إلى الدليل لا إلى الذاكرة — وقوة الحلقة من قوة هياكل بياناتها، ولهذا كان يومنا هذا عن كيفية تمثيل الحقائق السريرية أصلاً.

لماذا بايثون للبيانات الطبية

تيمم بايثون على علم البيانات الصحية لثلاثة أسباب. فهي مفتوحة المصدر: مجانية على كل حاسوب وزاري، بلا تراخيص وبلا ارتبان لموزد. وهي مقروءة: فالتعبير `mean_glucose = sum(readings) / len(readings)` يمكن قراءته بصوت عالٍ لزميل لم يبرمج قط — وهو أمر جوهري حيث يجب أن تكون الشيفرة قابلة للتدقيق سرريباً. ولها أعماق منظومة علمية: NumPy وpandas للبيانات، وMatplotlib للأشكال، وscikit-learn للتعلم الآلي، وكلاً منها يصونه مجتمع عالمي من المستشفيات والجامعات ووكالات

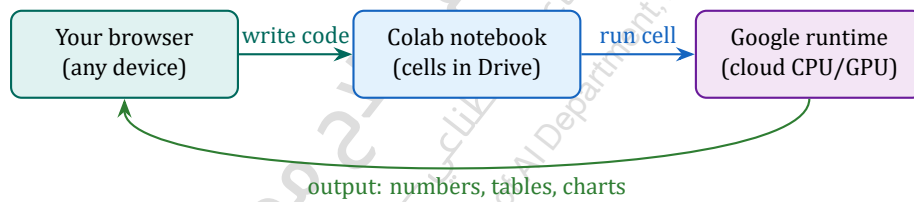
الصحة العامة. وعندما تتشارك منظمة الصحة العالمية والاتحادات الجينومية ومعظم أوراق الذكاء الاصطناعي الطبي المنشورة الشيفرة، فإنها تتشارك بايثون.

الدفاتر وبيئات التشغيل والخلايا

Google Colab بيئة دفاتر ملاحظات تعمل في المتصفح. الدفتر سلسلة من الخلايا؛ خلية الشيفرة تحوي بايثون وتعرض مخرجاتها تحتها مباشرة، وخلية النص تحوي الشرح — فالدفتر إذن هو التحليل وتقريره في آن واحد، وهو مثالي لمسار تدقيق مختبري. لا تعمل الشيفرة على حاسوبك؛ فالضغط على زر التشغيل (Run) يرسل الخلية إلى بيئة تشغيل، وهي آلة افتراضية مؤقتة في أحد مراكز بيانات Google، تنفذها وتعيد النتائج إلى متصفحك. استوعب نتيجتين لهذا: حتى الحاسوب المختبري المتواضع يستطيع تحليل مجموعات بيانات كبيرة، لأن السحابة هي التي تحسب؛ وبيئة التشغيل مؤقتة — فعندما يعاد تدويرها تختفي المتغيرات والملفات المرفوعة، بينما يبقى الدفتر نفسه آمناً في Google Drive. وتُشارك الدفاتر كما تُشارك المستندات، بصلاحيات عرض أو تحرير — وبهذه الطريقة ستسَلِّم واجباتك في هذه الدورة.

المكتبات والاستيراد وأداة pip

المكتبة (أو الحزمة) شيفرة كتبها آخرون واختبروها ويصونونها، تُلحق بجلستك بعبارة `import` — أي أنها استعارة محلل مُصادق عليه بدلاً من بناء محلل من الصفر. العبارة `import pandas as pd` تحمل `pandas` تحت الاسم المختصر العالمي `pd`. وتُجلب المكتبات الجديدة من فهرس حزم بايثون بأداة `pip`؛ غير أن بيئات تشغيل Colab تصل والمنظومة العلمية كلها مثبتة مسبقاً فيها، ولهذا لا تحتاج هذه الدورة إلى جلسة تثبيت. أما المشاريع الاحترافية فتجَمِّد اعتمادياتها في `requirements.txt` حتى تبقى النتائج قابلة لإعادة الإنتاج بعد سنوات — وهي الفكرة نفسها في تدوين أرقام وجبات الكواشف لعدة فحص مُصادق عليها.



شكل 1.1: سير العمل في Colab: الشيفرة المكتوبة في المتصفح تعمل على بيئة تشغيل مؤقتة لدى Google؛ ولا يعود إلى جهازك سوى النتائج. لا تثبيت، ولا حاجة إلى قدرة حوسبة محلية، ويبقى الدفتر محفوظاً بأمان في Drive.

أنواع بايثون هي أنواع البيانات السريية

تصنّف لغات البرمجة القيم إلى أنواع، والأنواع الأولية الأربعة في بايثون تقابل واحداً لواحد أصناف الحقائق الموجودة على أي استمارة طلب مختبري. واختيار النوع الصحيح ليس تزمناً لغوياً؛ فالإحصاءات لا تُحسب إلا على الأنواع العددية، وعمر المريض المخزن سهواً كنص يستم بصمت كل متوسط وكل نموذج يُبنى عليه.

جدول 1.1: مطابقة أنواع بايثون الأولية مع البيانات السريية

نوع بايثون	صنف القيمة	أمثلة سريية	الصيغة في الشيفرة
int	عدد صحيح	العمر، مدة الرقود، عدد الصفحات	58
float	عدد عشري	الكرياتينين، HbA1c، قيمة TSH	1.4
str	نص	معرف المريض، التشخيص، المحافظة	"MOH-00042"
bool	راية نعم/لا	مدخن، أُعيد إدخاله، PCR إيجابي	True

أنواع بايثون بوصفها أنواع بيانات سريية

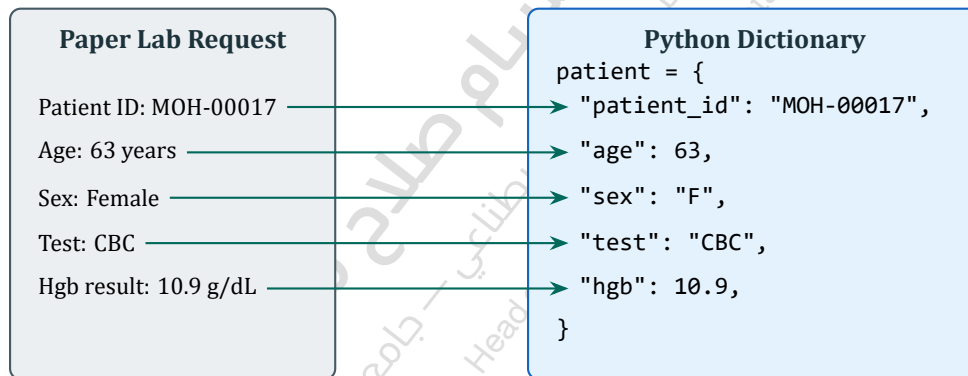
النوع `int` يعدّ الأشياء غير القابلة للتجزئة: سنوات العمر وأيام الرقود. والنوع `float` يحمل دقة القياس: كرياتينين 1.4 ملغ/دل. والنوع `str` (السلسلة النصية) نص بين علامتي اقتباس؛ فالقيمة "58" حرفان لا عدد، ولا يمكن حساب متوسط عمود منها. والنوع

bool هو بالضبط True أو False، وهو الصيغة الطبيعية لكل راية سريرية؛ والمقارنات تنتج قيمة منطقية مباشرة ($hgb < 13$)، وهذه الآلية سترفع، من اليوم الثالث فصاعداً، رايات آلاف النتائج غير الطبيعية بسطر موجّه واحد. أما المتغير فوعاء مسّعى لقيمة واحدة؛ وينبغي أن يبيّن الاسم ما هي القيمة، وأن يتضمن وحدتها إن كانت قياساً: فالاسم creatinine_mgdl يوتّق نفسه بنفسه، أما x فهو تقرير حادثة مؤجل.

القوائم والقواميس: أشكال السجلات السريرية

القوائم والقواميس بوصفها سجلات سريرية

القائمة (list) سلسلة مرتبة من القيم بين قوسين مربعين: [148, 162, 155]. الترتيب فيها مهم، تماماً كما في قراءات الجلوكوز المتسلسلة — فالقراءة الثالثة هي الثالثة بمعناها الكامل. والدوال الجاهزة (built-ins) تستهلك القائمة كلها دفعة واحدة: len (كم عددها)، max و min (الطرفان)، sum (المجموع)، و sorted (نسخة مرتبة). والقاموس (dict) مجموعة أزواج مفتاح-قيمة بين قوسين معقوسين: المفتاح اسم الحقل، والقيمة محتواه — وهذا بالضبط صف واحد من السجل الصحي الإلكتروني (EHR)، وهو نفسه الشكل (JSON) الذي تتبادل به أجهزة المختبر ووحدات البيانات الصحية الوطنية السجلات. والقواميس تتداخل: فسجل المريض يمكن أن يضم مفتاح "cbc" قيمته نفسها قاموس تحليل. وقائمة القواميس ردهة كاملة: سجلات كثيرة تتقاسم المفاتيح نفسها. احتفظ بهذه الصورة — فغداً تصبح إطار البيانات.



شكل 1.2: طلب مختبري ورقي يُطابق سطرًا بسطر مع قاموس بايثون: كل تسمية حقل تصبح مفتاحاً، وكل قيد مكتوب بخط اليد يصبح قيمة منقطة — وهذه هي بالضبط الطريقة التي تتبادل بها السجلات الصحية الإلكترونية ووحدات المختبرات البيانات (JSON).

★ لؤلؤة سريرية — القيم المرجعية المعتمدة في هذه الدورة

احفظ هذه الحدود الفاصلة للبالغين عن ظهر قلب؛ فكل مجموعات بيانات الدورة تستخدمها. صورة الدم الكاملة: كريات الدم البيض $4-11 \times 10^9/L$ ؛ الهيموغلوبين 13-17 غ/دل (للرجال) و12-15 غ/دل (للنساء)؛ الصفائح $150-450 \times 10^9/L$. سكر الدم: جلوكوز البلازما الصائم البالغ 126 ملغ/دل فأكثر في مناسبتين تشخيصيّ للسكري (100-125 خلل في جلوكوز الصيام)؛ والسكر التراكمي HbA1c البالغ 6.5% فأكثر سكري، و6.4-5.7 ما قبل السكري. RT-PCR: قيمة عتبة الدورة (Ct) هي دورة التضخيم التي تعبر عندها إشارة الهدف حد الكشف؛ وقيمة Ct المنخفضة تعني حملاً فيروسياً مرتفعاً. نطاقات الدورة: Ct < 25 حمل فيروسي مرتفع؛ 25-30 متوسط؛ 30-35 منخفض؛ 35-40 منخفض جداً قرب حد الكشف — ففكر بإعادة الفحص. لاحظ المنطق العكسي: كلما صغرت قيمة Ct زاد الفيروس. وكل النطاقات تعتمد على العدة التشخيصية؛ فتحقق منها لمنصتك أنت. وملفنا rtPCR_results.csv يرمّز النتائج السلبية بقيمة Ct فارغة مع "Negative".result =

يُختتم الجزء النظري بالانضباط المركزي للدورة كلها: كل تحليل هو عملية على مجموعة كاملة دفعة واحدة — قائمة اليوم، وعمود إطار بيانات غداً. والتطبيق الموجّه الآن يجعل هذا ملموساً.

1.2 التطبيق الموجّه (40 دقيقة)

■ عرض تطبيقي مباشر — من الخلية الأولى إلى ردهة من السجلات

الهدف: خلال 40 دقيقة يشغل كل متدرب خلاياه الأولى في Colab، ويخزن مريضاً واحداً في متغيرات منمطة، وينسق سطر تقرير بسلسلة f، ويلخص سلسلة غلوكوز بدوال القوائم الجاهزة، ويبنى قاموس مريض متداخلاً، ويجمع ردهة من ثلاثة مرضى كقائمة قواميس، ويرى — من دون شرح بعد — أن سطرًا واحداً من pandas يحول تلك الردهة إلى جدول. يكتب المتدربون كل سطر بأنفسهم؛ فذاكرة الأصابع للصياغة جزء من الهدف.

افتح الموقع colab.research.google.com، وسجل الدخول بحساب Google الخاص بالدورة، واختر دفتر ملاحظات جديد (New notebook). أعد تسميته `day01_yourname` من حقل العنوان في الأعلى. أمامك الآن خلية شيفرة فارغة؛ والخطوات التالية تملأ الدفتر.

الخطوة 1: أول تماس — شغل خلية

اكتب السطرين أدناه في الخلية الأولى واضغط `Shift+Enter`، أكثر ضغطة مفاتيح استخداماً في هذه الدورة: فهي تشغل الخلية وتنتقل إلى ما بعدها. السطر الأول يطبع رسالة؛ والثاني يُظهر بايثون كآلة حاسبة شفافة (120 دقيقة من التثفيل معبراً عنها بالساعات).

```
1 print("Hello from the Diagnostics Directorate!")
2 print(120 / 60)
```

شيفرة 1.2: أول سطرين في الدورة

```
Hello from the Diagnostics Directorate!
2.0
```

لاحظ أن القسمة تعطي دائماً قيمة float (ولهذا ظهرت 2.0 لا 2) — فحساب القياسات يجب ألا يتخلى عن الدقة بصمت أبداً.

الخطوة 2: متغيرات المريض وأنواعها

نخزن الآن أربع حقائيق عن مريض واحد، واحدة لكل نوع أولي. أسماء المتغيرات تتبع عرف الدورة: أحرف صغيرة، وكلمات تفصلها شرطات سفلية، والوحدة مضمّنة في الاسم كلما كانت القيمة قياساً. والدالة `type` تكشف كيف صُنفت بايثون كل قيمة.

```
1 patient_id = "MOH-00042"      # str - text identifier
2 age = 58                      # int - whole number
3 creatinine_mgdl = 1.4         # float - decimal measurement
4 smoker = True                 # bool - yes/no clinical flag
5
6 print(type(patient_id), type(age))
7 print(type(creatinine_mgdl), type(smoker))
```

شيفرة 1.3: أربع حقائيق سريرية، أربعة أنواع

```
<class 'str'> <class 'int'>
<class 'float'> <class 'bool'>
```

✖ خطأ شائع — أرقام متنكرة في هيئة نص

أكثر أخطاء البيانات شيوعاً في الجداول الصحية عدد مخزن كنص: فالتعبير `age = "58"` يبدو مطابقاً على الشاشة، لكن `"58"` + 1 = 59، وعمود كامل من هذه القيم لا متوسط له. اليوم الثالث يواجه هذا الخطأ على نطاق واسع في `messy_lab_data.csv`؛ ومنذ اليوم، الانضباط هو: علامات الاقتباس للمعرفات والنصوص، وأبداً ليست للقياسات.

الخطوة 3: سلسلة f تبني سطر التقرير

السلسلة المنسقة (f-string) سلسلة نصية تسيقها f وتضمن أقواسها المعقوفة قيمة حية؛ ومحدّات التنسيق بعد النقطتين تضبط العرض (f: 1 : منزلة عشرية واحدة، وf: 0. : بلا منازل). كل سطر تقرير مطبوع في هذه الدورة يُبنى بهذه الطريقة. ولاحظ كيف تُقسّم السلسلة الطويلة على أسطر داخل قوسين حتى لا يطول أي سطر فيتعذر قراءته.

```
1 name = "Ali Hassan"
2 glucose_mgdl = 182.0
3 hba1c_pct = 8.14
4
5 report_line = (
6     f"{patient_id} | {name} | {age} y | "
7     f"glucose {glucose_mgdl:.0f} mg/dL | HbA1c {hba1c_pct:.1f} %"
8 )
9 print(report_line)
```

شيفرة 1.4: سطر تقرير مختبري منسق

```
MOH-00042 | Ali Hassan | 58 y | glucose 182 mg/dL | HbA1c 8.1 %
```

توقف عند المضمون السريري: غلوكوز 182 ملغ/دل مع HbA1c 8.1% يصفان سكرياً راسخاً سيئ الضبط ($\text{HbA1c} \geq 6.5\%$) — فالسطر المنسق يُقرأ فعلاً وكأنه ورقة إجابة.

الخطوة 4: قائمة قراءات الغلوكوز ودوالها الجاهزة

سبع قيم غلوكوز صائم من أسبوع مراقبة داخل الردهة تدخل في قائمة. كل ملخص يأتي من دالة جاهزة تستهلك القائمة بكاملها دفعة واحدة؛ والمتوسط هو ببساطة المجموع مقسوماً على العدد.

```
1 glucose_readings = [148, 162, 155, 171, 149, 180, 158]
2
3 print("Number of readings:", len(glucose_readings))
4 print("Lowest :", min(glucose_readings), "mg/dL")
5 print("Highest:", max(glucose_readings), "mg/dL")
6 print("Sorted :", sorted(glucose_readings))
7 mean_glucose = sum(glucose_readings) / len(glucose_readings)
8 print(f"Mean : {mean_glucose:.1f} mg/dL")
```

شيفرة 1.5: تلخيص سلسلة قراءات دون أي حلقة

```
Number of readings: 7
Lowest : 148 mg/dL
Highest: 180 mg/dL
Sorted : [148, 149, 155, 158, 162, 171, 180]
Mean : 160.4 mg/dL
```

كل القراءات تتجاوز عتبة الصيام البالغة 126 ملغ/دل: الدوال الجاهزة أجابت عن السؤال السريري — فرط سكر دم مستمر — في خمسة أسطر.

✖ خطأ شائع — آخر حلقة تكرارية سترها هنا

معظم دورات البرمجة كانت ستعلّمك الآن مراكمة المتوسط عنصراً عنصراً. نحن نعرض هذا النمط مرة واحدة بالضبط، فقط لكي نتعرف عليه في الشيفرات القديمة — ثم نحظره. قارن:

```
1 # ANTI-PATTERN -- shown once, never used again in this course
2 total = 0
```

```

3 for g in glucose_readings:
4     total = total + g
5 mean_loop = total / len(glucose_readings)
6
7 # Course standard: one whole-collection expression, no loop
8 mean_builtin = sum(glucose_readings) / len(glucose_readings)
9
10 import numpy as np
11 mean_numpy = np.mean(glucose_readings)
12 print(mean_loop, mean_builtin, mean_numpy)

```

شيفرة 1.6: النمط المرفوض مقابل معيار الدورة

```
160.42857142857142 160.42857142857142 160.42857142857142
```

ثلاث إجابات متطابقة — لكن الحلقة التكرارية تستهلك أربعة أسطر، وتخفي قصدها، وعلى مليون نتيجة محلل تعمل أبطأ بمراتب من حساب NumPy المُجمّع. قاعدة الدورة: كل عملية تعمل على مجموعة كاملة دفعة واحدة — الدوال الجاهزة اليوم، والمعالجة الموجهة (Vectorization) بـ NumPy و pandas من اليوم الثاني. لن ترى كتلة for أو while أو if/else أخرى في أي شيفرة مُدرّسة في هذا الدليل.

الخطوة 5: مريض واحد في قاموس

ننتقل الآن إلى السجل نفسه. المفاتيح أسماء الحقول (سلاسل نصية بين علامتي اقتباس دائماً)؛ والقيم تحمل النوع المناسب من الجدول 1.1. الأقواس المربعة تقرأ قيمة بمفتاحها؛ والإسناد إلى مفتاح جديد يضيف حقلاً؛ والإسناد إلى مفتاح قائم يحدّثه — ثلاث عمليات تغطي معظم ما يفعله السجل الصحي الإلكتروني طوال يومه.

```

1 patient = {
2     "patient_id": "MOH-00042",
3     "name": "Ali Hassan",
4     "age": 58,
5     "sex": "M",
6     "smoker": True,
7 }
8
9 print(patient["name"], "is", patient["age"], "years old")
10 patient["governorate"] = "Baghdad" # add a new field
11 patient["age"] = 59 # update after a birthday
12 print(len(patient), "fields on record")
13 print(list(patient.keys()))

```

شيفرة 1.7: سجل مريض كأزواج مفتاح وقيمة

```
Ali Hassan is 58 years old
6 fields on record
['patient_id', 'name', 'age', 'sex', 'smoker', 'governorate']
```

الخطوة 6: قاموس متداخل يحمل لوحة صورة الدم

القيمة في القاموس يمكن أن تكون هي نفسها قاموساً. نلحق لوحة صورة الدم كسجل فرعي تحت المفتاح "cbc"؛ والأقواس المربعة المتسلسلة تنقب وصولاً إلى تحليل واحد بعينه، والمقارنة الأخيرة تحسب راية منطقية من البيانات — وهذه بذرة كل فرز آلي للنتائج غير الطبيعية في هذه الدورة.

```

1 patient["cbc"] = {"wbc": 12.7, "hgb": 10.9, "plt": 214}
2
3 print("WBC:", patient["cbc"]["wbc"], "x10^9/L")
4 print("Hgb:", patient["cbc"]["hgb"], "g/dL")
5 anemic = patient["cbc"]["hgb"] < 13 # adult male cut-off
6 print("Anemia flag:", anemic)

```

شيفرة 1.8: تداخل لوحة مختبرية داخل السجل

```

WBC: 12.7 x10^9/L
Hgb: 10.9 g/dL
Anemia flag: True

```

سريرياً: كريات بيض $12.7 \times 10^9/L$ كثرة كريات بيض خفيفة (الحد الأعلى 11)، وهيموغلوبين 10.9 غ/دل دون الحد الفاصل للذكور البالغ 13 غ/دل، فالرأية المحسوبة تبلغ عن فقر الدم بشكل صحيح.

الخطوة 7: الردهة قائمة قواميس

ثلاثة مرضى، ثلاثة قواميس بمفاتيح متطابقة، قائمة واحدة. الدليل [0] هو السجل الأول و[-1] الأخير — فبايثون تعدّ من الصفر، وهو عرف ستصادفه في كل مكان.

```

1 ward = [
2     {"patient_id": "MOH-00042", "age": 59, "sex": "M", "glucose": 182.0},
3     {"patient_id": "MOH-00107", "age": 34, "sex": "F", "glucose": 94.0},
4     {"patient_id": "MOH-00311", "age": 71, "sex": "M", "glucose": 143.0},
5 ]
6
7 print("Patients in ward:", len(ward))
8 print("First patient  :", ward[0]["patient_id"])
9 print("Last patient   :", ward[-1]["patient_id"])

```

شيفرة 1.9: ثلاثة سجلات، هيكل واحد

```

Patients in ward: 3
First patient  : MOH-00042
Last patient   : MOH-00311

```

الخطوة 8: التشويقة — سطر واحد يجعلها جدولاً

لا تشرح الخلية التالية اليوم؛ شغلها فقط وانظر. قائمة القواميس ذات المفاتيح المشتركة هيكل طبيعي إلى درجة أن pandas تحولته إلى جدول موسوم — إطار بيانات — في سطر واحد: المفاتيح تصبح أعمدة، والسجلات تصبح صفوفاً.

```

1 import pandas as pd
2
3 ward_table = pd.DataFrame(ward)
4 print(ward_table)

```

شيفرة 1.10: كائن الغد المحوري، في سطر واحد

```

patient_id age sex glucose
0 MOH-00042 59  M  182.0
1 MOH-00107 34  F   94.0
2 MOH-00311 71  M  143.0

```

والكائن نفسه يمتلئ مباشرة من ملف: فمجلد بيانات الدورة يضم 1,200 نتيجة صورة دم كاملة، تُقرأ بسطر واحد.

```
1 cbc = pd.read_csv("cbc_results.csv")
2 print(cbc[["patient_id", "wbc", "hgb", "plt"]].head(3))
```

شيفرة 1.11: لمحة مسبقة: قراءة مجموعة بيانات صورة الدم الخاصة بالدورة

	patient_id	wbc	hgb	plt
0	MOH-00001	6.7	16.3	292
1	MOH-00002	12.7	13.6	170
2	MOH-00003	8.2	13.9	234

ألف ومئتا مريض في سطرين. أما كيف يعمل هذا — وكل ما يمكنك فعله به — فهو موضوع اليوم الثاني بأكمله.

1.3 التحدي السريري (30 دقيقة)

التحدي السريري — ملف المريض الرقمي التجريبي في العناية المركزة

تجري المديرية تجربة استرشادية لملف مريض رقمي مهيكل في وحدة العناية المركزة في البصرة. بين يديك الملف الورقي للمريضة زينب كاظم، ملف MOH-00873، 63 عاماً، أنثى، البصرة، العناية المركزة. ملاحظات التمريض: درجة الحرارة 38.4 درجة مئوية، الضغط الانقباضي 148 ملم زئبق، ست قراءات نبض عبر الوريدية (92 و 101 و 88 و 110 و 97 و 104 نبضة في الدقيقة)، الوزن 82.0 كغم، الطول 1.74 م. صورة الدم الصباحية: كريات بيض 13.8، هيموغلوبين 10.4، هيماتوكريت 32.1، صفائح 512. فحص RT-PCR لكوفيد-19 (العينة PCR-0455، الجين الهدف ORF1ab): إيجابي، وقيمة Ct 18.4. بالعمل حصراً بالمتغيرات والقوائم والقواميس وسلاسل f — لا حلقات تكرارية ولا جملاً شرطية — أنجز المهام التالية:

1. أنشئ المتغيرين weight_kg و height_m واحسب $bmi = weight_kg / height_m^{*2}$.
2. ابن قاموساً متداخلاً واحداً باسم patient_file بأربعة أقسام بالضبط: "demographics" و "vitals" (يضم قراءات النبض الست في قائمة) و "cbc" و "rtPCR".
3. أدرج مؤشر كتلة الجسم المحسوب في قسم "vitals" كمفتاح جديد.
4. باستخدام الوصول بالأقواس المربعة المتسلسلة فقط، اطبع الهيموغلوبين وقيمة Ct لفحص PCR.
5. لخص قائمة النبض بالدوال len و min و max ومتوسط محسوب؛ واشتق راية منطقية tachycardic تكون True عندما تتجاوز أي قراءة 100 نبضة في الدقيقة (تلميح: قارن $\max(\dots)$ مع 100).
6. ركب ملخصاً سريرياً متعدد الأسطر بسلسلة f (خمس أسطر على الأقل: الهوية، والعمر والجنس، ومؤشر كتلة الجسم بمنزلة عشرية واحدة، وأبرز نتائج صورة الدم، ونتيجة RT-PCR ثم اطبعه).
7. بلغ عن مدى اكتمال السجل: عدد الأقسام العليا وعدد حقول صورة الدم.

✓ حل المدرب — ملف المريض الرقمي في العناية المركزة

يسير الحل بترتيب المهام تماماً. لاحظ كيف يكاد القاموس المتداخل يُقرأ مثل الملف الورقي نفسه — وهذه المقروئية هي جوهر الهيكل.

```
1 # Task 1 -- anthropometrics; ** is exponentiation (height squared)
2 weight_kg = 82.0
3 height_m = 1.74
4 bmi = weight_kg / height_m**2
5
6 # Task 2 -- the complete digital patient file
7 patient_file = {
8     "demographics": {
9         "patient_id": "MOH-00873",
10        "name": "Zainab Kadhim",
```

```

11     "age": 63,
12     "sex": "F",
13     "governorate": "Basra",
14     "department": "ICU",
15 },
16 "vitals": {
17     "temp_c": 38.4,
18     "systolic_bp": 148,
19     "pulse_readings": [92, 101, 88, 110, 97, 104],
20     "weight_kg": weight_kg,
21     "height_m": height_m,
22 },
23 "cbc": {"wbc": 13.8, "hgb": 10.4, "hct": 32.1, "plt": 512},
24 "rtPCR": {
25     "sample_id": "PCR-0455",
26     "target_gene": "ORF1ab",
27     "ct_value": 18.4,
28     "result": "Positive",
29 },
30 }
31
32 # Task 3 -- store the computed BMI inside the record
33 patient_file["vitals"]["bmi"] = bmi
34
35 # Task 4 -- targeted reads with chained access
36 print("Hemoglobin :", patient_file["cbc"]["hgb"], "g/dL")
37 print("PCR Ct      :", patient_file["rtPCR"]["ct_value"])
38
39 # Task 5 -- pulse statistics with list built-ins only
40 pulse = patient_file["vitals"]["pulse_readings"]
41 mean_pulse = sum(pulse) / len(pulse)
42 print("Pulse n/min/max:", len(pulse), min(pulse), max(pulse))
43 print(f"Mean pulse      : {mean_pulse:.1f} bpm")
44 tachycardic = max(pulse) > 100
45 print("Tachycardia episode:", tachycardic)

```

شيفرة 1.12: حل التحدي، المهام 1 إلى 5

```

Hemoglobin : 10.4 g/dL
PCR Ct      : 18.4
Pulse n/min/max: 6 88 110
Mean pulse   : 98.7 bpm
Tachycardia episode: True

```

```

1 # Task 6 -- multi-line clinical summary
2 demo = patient_file["demographics"]
3 summary = (
4     f"CLINICAL SUMMARY -- {demo['name']} ({demo['patient_id']})\n"
5     f"  Age/Sex : {demo['age']} y / {demo['sex']}, "
6     f"{demo['department']}, {demo['governorate']}\n"
7     f"  BMI      : {patient_file['vitals']['bmi']:.1f} kg/m^2\n"
8     f"  CBC       : Hgb {patient_file['cbc']['hgb']} g/dL, "
9     f"Plt {patient_file['cbc']['plt']} x10^9/L\n"
10    f"  RT-PCR    : {patient_file['rtPCR']['result']} "
11    f"(Ct {patient_file['rtPCR']['ct_value']}, "

```

```

12     f"{patient_file['rtPCR']['target_gene']} gene)"
13 )
14 print(summary)
15
16 # Task 7 -- how complete is the record?
17 print("Top-level sections:", len(patient_file))
18 print("CBC fields      :", len(patient_file["cbc"]))

```

شيفرة 1.13: حل التحدي، المهمتان 6 و 7

```

CLINICAL SUMMARY -- Zainab Kadhim (MOH-00873)
Age/Sex : 63 y / F, ICU, Basra
BMI      : 27.1 kg/m^2
CBC      : Hgb 10.4 g/dL, Plt 512 x10^9/L
RT-PCR   : Positive (Ct 18.4, ORF1ab gene)
Top-level sections: 4
CBC fields      : 4

```

القراءة السريرية. امرأة في الثالثة والستين محمومة (38.4 درجة مئوية) في العناية المركزة مع تسرع قلب متقطع (الحد الأقصى 110 نبضة في الدقيقة؛ الـ True). تُظهر صورة الدم كثرة كريات بيض (13.8 والحد الأعلى 11)، وفقر دم متوسطاً لامرأة بالغة (هيموغلوبين 10.4 والحد الفاصل 12)، وكثرة صفيحات (512 والحد الأعلى 450) — وهي ثلاثية تنسجم مع سرورة التهابية أو خمجية نشطة. ونتيجة RT-PCR إيجابية بقيمة Ct 18.4: تذكر المنطق العكسي — قيمة Ct المنخفضة تعني حملاً فيروسياً مرتفعاً، ما يستدعي عزلاً صارماً. ومؤشر كتلة الجسم 27.1 kg/m^2 يقع في نطاق زيادة الوزن (25–29.9). كل عبارة من هذه العبارات جاءت من وصول قاموسي ودوال قوائم جاهزة ومقارنات — دون حلقة واحدة أو جملة شرطية واحدة.

النقاش الجماعي

1. أي المستندات الورقية في مختبرك تنطبق بأكبر قدر من التلقائية على قاموس. وأي الحقول ستختارها مفاتيح؟ وهل سيختار كل مستشفى في محافظتك أسماء المفاتيح نفسها؟
2. الحقيقة الواحدة يمكن أن تُرمز "F" أو "female" أو "f" بمسافة زائدة. ما مصير إحصاء وطني يُحسب على ثماني محافظات تختلف على هذا النحو؟ (اليوم الثالث يهاجم هذه المشكلة بعينها في messy_lab_data.csv).
3. من في المديرية ينبغي أن يملك القائمة الرئيسية لأسماء الحقول ووحداتها — قاموس بيانات وطني — وكيف تفرض الالتزام به عند إدخال البيانات؟

1.4 الواجب المنزلي والتقييم الذاتي

الواجب المنزلي

1. ملف مريضك أنت. ابن قاموساً متداخلاً لمريض واحد من قسمك مجهول الهوية تماماً، بثلاثة أقسام على الأقل؛ واطبع ثلاث قيم بالوصول المتسلسل.
2. مراقبة البوتاسيوم. خزن القيم [3.1, 3.4, 3.9, 4.2, 5.6] ملي مول/لتر باسم potassium_readings؛ واطبع المتوسط بمزلتين عشريتين، ثم اشتق رايتين منطقيتين لنقص بوتاسيوم الدم ($\min(\dots) < 3.5$) وفرط بوتاسيوم الدم ($\max(\dots) > 5.1$).
3. سطر التقرير الكلوي. مع creatinine_mgdl = 1.37 و sodium_mmol = 141.6، ابن سطر سلسلة f واحداً يعرض الكرياتينين بمزلتين عشريتين والصوديوم بلا منازل.
4. وسّع الردهة. وسّع قائمة ward من الخطوة 7 إلى خمسة مرضى، وشغّل pd.DataFrame(ward) من جديد، وأسقط عمداً المفتاح "glucose" من أحد المرضى الجديدين. اكتب جملة واحدة تخمن فيها ما سيظهر في تلك الخانة. (اليوم الثالث يعطي ذلك اسماً).
5. إتقان Colab. أعد تسمية دفترتك، وأضف خلية نصية تصف ما يفعله الدفتر، وشاركه بصلاحيّة العرض مع زميل، ونزل نسخة بصيغة ..ipynb.

اختبار التحقق الذاتي

1. أي نوع بايثون ينبغي أن يخزن قيمة HbA1c البالغة 6.8%، وأي نوع يخزن عدد أسرة العناية المركزة؟
2. ماذا تعيد len(ward) عندما تكون ward قائمة قواميس مرضى؟
3. اكتب التعبير الذي يقرأ عدد الصفحات من patient = {"cbc": {"plt": 214}}.
4. اذكر الحد الفاصل لغلوكونز البلازما الصائم في تشخيص السكري، والحد الفاصل للسكر التراكمي HbA1c.
5. لماذا لا يحتاج دفتر Colab إلى pip install لمكتبي NumPy وpandas — وما الملف الصغير الذي يتيح لأي مشروع إعلان مكنتاته للآخرين؟

إجابات الاختبار. (1) float لقيمة HbA1c وint لعدد الأسرة. (2) عدد سجلات المرضى (القواميس) في القائمة. (3) patient["cbc"]["plt"] (4) غلوكونز صائم 126 ملغ/دل فأكثر في مناسبتين؛ وHbA1c 6.5% فأكثر (5.7–6.4% ما قبل السكري). (5) بيانات تشغيل Colab تصل والمنظومة العلمية مثبتة مسبقاً فيها؛ وملف requirements.txt يسرد اعتماديات المشروع حتى يستطيع أي جهاز إعادة إنشاء البيئة. النتائج المتوقعة للواجب 2: المتوسط 4.04 ملي مول/لتر، نقص البوتاسيوم True (الحد الأدنى 3.1)، فرط البوتاسيوم True (الحد الأعلى 5.6).

♣ القيمة لمديرية التشخيص — من الأرشفة الورقي إلى سجلات قابلة للاستعلام

تشتري دقائق اليوم التسعون للمديرية ثلاثة مكاسب. أولاً، بيئة تحليلات بلا كلفة: لا يحتاج Colab إلا إلى متصفح، فيستطيع كل إحصائي مستشفى في كل محافظة أن يبدأ غداً دون مشتريات أو تراخيص أو عتاد جديد. ثانياً، مفردات سجل مشتركة: فالموظف الذي يرى ملف المريض كأزواج مفتاح-قيمة منقطة يكتب سجلات أنظف، ويصمم قوالب أفضل، ويتواصل بدقة مع موردي أنظمة السجلات الصحية الإلكترونية (EHR) وأنظمة معلومات المختبر (LIS) — وهذه أول خطوة نحو قاموس بيانات وطني. ثالثاً، عادة قابلة للتدقيق: فالدفتر الذي يمزج الشيفرة والنتائج والشرح تقرير قابل لإعادة الإنتاج — وملخص مكافحة العدوى للشهر القادم يُعاد توليده في دقائق عندما تتحدث البيانات، بدلاً من إعادة طبعه من الورق.

خلاصات اليوم

- القيمة التشخيصية تتضاعف عندما تنتقل النتائج من الورق إلى سجلات مهيكلية قابلة للاستعلام؛ وهذه الدورة تبني ذلك المسار من طرفه إلى طرفه.
 - بايثون هي معيار البيانات الصحية: مجانية، وقابلة للتدقيق سرياً، ومحاطة بأعمق منظومة علمية.
 - Colab يشغل الدفاتر على بيئات تشغيل سحابية مؤقتة — بلا تثبيت، وقابلة للمشاركة كالمستندات؛ وملف requirements.txt يعلن الاعتماديات في الأماكن الأخرى.
 - الأنواع الأولية الأربعة تقابل البيانات السريية: int للعمر، float للكرياتينين، str لمعرف المريض، وbool لحالة التدخين — والأرقام بين علامتي اقتباس نصوص لا قياسات.
 - سلاسل f مع محدّدات التنسيق (f: .1f): تحوّل المتغيرات إلى أسطر تقارير مضبوطة بدقة.
 - الدوال len وmin وmax وsorted تلخص القائمة كلها دفعة واحدة — لا حلقات تكرارية، اليوم ولا في أي يوم قادم هنا.
 - القاموس سجل مريض واحد (مفتاح-قيمة، وهو شكل JSON الذي تتخذه السجلات الصحية الإلكترونية وواجهات المختبرات)؛ والقواميس تتداخل، وقائمة القواميس ردهة.
 - السطر pd.DataFrame(ward) حوّل تلك الردهة إلى جدول في سطر واحد — وهو الكائن الذي سيتمحور حوله كل ما هو قادم.
- غداً، في الفصل 2، نتوقف عن كتابة السجلات بأيدينا: فالدالة pd.read_csv ستحمل 1,200 سجل مريض حقيقي دفعة واحدة، وستجري أول تحليل استكشافي لمجموعات بيانات الدورة — متوسطات وأعداد وتوزيعات عبر كل محافظة في بضعة أسطر.

باب 2

اليوم 2 — استيراد السجلات الطبية وأول تحليل استكشافي

كل نظام من أنظمة معلومات المختبرات (LIS) في المديرية قادر على تصدير سجلاته في هيئة ملف مفصول بالفواصل، وكل ملف من تلك الملفات المصدرة يستقر اليوم في جدول بيانات إلكتروني. لقد خدمتنا جداول البيانات جيداً عندما كانت المحافظة تبليغ عن بضع مئات من العينات أسبوعياً، غير أنها لم تُصمَّم يوماً لتكون بنية تحتية تشخيصية: فهي تبتر الملفات الكبيرة، وتعيد تفسير المعرفات بصمت، ولا تترك أي أثر يدل على من غير أي خلية. في هذه الجلسة نستبدل بجدول البيانات الأداة التي توحدت عليها الوبائيات الحديثة والبحوث السريرية وتحليلات المستشفيات جميعاً: إطار البيانات (DataFrame) في مكتبة Pandas. ستُحمِل مجموعة بيانات تشخيصية حقيقية، وخلال خمسة أوامر فقط ستعرف كم مريضاً تضم، وأي المتغيرات رقمية وأيها نصية، وهل توجد قيم مفقودة (Missing Values)، وكيف يبدو الملف الإحصائي الكامل لكل مادة مقيسة — وهو المكافئ الحاسوبي لرفع الشريحة إلى المجهر قبل كتابة تقرير علم الأمراض. وبنهاية الدقائق التسعين ستنتج، من البيانات الخام إلى المذكرة المكتوبة، إحاطة يستطيع المدير العام أن يتخذ قراراً بناءً عليها.

الدقائق	الجزء	المحتوى والأنشطة
00-05	مراجعة اليوم الأول	هياكل البيانات الصحية، وتشريح ملفات CSV، وسير عمل الدفتر التفاعلي من الفصل 1: هدف اليوم: من الملف إلى الملف الإحصائي
05-10	النظرية الأولى	لماذا يتهار Excel: حدود الصفوف، والتحويل الصامت للأنواع، وانعدام قابلية إعادة الإنتاج، وغياب سجل التدقيق
10-20	النظرية الثانية	تشرح إطار البيانات (الفهرس والأعمدة والقيم والسلسلة): أنواع البيانات بوصفها بيانات وصفية تشخيصية؛ بروتوكول النظرية الأولى: المئينات مقابل المجالات المرجعية
20-30	التطبيق الموجه الأول	تحميل diabetes_dataset.csv: الأبعاد، والصفوف الأولى والأخيرة، والصورة الشعاعية التقنية بالأمر info()
30-40	التطبيق الموجه الثاني	تفسير describe() صففاً صففاً كتقرير مختبري؛ معدل الانتشار بالأمر value_counts()
40-50	التطبيق الموجه الثالث	إحصاءات السلسلة (المتوسط والوسيط والانحراف المعياري)، واختيار الأعمدة، والترتيب لإيجاد المرضى المتطرفين
50-60	التطبيق الموجه الرابع	أول قناع منطقي: عدد الغلوكوز ≥ 126 ملغ/دل؛ نقل البروتوكول كاملاً إلى cbc_results.csv
60-70	التحدي السريري الأول	الفرق تستخرج ملف chemistry_panel.csv: الأبعاد، و info() و describe() الكامل، والقيم المتطرفة المريبة
70-80	التحدي السريري الثاني	معدل انتشار سكر الدم بالقناع؛ خمسة مرضى متطرفين بالترتيب؛ صياغة مذكرة الجمل الخمس
80-90	الاستخلاص والختام	قراءة المذكرات، والنقاش الجماعي، وشرح الواجب المنزلي، وملحة عن تنظيف البيانات في اليوم الثالث

✓ الأهداف التعليمية

بنهاية هذه الجلسة ستكون قادراً على:

1. أن تشرح، بحالات فشل ملموسة، لماذا تُعد جداول البيانات غير مأمونة أداة رئيسية لتحليل البيانات المخترية على مستوى المديرية.
2. أن تُحمّل أي تصدير من نظام معلومات المختبر إلى إطار بيانات Pandas بالأمر `pd.read_csv()` وأن تذكر أبعاده من `df.shape`.
3. أن تسمي الأجزاء التشريحية الثلاثة لإطار البيانات (الفهرس والأعمدة والقيم) وأن تميز السلسلة (Series) بوصفها متغيراً سريياً واحداً.
4. أن تقرأ تقرير `df.info()` وتستخرج منه أنواع البيانات وأعداد القيم غير الفارغة وبصمة الذاكرة — وأن تشرح أهمية كل منها سريياً.
5. أن تفسر كل صف من صفوف `df.describe()` (العدد والمتوسط والانحراف المعياري والحد الأدنى والربيعيات والحد الأقصى) بلغة المجالات المرجعية.
6. أن تقيس معدل انتشار المرض بالأمر `value_counts(normalize=True)` وبمتوسط قناع منطقي (Boolean Mask).
7. أن تُنتج مذكرة مكتوبة من خمس جمل عن جودة البيانات ومعدل الانتشار من تصدير كيمياء خام في أقل من 20 دقيقة.

➔ حقيقة أدوات الذكاء الاصطناعي والاعتماديات — مكتبات اليوم

نعمل اليوم داخل Pandas بالكامل تقريباً، ونحتجها NumPy.

```
1 import pandas as pd
2 import numpy as np
3
4 pd.set_option("display.width", 100)
```

شيفرة 2.1: استيرادات اليوم الثاني

```
pip install pandas numpy openpyxl
```

pandas هي محرك الجداول القياسي عالمياً لتحليل البيانات: تقرأ تصديرات CSV وExcel من أي نظام معلومات مختبر، وتحفظ بها أطر بيانات، وتمنحنا بروتوكول النظرة الأولى كله بأوامر من سطر واحد. وبالنسبة إلى المديرية يعني هذا أن ملف الترخيص الذي كان يستهلك يوماً كاملاً من عمل جداول البيانات يصبح سكريتاً قابلاً لإعادة الإنتاج من خمسة أوامر. أما **numpy** فتوفر المصفوفات العددية السريعة التي تسكن داخل كل عمود من أعمدة إطار البيانات؛ وهي السبب في قدرة Pandas على تلخيص 1200 لوحة كيمياء في أجزاء من الثانية، وفي قابليتها للتوسع إلى ملايين السجلات دون تغيير سطر واحد من شيفرتنا. و**openpyxl** هي المساعد الهادئ الذي يتيح للأمر `pd.read_excel()` فتح ملفات `.xlsx`، فتدخل تصديرات Excel القديمة من المختبرات الإقليمية المسار نفسه الذي تدخله ملفات CSV. وخيار `display.width` يخبر Pandas ببساطة بعرض شاشتنا كي تبقى الجداول المطبوعة مقروءة.

2.1 السياق السريي والإطار النظري (20 دقيقة)

2.1.1 لماذا يجب أن ينتهي عصر جداول البيانات

اسأل أي مشرف مختبر عن طريقة إعداد التقرير الشهري وستسمع غالباً الجواب نفسه: تصدير من نظام معلومات المختبر، وفتح في Excel، وفرز، وفلتر، ونسخ، ولصق، ودعاء. والدعاء ضروري لأن جداول البيانات تفشل بطرائق تبقى خفية إلى أن تصبح علنية. في تشرين الأول 2020 فقدت هيئة الصحة العامة في إنكلترا (Public Health England) قرابة 16,000 حالة COVID-19 مؤكدة من التبليغ الوطني لأن النتائج مُرّرت عبر صيغة Excel قديمة لم تعد ورقة العمل فيها تتسع لمزيد من الصفوف — لم يحذر الملف أحداً؛ بل توقف فحسب. وقد حاربت دوريات الجينوم العدو نفسه سنوات: جداول تكميلية مليئة بأسماء جينات حُوّلت بصمت إلى تواريخ تقويمية. وإذا كان Excel

قادراً على إفساد رمز جين، فهو قادر على إفساد معرف مريض.

سقف جداول البيانات

أربعة أعطال بنيوية تجعل جداول البيانات غير مأمونة بوصفها بنية تحتية تشخيصية. (1) حدود صارمة لعدد الصفوف: تتوقف ورقة العمل الحديثة عند 1,048,576 صفراً (وصيغة x1s . القديمة عند 65,536)؛ والصفوف التي تتجاوز الحد تُهمل دون أي رسالة خطأ. (2) تحويل صامت للأنواع: معرف مثل 00042 يصبح الرقم 42، والنص الذي يشبه تاريخاً يصبح تاريخاً؛ والقيمة الأصلية لا سبيل إلى استرجاعها. (3) انعدام قابلية إعادة الإنتاج: سلسلة من النقرات اليدوية لا يمكن إعادة تشغيلها ولا فحصها ولا تسليمها إلى زميل؛ فالتحليل يعيش في الذاكرة العضلية لشخص واحد فقط. (4) غياب سجل التدقيق: الخلية التي كُتِب فوقها بضغطة زر طائشة لا تترك أي أثر لما كانت عليه، وهو أمر لا يمكن الدفاع عنه في تدقيق على مستوى الوزارة. أما السكرتير فيقلب الخصائص الأربع جميعاً: صفوف بلا حدود، وأنواع صريحة، وإعادة تنفيذ مطابقة، وسجل دائم قابل للمراجعة لكل خطوة.

مكتبة Pandas (من *panel data*)، أي بيانات اللوحات) هي الأداة التي أجمع عليها مجتمع تحليلات الصحة العالمي: منظمة الصحة العالمية، ومكاتب الإحصاء الوطنية، ووحدات البحوث في المستشفيات، وكل مسار تعلم آلي في هذه الدورة، جميعها تتحدث لغة DataFrame.

2.1.2 تشريح إطار البيانات

إطار البيانات: جدول بعقد صارم

إطار البيانات جدول ثنائي الأبعاد له ثلاثة أجزاء تشريحية. الفهرس (index) يعنون الصفوف؛ وفي عملنا يمثل الصف الواحد مريضاً واحداً أو عينة واحدة. الأعمدة (columns) هي المتغيرات السريرية — الغلوكوز ومؤشر كتلة الجسم والعمر والمخرجات — ولكل منها اسم ونوع بيانات واحد. القيم (values) هي القياسات نفسها، مخزنة داخلياً في مصفوفات NumPy سريعة. واختيار عمود واحد، مثل ["glucose", df]، يعيد سلسلة (Series): متجه أحادي البعد معنواً يمثل متغيراً سريرياً واحداً عبر جميع المرضى. وكل إحصاءة نحسبها اليوم تقريباً هي سؤال يُطرح على سلسلة. والعقد الذي يجعل هذا أكثر أماناً من جدول البيانات هو أن لكل عمود نوعاً واحداً معلناً بالضبط، وأن كل عملية تُطبَّق على أعمدة كاملة دفعة واحدة — فلا توجد خلية منفردة يمكن لضغطة زر طائشة أن تفسدها.

columns = clinical variables

	glucose	bmi	age	outcome	
index: one row = one patient	0	134	41.3	31	0
	1	149	34.7	37	0
	2	106	31.5	46	1

one column = a Series
(here: bmi)

شكل 2.1: تشريح إطار البيانات باستخدام المرضى الثلاثة الأوائل من مجموعة بيانات فحص السكري. الفهرس الرمادي يعنون الصفوف، وصف الترويسة الأزرق يسمي المتغيرات السريرية، والعمود المظلل يبين أن المتغير الواحد المستخرج من الجدول سلسلة (Series).

2.1.3 أنواع البيانات بيانات وصفية تشخيصية

أنواع البيانات: أي نوع من القيم يسكن هذا العمود؟

يحمل كل عمود نوع بيانات (dtype) يحدد ما يستطيع الإجابة عنه من أسئلة. فالعمود الرقمي (float64, int64) يدعم المتوسطات والمئينات ومقارنات العتبات؛ والعمود النصي (object) يدعم عدّ الفئات؛ وعمود التواريخ (datetime64) يدعم البوابات الزمنية. وتهتم أنواع البيانات سريرياً لأن النوع الخاطئ يعطل التحليل الصحيح بصمت: فإذا وصل عمود الغلوكوز نصاً

(ربما لأن إحدى الخلايا تحتوي pending)، فلن تستطيع Pandas حساب متوسطه، وسيتخطاه (`describe()`) ببساطة. لذلك فإن فحص أنواع البيانات بالأمر (`df.info()`) ليس طقساً برمجياً — إنه خطوة قبول العينة في تحليل البيانات: تحقق مما استلمته قبل أن تفحصه.

جدول 2.1: أنواع بيانات Pandas التي ستقابلها في بيانات المديرية.

النوع	ما يحتويه	مثال سريري
int64	أعداد صحيحة	العمر، عدد الصفحات، مدة الرقود
float64	أعداد عشرية	HbA1c، الكرياتينين، قيمة Ct
object	نصوص	patient_id، الجنس، المحافظة
bool	True / False	قناع الغلوكوز ≥ 126 ملغ/دل
datetime64	تواريخ وأوقات	admission_date، sample_date

نقطة ستهماً غداً: float64 هو أيضاً النوع الذي يحمل القيم المفقودة (NaN)؛ فإذا كان عدد المدخلات غير الفارغة في (`df.info()`) أقل من عدد الصفوف فذلك يعني نتائج غائبة — وفي ملف RT-PCR لدينا تكون قيم Ct فارغة للعينات السالبة بحكم التصميم.

2.1.4 بروتوكول النظرة الأولى

خمس أوامر، صفر معادلات

قبل أي تحليل، يشغل المحلل المنضبط الأوامر الخمسة نفسها، بالترتيب نفسه، على كل ملف جديد — تماماً كما يسمح اختصاصي أمراض الدم الشريحة بالعدسة الصغرى قبل الانتقال إلى العدسة الزيتية. (1) `df.tail()` / `df.head()`: تفقد الصفوف الأولى والأخيرة بعينيك؛ هل هؤلاء هم المرضى والمتغيرات التي توقعتها؟ (2) `df.shape`: كم عدد الصفوف (المرضى) والأعمدة (المتغيرات)؟ وهل يطابق العدد ما يدعي نظام معلومات المختبر أنه صدره؟ (3) `df.info()`: الصورة الشعاعية التقنية — أنواع البيانات وأعداد القيم غير الفارغة والذاكرة. (4) `df.describe()`: ملف إحصائي كامل لكل متغير رقمي. (5) `df["col"].value_counts()`: جدول تكرارات المتغير الفئوي؛ ومع `normalize=True` يتحول إلى معدل انتشار. خمسة أوامر، وتكون قد أجريت فحصاً استكشافياً كاملاً دون كتابة معادلة واحدة.

جدول 2.2: بروتوكول النظرة الأولى مقابل بالأسئلة السريرية.

الأمر	السؤال الذي يجيب عنه	النظير المختبري
<code>tail()</code> , <code>head()</code>	هل هذا هو الملف الذي أظنه؟	الفحص العياني
<code>shape</code>	كم مريضاً وكم متغيراً؟	عدّ العينات
<code>info()</code>	هل الأنواع مكتملة؟ وهل من قيم مفقودة؟	فحص القبول
<code>describe()</code>	ما توزيع كل مادة مقيسة؟	ملف ضبط الجودة
<code>value_counts()</code>	ما تكرار كل فئة؟	العد التفريقي

المئينات مجالات مرجعية متكررة

يعرض ناتج (`describe()`) المئينات 25 و50 و75. وهذه هي الكائنات الرياضية نفسها التي يستخدمها مختبرك أصلاً: فالمجال المرجعي التقليدي هو النسبة المركزية البالغة 95% من جمهرة مرجعية سليمة — أي المدى من المئين 2.5 إلى المئين 97.5. صف 50% هو الوسيط، القيمة التي تشطر المرضى نصفين، وهو مئين: قيمة شاذة عبثية واحدة (غلوكوز 9000 من خطأ إدخال بيانات) تجر المتوسط إلى الأعلى لكنها لا تكاد تحرك الوسيط. وعندما يختلف المتوسط والوسيط اختلافاً شديداً، فاشتبه إما في توزيع بيولوجي ملتو (الشحوم الثلاثية، قيم Ct) وإما في تلوث بقيم خاطئة. ومقارنة `max` و `min` بالحدود الفيزيولوجية المعروفة هي أسرع فحص لجودة البيانات على الإطلاق — وهي بالضبط الطريقة التي ستكتشف بها القيم المريبة في تحدي اليوم.

★ لؤلؤة سريرية — القيم المرجعية المستخدمة اليوم

جميع العتبات في هذا الفصل قيم قياسية للبالغين. تشخيص سكر الدم (وفق الجمعية الأمريكية للسكري (ADA): غلوكوز البلازما الصائم $100 < \text{ملغ/دل طبيعي}$ ؛ و $100-125 \text{ ملغ/دل}$ اختلال (ما قبل السكري)؛ و $126 \geq \text{ملغ/دل}$ مدى سكري يتطلب التأكيد في مناسبة ثانية. السكر التراكمي: $\text{HbA1c} < 5.7\%$ طبيعي؛ $5.7-6.4\%$ ما قبل السكري؛ $6.5\% \geq$ سكري. صورة الدم الكاملة (CBC): الكريات البيض $4-11 \times 10^9/\text{L}$ ؛ الهيموغلوبين $13-17 \text{ غ/دل}$ (رجال) و $12-15 \text{ غ/دل}$ (نساء)؛ الصفائح $150-450 \times 10^9/\text{L}$. الشوارد: الصوديوم $135-145 \text{ ملي مول/لتر}$ ؛ البوتاسيوم $3.5-5.1 \text{ ملي مول/لتر}$ — والقيم دون 2.5 أو فوق 6.0 ملي مول/لتر نتائج حرجة مصحوبة بخطر اضطراب النظم. القيمة المختبرية الواحدة خارج هذه الحدود راية للمراجعة، وليست تشخيصاً بذاتها أبداً: فالتشخيص يتطلب السياق السريري، ويتطلب للأمراض المزمنة تأكيداً.

2.2 التطبيق الموجّه (40 دقيقة)

■ عرض تطبيقي مباشر — استخراج الملف الإحصائي لتصدير فحص السكري

الهدف: تطبيق بروتوكول النظرة الأولى كاملاً على diabetes_dataset.csv (768 مريضاً مفحوصاً، على نمط بيانات Pima)، وتفسير كل ناتج تفسيراً سريرياً، وقياس معدل انتشار السكري بطريقتين مستقلتين، ثم إثبات أن البروتوكول ينتقل دون تغيير إلى تصدير صورة دم كاملة حقيقي (cbc_results.csv). يكتب المتدربون كل أمر بأيديهم؛ لا شيء يُنسخ ويُصق.

الخطوة 1: حُلّ الملف، وتحقق من الأبعاد، وألق نظرة أولى

يجب بدء الدفتر التفاعلي في المجلد الذي يحتوي مجموعات البيانات؛ عندئذٍ يُقرأ الملف باسمه المجرد. يحل الأمر read_csv() الملف النصي ويعيد كائن DataFrame.

```
1 import pandas as pd
2
3 pd.set_option("display.width", 100)
4
5 df = pd.read_csv("diabetes_dataset.csv")
6 print(type(df))
7 print(df.shape)
```

شيفرة 2.2: قراءة تصدير الفحص والتحقق من حجمه

```
<class 'pandas.core.frame.DataFrame'>
(768, 9)
```

يعيد shape دائماً (الصفوف، الأعمدة): 768 مريضاً وتسعة متغيرات سريرية. طابق هذا مع النظام المصدر — فإذا قال نظام معلومات المختبر إنه صدر 770 مراجعة واستلمت أنت 768، فتوقف وحقق في الأمر قبل حساب أي إحصاءة.

✖ خطأ شائع — الملف غير موجود

أكثر أخطاء اليوم الثاني شيوعاً هو FileNotFoundError: diabetes_dataset.csv. وهو لا يعني تقريباً أبداً أن الملف اختفى؛ بل يعني أن مجلد العمل الخاص بالدفتر ليس مجلد مجموعات البيانات، فتبحث Pandas في الغرفة الخطأ. تحقق من موقعك بالأمر `import os; os.getcwd()` وابدأ الدفتر من مجلد مجموعات البيانات (أو مرر المسار الكامل). والفرق التي توحد تخطيطاً واحداً لمجلد المشروع — كما فعلنا في اليوم الأول — لا تكاد ترى هذا الخطأ مرة أخرى.

بعد تأكيد الأبعاد، ألق نظرة على الصفوف نفسها. يعرض head() الصفوف الخمسة الأولى (هنا ثلاثة، لتناسب الصفحة)؛ ويعرض tail() الصفوف الأخيرة حيث تختبئ حالات بتر التصدير والسطور الذيلية التالفة. ومع تسعة أعمدة تختصر Pandas الوسط بعلامة حذف — البيانات سليمة، والعرض وحده مطوي.

```
1 print(df.head(3))
```

شيفرة 2.3: الصفوف الأولى من الملف المصدر

	pregnancies	glucose	blood_pressure	...	diabetes_pedigree	age	outcome
0	3	134	82	...	0.865	31	0
1	1	149	67	...	0.395	37	0
2	3	106	65	...	0.596	46	1

[3 rows x 9 columns]

نستطيع منذ الآن قراءة قواعد الجدول: المريض 0 يبلغ 31 عاماً وغلوكوزه 134 ملغ/دل ولا يحمل تشخيص السكري (outcome تساوي 0)؛ والمريض 2 يحمل التشخيص (outcome تساوي 1). والأرقام غير المعنونة في أقصى الطرف هي الفهرس.

الخطوة 2: الصورة الشعاعية التقنية بالأمر `info()`

```
1 df.info()
```

شيفرة 2.4: أنواع البيانات والاكتمال والذاكرة في تقرير واحد

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
#   Column                Non-Null Count  Dtype
---  -
0   pregnancies            768 non-null   int64
1   glucose                768 non-null   int64
2   blood_pressure         768 non-null   int64
3   skin_thickness         768 non-null   int64
4   insulin                768 non-null   int64
5   bmi                    768 non-null   float64
6   diabetes_pedigree      768 non-null   float64
7   age                    768 non-null   int64
8   outcome                768 non-null   int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

اقرأه كتقرير قبول: كل عمود رقمي (لم يصل أي غلوكوز نصاً)، وكل عمود يعرض 768 غير فارغة (لا نتائج مفقودة)، والجدول كله يكلف 54 كيلوبايت — وسنة كاملة من بيانات محافظة بأسرها ستظل تتسع في حاسوب محمول.

الخطوة 3: الملف الإحصائي الكامل بالأمر `describe()`

يحسب `describe()` ثماني إحصاءات لكل عمود رقمي. نقرب إلى منزلة عشرية واحدة وننقل الجدول بالأمر `T`. بحيث يصبح كل متغير صفراً — الوجهة نفسها التي يُطبع بها التقرير المختبري.

```
1 print(df.describe().round(1).T)
```

شيفرة 2.5: ملف إحصائي كامل دون أي معادلة

	count	mean	std	min	25%	50%	75%	max
pregnancies	768.0	2.7	1.7	0.0	2.0	3.0	4.0	8.0
glucose	768.0	116.2	30.1	44.0	96.0	114.0	138.0	211.0
blood_pressure	768.0	69.5	12.0	40.0	61.0	69.0	77.0	105.0
skin_thickness	768.0	26.5	7.8	7.0	21.0	27.0	32.0	50.0

insulin	768.0	115.5	90.9	15.0	53.0	97.0	154.2	850.0
bmi	768.0	31.7	6.7	15.0	27.2	31.6	36.2	52.5
diabetes_pedigree	768.0	0.5	0.3	0.1	0.2	0.4	0.6	2.5
age	768.0	38.7	11.2	21.0	30.0	37.0	44.0	81.0
outcome	768.0	0.4	0.5	0.0	0.0	0.0	1.0	1.0

انزل على صف الجلوكوز تماماً كما تقرأ مطبوعة ضبط جودة: العدد (count) 768 — لكل مريض قيمة؛ المتوسط (mean) 116.2 ملغ/دل — متوسط الجمهرة يقع أصلاً في مدى الاختلال: الانحراف المعياري (std) 30.1 — انتشار بيولوجي واسع؛ الحد الأدنى (min) 44 — هبوط سكر شديد، قيمة يجب التحقق منها في الملف السريري؛ صف 50% (الوسيط) 114 — قريب من المتوسط، فلا التواء جسيماً؛ صف 75% 138 — ربع كامل من المفحوصين عند عتبة التشخيص 126 ملغ/دل أو فوقها؛ الحد الأقصى (max) 211 — فرط سكر لا لبس فيه. وصف مؤشر كتلة الجسم يروي قصته الخاصة: وسيط قدره 31.6 كغ/م² يعني أن أكثر من نصف جمهرة الفحص هذه بدينة. ولاحظ أيضاً الإنسولين الذي يقع متوسطه (115.5) فوق وسيطه (97.0) بفارق كبير — توزيع ملتوٍ إلى اليمين، نموذجي للهرمونات، والوسيط فيه هو الملخص الأمين.

✖ خطأ شائع — نقاط عمياء في describe()

سلوكان صامتان يضللان المبتدئين. أولاً، يتخطى describe() الأعمدة غير الرقمية كلياً: في ملف صورة الدم الكاملة لن يقول شيئاً عن patient_id أو sample_date، ولن ينهك إلى ذلك. ثانياً، يحسب بكل سرور إحصاءات للفئات المرمزة: صف outcome أعلاه له متوسط قدره 0.4، وهو ذو معنى فقط لأن الترميز 0/1 (إنه معدل الانتشار!)، بينما ربيعياته وانحرافه المعياري بلا معنى سريري. القاعدة: describe() للقياسات و value_counts() للفئات — واعرف دائماً إلى أي النوعين ينتمي العمود قبل اقتباس أي إحصاء منه.

الخطوة 4: معدل الانتشار بالأمر value_counts()

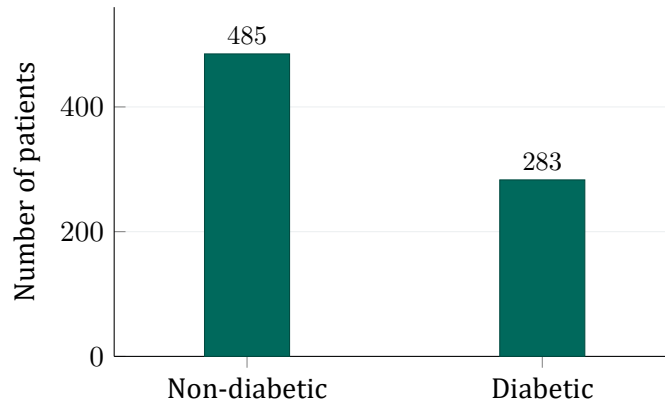
بالنسبة إلى المتغير الفئوي، جدول التكرارات هو التحليل. والخيار normalize=True يحول الأعداد إلى نسب — معدل انتشار جاهز لشرحته إحاطة.

```
1 print(df["outcome"].value_counts())
2 print((df["outcome"].value_counts(normalize=True) * 100).round(1))
```

شيفرة 2.6: الأعداد ثم معدل الانتشار لمخرجات السكري

```
outcome
0    485
1    283
Name: count, dtype: int64
outcome
0    63.2
1    36.8
Name: proportion, dtype: float64
```

283 من 768 مريضاً مفحوصاً (36.8%) يحملون وسم السكري — هذه فئة فحص عالية الخطورة وليست الجمهرة العامة، وذلك سياق يجب أن يظهر في أي تقرير يقتبس الرقم.



شكل 2.2: توزيع المخرجات في مجموعة بيانات فحص السكري كما أعادها الأمر `value_counts()` تماماً: 485 مريضاً غير سكري مقابل 283 مريضاً سكرياً (معدل انتشار 36.8% في هذه الفئة عالية الخطورة).

الخطوة 5: متغير واحد تحت المجهر — إحصاءات السلسلة

اختيار عمود واحد بالأمر `df["glucose"]` يعيد سلسلة، والسلسلة تجيب عن الأسئلة الإحصائية مباشرة.

```
1 print(round(df["glucose"].mean(), 1))
2 print(round(df["glucose"].median(), 1))
3 print(round(df["glucose"].std(), 1))
4 print(df["glucose"].min(), df["glucose"].max())
```

شيفرة 2.7: المتوسط والوسيط والتشتت لمتغير سريري واحد

```
116.2
114.0
30.1
44 211
```

المتوسط 116.2 مقابل الوسيط 114.0: يتفق الاثنان ضمن 2 ملغ/دل، فالتوزيع متناظر تقريباً وكلا المخلصين مقبول. وعندما تباعد عن متغير يفتقران فيه (تذكر الإنسولين) فاقتبس الوسيط.

الخطوة 6: اختيار الأعمدة ذات الصلة السريرية

الأقواس المزدوجة تختار قائمة أعمدة وتعيد إطار بيانات أصغر — وهي الطريقة القياسية لتركيز تصدير LIS عريض على المتغيرات التي تحتاجها السؤال فعلاً.

```
1 print(df[["glucose", "bmi"]].describe().round(1))
```

شيفرة 2.8: ملف ثنائي المتغيرات: سكر الدم والسمنة

```
count    glucose    bmi
mean      116.2    31.7
std        30.1     6.7
min        44.0    15.0
25%        96.0    27.2
50%       114.0    31.6
75%       138.0    36.2
max       211.0    52.5
```

تذكره للأقواس: `df["glucose"]` (زوج واحد) يعيد سلسلة؛ و `df[["glucose", "bmi"]]` (زوجان، وفي الداخل قائمة) يعيد إطار بيانات.

الخطوة 7: الترتيب لإيجاد المرضى المتطرفين

يرتب `sort_values()` الجدول كله وفق عمود واحد. والترتيب التنازلي يُظهر المرضى الأشد فرطاً في سكر الدم — الصفوف التي يريد الطبيب سحبها أولاً.

```
1 top5 = df.sort_values("glucose", ascending=False)
2 print(top5[["glucose", "bmi", "age", "outcome"]].head())
```

شيفرة 2.9: المرضى الخمسة الأعلى غلوكوزاً بين المفحوصين

	glucose	bmi	age	outcome
510	211	52.5	26	1
206	207	48.0	45	1
659	205	46.6	31	1
432	204	43.1	45	1
162	203	35.8	34	1

المرضى الخمسة جميعاً فوق 200 ملغ/دل يحملون وسم السكري، وأربعة منهم بدينون بشدة — أول تلميح بصري إلى العلاقة بين الغلوكوز ومؤشر كتلة الجسم والمخرجات، وهي العلاقة التي ستكتملها أيام التعلم الآلي. وأرقام الفهرس تسافر مع صفوفها: المريض 510 يبقى المريض 510 بعد الترتيب — وهذه بالضبط حماية الهوية التي يمكن لفرز في جدول بيانات أن يدمرها بصمت.

الخطوة 8: أول مذاق للقناع المنطقي — العدّ

مقارنة سلسلة كاملة بعتبة تُنتج سلسلة من قيم True/False — أي قناعاً منطقياً. ولأن True تُحسب 1، فإن `sum()` يعدّ المرضى فوق العتبة و `mean()` يعطي النسبة مباشرة. اليوم نكتفي بالعدّ بالأقنعة؛ وفي اليوم الرابع سنستخدمها لسحب صفوف المرضى أنفسهم.

```
1 high_glucose = df["glucose"] >= 126
2 print(high_glucose.head(3))
3 print(high_glucose.sum())
4 print(round(high_glucose.mean() * 100, 1))
```

شيفرة 2.10: كم مريضاً يقع في المدى الغلوكوزي السكري؟

0	True
1	True
2	False
Name: glucose, dtype: bool	
292	
38.0	

292 مريضاً (38.0%) يستوفون معيار الغلوكوز ≥ 126 ملغ/دل — رقم قريب من التشخيصات المسجلة البالغة 283 (36.8%) لكنه غير مطابق لها: نتعامل مع غلوكوز هذه المجموعة التعليمية بوصفه قيمة صائمة، والتشخيص يتطلب فوق ذلك تأكيداً وتقييماً سريرياً. وشبه التوافق بين الرقمين هو بذاته فحص جودة.

الخطوة 9: البروتوكول ينتقل دون تغيير — تصدير CBC حقيقي

لم يكن أي مما سبق خاصاً بالسكري: فعند توجيه الأوامر نفسها إلى تصدير صورة الدم الكاملة في المديرية، تستخرج ملف أمراض الدم بالطلاقة نفسها.

```
1 cbc = pd.read_csv("cbc_results.csv")
2 print(cbc.shape)
3 print(cbc[["patient_id", "wbc", "hgb", "plt"]].head())
4 print(cbc[["wbc", "hgb", "plt"]].describe().round(1))
```

شيفرة 2.11: نظرة أولى على 1200 صورة دم كاملة

```
(1200, 9)
  patient_id  wbc  hgb  plt
0  MOH-00001   6.7  16.3  292
1  MOH-00002  12.7  13.6  170
2  MOH-00003   8.2  13.9  234
3  MOH-00004   7.7  12.7  212
4  MOH-00005   5.2  15.9   40
      wbc      hgb      plt
count  1200.0  1200.0  1200.0
mean     7.6    13.9   268.5
std     2.5     1.6    78.4
min     2.7     9.8    40.0
25%     5.8    12.8   218.0
50%     7.1    13.8   268.5
75%     8.9    15.0   320.0
max    20.0    19.5   508.0
```

حتى الصفوف الخمسة الأولى حية سريرياً: المريض MOH-00002 كرياتة البيض 12.7 (كثرة كريات بيض) والمريض MOH-00005 صفيحاته 40 — نقص صفيحات شديد يطلق في النظام الحي نداء قيمة حرجة. وكتلة describe تؤكد ذلك على نطاق واسع: أدنى صفيحات 40، وأدنى هيموغلوبين 9.8 غ/دل. وعدان بالقناع يكملان التقرير:

```
1 print((cbc["wbc"] > 11).sum())
2 print((cbc["plt"] < 150).sum())
```

شيفرة 2.12: عد كثر الكريات البيض ونقص الصفيحات بالقناع

```
104
83
```

في شاشة واحدة: 104 من 1200 عينة (8.7%) تُظهر كثرة كريات بيض و83 عينة (6.9%) نقص صفيحات — تقرير صباحي على مستوى المختبر كله، قابل للتكرار غداً بالأسطر الأربعة نفسها.

2.3 التحدي السريري (30 دقيقة)

التحدي السريري — استخراج ملف لوحة الكيمياء لإحاطة المديرية

طلب المدير العام إحاطة في اليوم نفسه عن الجودة ومعدل الانتشار بشأن تصدير الكيمياء الموحد chemistry_panel.csv (لوحة واحدة لكل مريض) قبل إحالته إلى البرنامج الوطني للأمراض غير الانتقالية (NCD). اعملوا في أزواج لإنتاج الأدلة والمذكرة:

1. استوردوا Pandas، وحملوا chemistry_panel.csv، وبلغوا عن عدد المرضى والمتغيرات من shape.
2. شغلوا info(): أكدوا نوع بيانات كل مادة مقيسة وما إذا كانت هناك قيم مفقودة.
3. أنتجوا الملف الإحصائي الكامل بالأمر T.describe().round(1).
4. من عمودي الحد الأدنى والحد الأقصى، حددوا ثلاثة متغيرات قيمها المتطرفة مريبة سريرياً بدرجة تستوجب مراجعتها مقابل سجل المريض (وهي موجودة فعلاً — استعملوا القيم المرجعية من قسم النظرية).
5. باستخدام الأقنعة المنطقية مع sum() وmean()، احسبوا عدد ونسبة المرضى ذوي الغلوكوز الصائم ≥ 126 ملغ/دل.
6. كرروا ذلك للسكر التراكمي HbA1c $\geq 6.5\%$.
7. رتبوا وفق مادة مقيسة مريبة واحدة واعرضوا المرضى الخمسة الأشد تطرفاً مع معرفاتهم.

8. اكتبوا مذكرة نتائج من خمس جمل للمدير العام: الأبعاد والاكتمال، ورقما الانتشار، والقيم المتطرفة المرببة، وإجراء واحد موصى به.

✓ حل المدرب — ملف لوحة الكيمياء خطوة بخطوة

السكرتير الكامل للاستخراج، ومخرجاته الحقيقية، والمذكرة.

```
1 import pandas as pd
2
3 pd.set_option("display.width", 100)
4
5 chem = pd.read_csv("chemistry_panel.csv")
6 print(chem.shape)
7 chem.info()
```

شيفرة 2.13: حل التحدي: التحميل والتحقق والاستخراج

```
(1200, 13)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1200 entries, 0 to 1199
Data columns (total 13 columns):
#   Column                Non-Null Count  Dtype
---  -
0   patient_id            1200 non-null  object
1   glucose_mgdl          1200 non-null  int64
2   hba1c_pct             1200 non-null  float64
3   creatinine_mgdl       1200 non-null  float64
4   urea_mgdl             1200 non-null  int64
5   alt_ul                1200 non-null  int64
6   ast_ul                1200 non-null  int64
7   total_chol_mgdl       1200 non-null  int64
8   triglycerides_mgdl    1200 non-null  int64
9   hdl_mgdl              1200 non-null  int64
10  ldl_mgdl              1200 non-null  int64
11  sodium_mmol           1200 non-null  int64
12  potassium_mmol        1200 non-null  float64
dtypes: float64(3), int64(9), object(1)
memory usage: 122.0+ KB
```

1200 مريض و13 متغيراً، وجميع المواد المقيسة رقمية، ولا قيم مفقودة — التصدير يجتاز القبول. والآن الملف الإحصائي:

```
1 print(chem.describe().round(1).T)
```

شيفرة 2.14: حل التحدي: الملف الإحصائي الكامل

	count	mean	std	min	25%	50%	75%	max
glucose_mgdl	1200.0	97.2	27.4	55.0	77.0	97.0	115.0	186.0
hba1c_pct	1200.0	5.6	1.0	4.2	4.7	5.5	6.2	8.8
creatinine_mgdl	1200.0	0.9	0.2	0.4	0.7	0.9	1.0	1.6
urea_mgdl	1200.0	27.1	10.1	10.0	20.0	27.0	34.0	68.0
alt_ul	1200.0	29.4	16.1	5.0	18.0	26.0	37.0	158.0
ast_ul	1200.0	23.6	14.1	5.0	13.0	21.0	31.0	118.0
total_chol_mgdl	1200.0	185.0	38.9	90.0	157.0	185.0	212.0	309.0
triglycerides_mgdl	1200.0	146.6	70.6	40.0	95.0	132.0	182.0	525.0

hdl_mgdl	1200.0	47.6	9.8	20.0	41.0	48.0	54.0	83.0
ldl_mgdl	1200.0	108.7	41.1	30.0	78.0	108.5	138.0	246.0
sodium_mmol	1200.0	139.5	3.0	130.0	137.0	139.0	141.0	152.0
potassium_mmol	1200.0	4.2	0.5	2.9	3.9	4.2	4.5	5.6

القيم المتطرفة المريبة (المهمة 4). بقراءة الحدين الأدنى والأقصى مقابل القيم المرجعية: (أ) glucose_mgdl بحد أدنى 55 — هبوط سكر عند مستوى يسبب أعراضاً؛ وفي لوحة مرضى خارجيين صائمين يستوجب هذا مراجعة الملف (هبوط سكر حقيقي؟ أم تأخر في فصل العينة؟). (ب) potassium_mmol بمدى 2.9-5.6 — كلا الطرفين شاذ: 2.9 نقص بوتاسيوم واضح، و5.6 يتجاوز الحد الأعلى 5.1؛ وكلاهما قد يكون بيولوجياً أو قبل تحليلي (انحلال الدم يرفع البوتاسيوم كاذباً). (ج) sodium_mmol بمدى 130-152 — يمتد من نقص صوديوم حقيقي إلى فرط صوديوم عند مستوى (> 150) يقترن بتجفاف شديد. وتنويه مستحق: الحد الأقصى لإنزيم ALT البالغ 158 وحدة/لتر، نحو أربعة أضعاف الحد المرجعي الأعلى، متوافق مع أذية خلوية كبدية حادة. لا شيء من هذه القيم مستحيل — وهذا بالضبط سبب حاجتها إلى تحقق سريري بدل الحذف الصامت.

```
1 diabetic_glc = chem["glucose_mgdl"] >= 126
2 diabetic_a1c = chem["hba1c_pct"] >= 6.5
3 print(diabetic_glc.sum())
4 print(round(diabetic_glc.mean() * 100, 1))
5 print(diabetic_a1c.sum())
6 print(round(diabetic_a1c.mean() * 100, 1))
7
8 worst_k = chem.sort_values("potassium_mmol", ascending=False)
9 print(worst_k[["patient_id", "potassium_mmol", "creatinine_mgdl"]].head())
```

شيفرة 2.15: حل التحدي: معدلا الانتشار الغلوكوزيان والقيم المتطرفة

191		
15.9		
237		
19.8		
patient_id	potassium_mmol	creatinine_mgdl
599 MOH-00600	5.6	1.13
240 MOH-00241	5.6	0.86
917 MOH-00918	5.5	0.97
532 MOH-00533	5.5	1.03
43 MOH-00044	5.4	1.08

التفسير السريري. يتفق المؤشران الغلوكوزيان في الاتجاه لا في المقدار: 191 مريضاً (15.9%) يتجاوزون عتبة الغلوكوز الصائم بينما يتجاوز 237 (19.8%) عتبة السكر التراكمي HbA1c. وهذا متوقع — فالسكر التراكمي يدمج ثلاثة أشهر من سكر الدم ويلتقط المرضى الذين صادف أن عينتهم الصائمة الوحيدة كانت مقبولة. وقيم البوتاسيوم الخمس الأعلى (5.4-5.6 ملي مول/لتر) تحدث مع كرياتينين طبيعي، ما يبعد احتمال الفشل الكلوي ويرجح إما تغيراً بيولوجياً مبكراً وإما عينات منحلّة الدم؛ وعلى المختبر فحص مؤشر انحلال الدم لتلك العينات الخمس.

المذكرة (المهمة 8). إلى المدير العام، مديرية التشخيص. يغطي تصدير الكيمياء الموحد 1200 مريض عبر 13 مادة مقيسة دون قيم مفقودة، وهو يجتاز القبول التقني. الدليل المختبري على السكري كبير: 191 مريضاً (15.9%) غلوكوزهم الصائم 126 ملغ/دل أو أعلى، و237 مريضاً (19.8%) سكرهم التراكمي 6.5% أو أعلى، أي أن نحو واحد من كل خمسة مرضى مفحوصين يستوفي معياراً تشخيصياً. تُظهر ثلاث مواد مقيسة قيماً متطرفة تستوجب مراجعة على مستوى السجل قبل الإصدار العام: الغلوكوز نزولاً إلى 55 ملغ/دل، والبوتاسيوم الممتد من 2.9 إلى 5.6 ملي مول/لتر، والصوديوم الممتد من 130 إلى 152 ملي مول/لتر، وكلها خارج المجال المرجعي بوضوح، مع اقتراب غلوكوز 55 ملغ/دل من منطقة القيم الحرجة. نوصي بفحص تأكيد للحوالات الغلوكوزية المحددة حديثاً وبفحص مؤشر انحلال الدم للعينات الخمس الأعلى بوتاسيوماً. وقد أنتج هذا الملف بسكريت قابل لإعادة الاستخدام ويمكن إصداره شهرياً لكل مختبر إقليمي دون أي جهد إضافي.

نقاش جماعي.

1. متوسط السكر التراكمي HbA1c في هذه المجموعة 5.6% — دون عتبة ما قبل السكري — ومع ذلك فإن 19.8% من المرضى في

- المدى السكري. كيف تصح العبارتان معاً، وأي الرقمين ينتهي إلى إحاطة موجهة إلى صناع السياسات؟
2. أي القيم المتطرفة المريبة أرجح أن تكون قبل تحليلية (انحلال دم، تأخر فصل، تمديد من ذراع موصولة بمحلول وريدي) وأنها أرجح أن تكون بيولوجية؟ وما البيانات الإضافية التي تحسم كل حالة؟
3. لو أنتجت هذه المذكرة في جدول بيانات، فما الذي لا يستطيع المدقق التحقق منه بالضبط بينما يستطيع التحقق منه مع سكرتنا؟

2.4 الواجب المنزلي والتقييم الذاتي

- تمارين الواجب المنزلي. اعمل في مجلد مجموعات البيانات؛ كل تمرين خلية جديدة في الدفتر.
- الملف القلبي. حمّل `heart_risk.csv`. بلّغ عن أبعاده، وشغّل `info()`، وأنتج `T.describe().round(1)`. ثم استخدم `value_counts()` على `chest_pain_type`: أي نمط تقديم سريري يهيمن على هذه الفئة، ولماذا يُعد ذلك لافتاً سريرياً في مجموعة بيانات خاصة بنقص التروية؟
 - تقرير المختبر الجزيئي. حمّل `rtPCR_results.csv` وشغّل `info()`. فسر في جملة واحدة لماذا لا يملك العمود `ct_value` سوى 429 مدخلة غير فارغة من أصل 800 — هل هذه مشكلة في جودة البيانات؟ ثم احسب نسبة كل فئة من فئات `result` بالأمر `value_counts(normalize=True)`.
 - عداد إعادة الرقود. حمّل `patients_demographics.csv` واستخدم قناعاً منطقياً (المساواة تعمل تماماً كالعتبة: `df["readmitted_30d"] == "Yes"`) للتبليغ عن عدد ونسبة حالات إعادة الرقود خلال 30 يوماً.
 - فحص فقر الدم. في `cbc_results.csv`، بلّغ عن المتوسط والوسيط والانحراف المعياري للعمود `hgb`، ثم عدّ العينات التي يقل فيها `hgb` عن 12 غ/دل وعبر عنها نسبة مئوية.
 - تأمل. في خمسة أسطر مكتوبة، عدّد ثلاث طرائق ملموسة كان يمكن بها لسير عمل قائم على Excel أن يفسد `patients_demographics.csv` دون أن يلاحظ أحد، وأي أمر من أوامر النظرة الأولى كان سيكشف كل إفساد منها.

اختبار التقييم الذاتي.

- يعيد `df.shape` القيمة (13, 1200). ماذا يعني الرقمان بالضبط، وبأي ترتيب؟
- أي أمر واحد يكشف، لجميع الأعمدة دفعة واحدة، نوع بيانات كل عمود وكم قيمة مفقودة فيه؟
- في ناتج `describe()`، أي صف هو الوسيط، ولماذا الوسيط أجدر بالثقة من المتوسط عندما يختبئ في العمود خطأ إدخال بيانات مثل غلوكوز 9000؟
- ما الذي يتغير في ناتج `value_counts()` عندما تمرر `normalize=True`، وما الاسم الوبائي للنتيجة في متغير مرضي؟
- دون تشغيله، اذكر ما يعيده `mean().chem["hba1c_pct"] >= 6.5` وفسر لماذا تنجح هذه الحيلة.

إجابات الاختبار. (1) 1200 صف (مرضى) و13 عموداً (متغيرات)، والصفوف أولاً دائماً. (2) `df.info()` — أنواع البيانات مع أعداد القيم غير الفارغة في تقرير واحد. (3) صف 50%؛ فالوسيط لا يعتمد إلا على ترتيب الرتب، لذا تحركه القيمة الشاذة العبثية الواحدة تحريكاً ضئيلاً بينما ينجر المتوسط نحو الخطأ. (4) تتحول الأعداد إلى نسب مجموعها 1؛ وفي متغير مرضي تكون نسبة الفئة الإيجابية هي معدل الانتشار. (5) يعيد 0.1975، أي نسبة المرضى ذوي `HbA1c` $\geq 6.5\%$ — قيمة True تُحسب 1 وFalse تُحسب 0، لذا فمتوسط القناع هو كسر القيم True.

♣ القيمة لمديرية التشخيص — من التصدير إلى الإحاطة في عشرين دقيقة

مهارة اليوم تحول أي تصدير من نظام معلومات المختبر — دموياً كان أم كيميائياً أم جزيئياً — إلى ملف إحصائي موثّق بخمسة أوامر، وقد أثبت التحدي الدورة الكاملة: من CSV خام إلى مذكرة للمدير العام في أقل من عشرين دقيقة. ومؤسسياً يعني هذا أن إحاطات الترصد التي تستهلك حالياً يوم عمل محلل كامل في جداول البيانات تصبح سكريات قابلة لإعادة الإنتاج يستطيع أي موظف مدرب إعادة تشغيلها شهرياً، لكل محافظة، مع سجل تدقيق تستطيع الوزارة الدفاع عنه. والبروتوكول نفسه هو خطوة القبول الإلزامية قبل مرحلتى التنظيف (اليوم الثالث) والتعلم الآلي (الأيام من السادس إلى العاشر): لن يُدرّب أي نموذج يُبنى في هذه الدورة على ملف لم يجتز هذا الفحص ذا الأوامر الخمسة.

خلاصات اليوم

- جداول البيانات تفشل بصمت — حدود صفوف، وتحويل أنواع، وغياب سجل التدقيق — ولذلك فهي غير مأمونة بنيةً تحتية تشخيصية رئيسية؛ أما السكريبتات فقابلية لإعادة الإنتاج والمراجعة.
- إطار البيانات هو الفهرس والأعمدة والقيم معاً؛ والعمود الواحد سلسلة، والصف الواحد مريض واحد أو عينة واحدة.
- `df.info()` هو خطوة قبول العينة: يجب التحقق من أنواع البيانات وأعداد القيم غير الفارغة قبل اقتباس أي إحصاء.
- بروتوكول النظرة الأولى — `head` و `shape` و `info` و `describe` و `value_counts` — يقدم فحصاً استكشافياً كاملاً بصفر معادلات.
- مئينات `describe()` هي الكائنات نفسها التي تمثلها المجالات المرجعية؛ ومقارنة الحدين الأدنى والأقصى بالفيزيولوجيا أسرع فحص متاح لجودة البيانات.
- بلغ عن الوسيط لا المتوسط عندما يكون التوزيع ملتوياً أو ملوثاً — وقد أثبت الإنسولين والشحوم الثلاثية ذلك اليوم.
- القناع المنطقي المقارن بعبئة سريرية يعدّ المرضى بالأمر `sum()`. ويعطي معدل الانتشار بالأمر `mean()`: الغلوكوز ≥ 126 ملغ/دل رصد 292 من 768 مريضاً مفحوصاً (38.0%).
- البروتوكول لا يتقيد بمجموعة بيانات بعينها: الأوامر الخمسة المتطابقة استخرجت ملف فحص سكري وتصدير صورة دم كاملة ولوحة كيمياء.
- غداً (الفصل 3) نواجه الملف الذي بُني هذا البروتوكول لاصطياده: `messy_lab_data.csv` بقيمه المفقودة، وقيمه الشاذة (Outliers) المستحيلة، وترميز الجنس غير المتسق فيه، وصفوفه المكررة — وتتعلم تنظيفه دون أن نلمس خلية واحدة بأيدينا.

باب 3

اليوم 3 — تنظيف البيانات الطبية

يرفع مختبر أحد الأقضية تصديره الشهري، فتعلن لوحة متابعة الترصد أن متوسط سكر الدم في منطقة التغطية قد تضاعف. تُراجع خطة شراء الأنسولين، وتُصاغ مسودة حملة توعية بالسكري — ثم يفتح أحدهم الملف فيجد قيمة غلوكوز واحدة مقدارها 9000 ملغ/دل، وهي قيمة لم ينتجها إنسان حي قط. خطأ نسخ واحد، قُبِل دون تمحيص، كاد يعيد توجيه ميزانية كاملة. هذا هو نمط الفشل الصامت في المعلوماتية الصحية: الحاسوب لا يرفض البيانات الملوثة أبداً، بل يحسب منها هراءً واثقاً من نفسه. بالأمس، في الفصل 2، تعلمت استيراد الجداول المختبرية ووصفها؛ واليوم تتعلم أن ترتبها ترتيباً منتجاً. سنأخذ تصديراً مختبرياً أُفسد عمداً — نتائج مفقودة، وإرسالات مكررة، وتسع كتابات مختلفة لجنسين اثنين، وقيم مستحيلة فيسيولوجياً — ونحوّله إلى مجموعة بيانات تستطيع المديرية اعتمادها، مع عدّ كل تصحيح وتسويغه وتدوينه في سجل التنظيف. إن كل نموذج ذكاء اصطناعي تشخيصي سنبنيه من اليوم السادس فصاعداً سيقف على الانضباط الذي تتمرن عليه اليوم.

الدقائق	الجزء	المحتوى والأنشطة
00-05	المراجعة والمدخل السريري	مراجعة اليوم الثاني: <code>read_csv</code> و <code>describe</code> : قصة قيمة 9000 ملغ/دل — بيانات ملوثة في المدخلات، تشخيص خاطئ في المخرجات
05-12	النظرية الأولى: لماذا تكون البيانات ملوثة	الإدخال اليدوي، وأعطال المحلات، والخلط بين الوحدات، ودمج الأنظمة القديمة؛ الكلفة السريرية للتلوث في الذكاء الاصطناعي التشخيصي
12-20	النظرية الثانية: العيوب الأربعة	القيم المفقودة وقرار الحذف مقابل التعويض؛ التكرارات؛ الترميز غير المتسق؛ القيم الشاذة وسياج المدى الربيعي
20-28	التطبيق الموجه 1	تحميل <code>messy_lab_data.csv</code> ؛ الأبعاد وأنواع الأعمدة؛ إحصاء القيم المفقودة باستخدام <code>isna().sum()</code> والنسب المئوية
28-38	التطبيق الموجه 2	بوابة الهوية باستخدام <code>dropna(subset=)</code> ؛ التعويض بوسيط الغلوكوز مع راية تدقيق
38-46	التطبيق الموجه 3	التكرارات التامة باستخدام <code>duplicated()</code> ؛ الترميز القياسي للجنس بدوال <code>map</code> و
46-54	التطبيق الموجه 4	سياج المدى الربيعي من <code>quantile()</code> ؛ قناع القيم الشاذة؛ قرار وضع الراية مقابل الحذف باستخدام <code>np.where</code>
54-60	التطبيق الموجه 5	بوابات المعقولة الفسيولوجية باستخدام <code>between()</code> ؛ مقارنة <code>describe</code> قبل التنظيف وبعده
60-80	التحدي السريري	الفرق تعتمد مجموعة البيانات: البروتوكول الكامل، وسجل التنظيف، و <code>to_csv</code> ، ودفاع سريري عن قرارات
80-90	الاستخلاص والختام	عرض حل المدرب، ونقاش الدفاعات، وشرح الواجب المنزلي، ولمحة عن اليوم الرابع

✓ الأهداف التعليمية

بنهاية هذه الجلسة ستكون قادراً على:

1. تحديد كمية البيانات المفقودة في كل عمود باستخدام `isna().sum()` والتعبير عنها كنسبة مئوية من السجلات، وتسمية الأعمدة الحرجة للتحليل.

2. تطبيق قاعدة القرار السريري للقيم المفقودة — حذف الصف عندما تضيع الهوية، والتعويض بالوسيط عندما تكون التحليلة قابلة للاسترجاع — وتوثيق كل تعويض براية تدقيق.
3. كشف السجلات المكررة تكراراً تاماً وإزالتها باستخدام `drop_duplicates()` و `duplicated()`، وشرح الكلفة البوائية للعد المزدوج.
4. توحيد الترميز الفئوي غير المتسق إلى المعيار القياسي M/F باستخدام الدوال الموجهة `str.strip()` و `str.upper()` و `str[0]` و `map`.
5. حساب سياجي توكي للمدى الربيعي باستخدام `quantile()` وتصنيف كل قيمة شاذة إلى مستحيلة أو متطرفة لكنها حقيقية، واختيار الحذف أو القص أو وضع الراية تبعاً لذلك.
6. تسليم مجموعة بيانات نظيفة معتمدة باستخدام `to_csv` مع سجل تنظيف يصمد أمام تدقيق الوزارة.

➤ حقبة أدوات الذكاء الاصطناعي والاعتماديات — حزمة تنظيف البيانات

تستخدم جلسة اليوم مكتبتين فقط — وهما المكتبتان نفسهما اللتان ستحملان طبقة ضبط الجودة بأكملها في مسارات عمل المديرية.

```
1 import pandas as pd
2 import numpy as np
3
4 pd.set_option("display.width", 100)
```

شيفرة 3.1: استيرادات اليوم الثالث.

```
pip install pandas numpy
```

pandas توفر إطار البيانات (**DataFrame**) وكل أفعال التنظيف لهذا اليوم: `isna` و `dropna` و `fillna` و `duplicated` و `drop_duplicates` و `map` و `quantile` و `between`. وبالنسبة إلى مديريةية صحية يعني هذا أن تصديراً مختبرياً كاملاً يُدقق ويُصلح في ثوانٍ، وبالطريقة نفسها كل شهر، بدلاً من ساعات من التحرير اليدوي المعرض للخطأ في جداول البيانات. **Numpy** توفر انتقاءً موجهاً سريعاً؛ ونستخدم اليوم `np.where` لإلحاق راية مراجعة بالنتائج المشبوهة في عملية واحدة. وبالنسبة إلى المديرية هذا هو الفرق بين مراجع يتسلم قائمة عمل قصيرة بالعينات المؤشرة ومراجع يعيد قراءة ستمئة صف بعينه المجردة.

3.1 السياق السريري والإطار النظري (20 دقيقة)

لماذا تكون بيانات الرعاية الصحية ملوثة

لا يعتمد أي مختبر إنتاج تصدير فاسد، ومع ذلك يكاد كل تصدير حقيقي يكون فاسداً في موضع ما. فالأسباب بنيوية لا شخصية. أولاً، النسخ اليدوي: التقني الذي ينقل نتيجة من شاشة المحلل إلى نموذج ورقي سيرتكب، بمعدل موثق جيداً يقارب خطأ ضغطة مفتاح واحدة لكل بضع مئات من الإدخالات، انزياح فاصلة عشرية (هيموغلوبين 9.9 يصبح 99) أو إضافة رقم (غلوكوز 900 يصبح 9000). ثانياً، أعطال الأجهزة والواجهات: قد يؤدي خلل عابر في الطبقة الوسيطة للمحلل إلى إرسال رمز خام من الجهاز، أو قيمة سالبة كعنصر نائب، أو الدفعة نفسها مرتين. ثالثاً، الخلط بين الوحدات: يُبلغ عن الغلوكوز بوحدة ملغ/دل في مستشفى وبوحدة ملي مول/لتر في مستشفى آخر (بينهما قسمة على 18)، فتخلط الملفات المدموجة المقياسين بصمت. رابعاً، دمج الأنظمة القديمة: عندما يُوحَّد سجل ورقي مع نظام LIS قديم ونظام HIS جديد، يجلب كل منها قاموسه الخاص — أحدها رمز الجنس M/F، وآخر رمزه male/female، وثالث ترك الكتبة يكتبون نصاً حراً بمسافات ضالة.

الكلفة السريرية ليست مجردة. فنموذج الذكاء الاصطناعي التشخيصي المدرب على بيانات ملوثة يتعلم التلوث: مريض إثنان مكرر يضاعف الوزن الظاهري لتلك الحالة السريرية؛ وغلوكوز 9000 يمدد مقياس كل ميزة معايرة؛ والسجلات المرمزة f بدل F تختفي بصمت من فحص مجال مرجعي خاص بالإناث. يلخص الإحصائيون ذلك بعبارة قمامة في المدخلات، قمامة في المخرجات؛ أما في ميداننا فليست المخرجات قمامة بل معدل خطأ تشخيصي، لذا نقول: بيانات ملوثة في المدخلات، تشخيص خاطئ في المخرجات.

البيانات المفقودة وقرار الحذف مقابل التعويض

تُكشف الخلية المفقودة بالتعبير `df.isna()`، وتُعد لكل عمود بالتعبير `df.isna().sum()`، ويُعبّر عنها كعيب نسبي بالتعبير `df.isna().mean() * 100`. أما ما يجب فعله حيالها فهو قرار سريري لا تقني:

- **احذف الصف** (`dropna(subset=[...])`) عندما يهدم الحقل المفقود معنى السجل — وفي مقدمة ذلك فقدان `patient_id`، لأن نتيجة لا يمكن ربطها بمرضى لا يمكنها أبداً أن تؤثر في الرعاية ولا يمكن التحقق منها أبداً.
- **عوّض** (`fillna`) عندما يكون السجل غنياً فيما عدا ذلك ويمكن استبدال التحليلة المفقودة ببديل يمكن الدفاع عنه — وعادةً هو وسيط العمود، لأنه حصين ضد القيم الشاذة ذاتها التي لم نزلها بعد. التعويض يحفظ أعمدة المعلومات السبعة الأخرى التي كان الحذف سيبددها. وهو مسوّغ عندما تتناثر الفراغات عشوائياً بين السجلات (الفقدان العشوائي التام، MCAR)، كما هي حال ثغرات النسخ عادة.
- **لا تعوّض** بصمت أبداً. كل خلية معوّضة يجب أن تترك أثراً: عمود راية منطقية وقيد في السجل. فالمدقق — أو الطبيب الذي يطعن في نصيحة نموذج — يجب أن يكون قادراً على فصل القيم المقيسة عن القيم المصنوعة.
- **لا تعوّض** أبداً الفقدان ذا الدلالة. بعض الفراغات بيانات بحد ذاتها: قيمة `ct_value` في فحص RT-PCR فارغة تحديداً لأن الهدف لم يُكتشف. وملؤها بالوسيط يعني اختراع حمل فيروسي لمرضى سلب.

السجلات المكررة

تنشأ التكرارات التامة من الإرسال المزدوج، وإعادة محاولات الواجهات، ودمج الملفات. يعيد `df.duplicated()` سلسلة منطقية تؤثر كل إعادة لصنف سابق (يبقى الظهور الأول `False`)، ولذلك يُعدّ `df.duplicated().sum()` الفائض ويزيله `df.drop_duplicates()`. الكلفة الوبائية لتركها هي العدد المزدوج: عشرون صفّاً مكرراً في تصدير من 600 صف تعني تضخيماً بنسبة 3.3% في كل عدد مرضي وكل تقدير انتشار وكل إحصائية عيب عمل تُحسب من الملف — والنموذج المدرب عليها يرى أولئك المرضى أهم مرتين من أي شخص آخر.

الترميز الفئوي القياسي

لا تكون الفئة صالحة للاستعمال إلا عندما يكون للمفهوم الواحد رمز واحد بالضبط. ومسار النظافة النصية الموجه هو دائماً الحركات الثلاث نفسها، سلسلة على الموصّل `str.strip()`: يزيل المسافات الخفية في البداية والنهاية، و(`str.upper()`) يوحد حالات الأحرف، و(`str[0]`) يبقى الحرف الأول المميّز (MALE وM يصبحان معاً M). أما `map` الختامية على قاموس قياسي لا يحوي سوى F وM فهي شبكة الأمان: أي رمز غير موجود في القاموس يصبح NaN بدلاً من أن يُخمن، وهكذا تطفو الرموز المجهولة في تدقيق الفقدان التالي بدلاً من الاختباء داخل فئة. قارن دائماً (`value_counts()`) قبل العملية وبعدها — يجب أن تتطابق المجاميع، وإلا فقد أسقط التحويل مرضى بصمت.

القيم الشاذة وسياج المدى الربيعي

القيمة الشاذة (Outlier) قيمة بعيدة عن كتلة توزيعها. وطريقة سياج توكي تُكمّم البعد بطريقة حصينة: مع Q_1 من $quantile(0.25)$ و Q_3 من $quantile(0.75)$ ، يكون المدى الربيعي (IQR) هو $IQR = Q_3 - Q_1$ ، والسيجان هما

$$lower = Q_1 - 1.5 \times IQR, \quad upper = Q_3 + 1.5 \times IQR.$$

لا تكاد الربيعيات تتحرك عندما تدخل قيمة 9000 إلى العمود — على خلاف الوسط الحسابي والانحراف المعياري — ولذلك يبقى السياجان صادقين حتى في البيانات الملوثة. غير أن السياج يكشف فقط؛ ولا يقرر. فالطب يميز نوعين من القيم الشاذة: القيمة المستحيلة (عمر 250، كرياتينين سالب)، وهي خطأ تسجيل يجب إزالته أو تصحيحه عند المصدر، والقيمة المتطرفة لكنها حقيقية (غلوكوز 500 ملغ/دل في الحماض الكيتوني السكري)، وهي بالضبط المريض الذي وُجد النظام التشخيصي لالتقاطه ويجب الإبقاء عليها، مع وضع راية مراجعة عليها في أقصى الحالات. وتنتج عن ذلك ثلاث استراتيجيات: الحذف (لا يُسوَّغ إلا للاستحالة)، والقص إلى حد معين (نادراً ما يُسوَّغ في الطب لأنه يلفّق قيمة)، ووضع الـ **الراية** (الخيار الافتراضي للمتطرفة لكنها حقيقية).

★ لؤلؤة سريرية — عتبات الغلوكوز وحدود الحياة

غلوكوز البلازما الصائم عند البالغين: 70–99 ملغ/دل طبيعي، و125–100 ملغ/دل خلل في غلوكوز الصيام، و ≥ 126 ملغ/دل في مناسبتين تشخيص لداء السكري (معايير ADA)؛ والغلوكوز العشوائي ≥ 200 ملغ/دل مع أعراض تشخيصي أيضاً، وكذلك السكر التراكمي $\geq 6.5\%$ HbA1c. وفي حالات عدم المعاوضة يرتفع السقف ارتفاعاً هائلاً: الحمض الكيتوني السكري يتظاهر نمطياً عند 800–350 ملغ/دل، وحالة فرط الأسمولية وفرط سكر الدم قد تتجاوز 1000 ملغ/دل. أما القيم التي تتعدى نحو 2500 ملغ/دل فهي غير متوافقة مع البقاء على قيد الحياة ومكانها تقارير الحالات. وعليه، فإن غلوكوزاً مسجلاً مقداره 193 ملغ/دل هو مريض؛ وغلوكوزاً مسجلاً مقداره 9000 ملغ/دل هو لوحة مفاتيح. ويجب أن يعامل بروتوكول التنظيف هذين الصنفين معاملتين مختلفتين، وهذا ما سيفعله اليوم.

يلخص الجدول 3.1 النظرية في مصفوفة القرارات التي سننفذها في التطبيق الموجه، ويسرد الجدول 3.2 بوابات المعقولية الفسيولوجية التي سنفرضها باستخدام `between()` — وهي حدود متممة السخاء لا ترفض إلا ما لا يستطيع أي كائن بشري إنتاجه، حتى لا يُستبعد أبداً أي مريض متطرف لكنه حقيقي بفعل التدقيق.

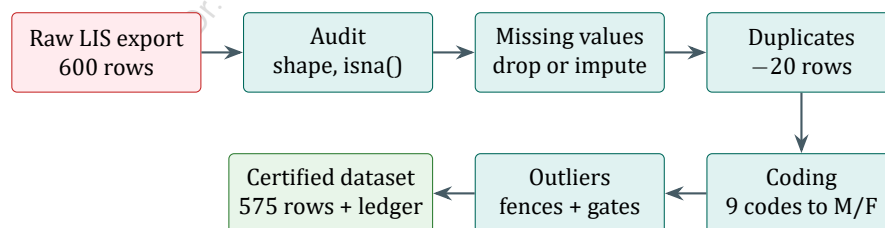
جدول 3.1: مصفوفة قرارات التنظيف لليوم الثالث.

العيب	اكتشفه باستخدام	الإجراء الافتراضي	التحوط السريري
معرف مفقود	<code>isna()</code> على عمود المفتاح	حذف الصف	الهوية لا تُعوّض أبداً
تحليلة مفقودة	<code>isna().sum()</code> وعبء النسبة المئوية	تعويض بالوسيط عبر <code>fillna()</code> مع راية	لا تُعوّض أبداً الفراغات ذات الدلالة
تكرارات تامة	<code>sum().duplicated()</code>	<code>drop_duplicates()</code>	المريض نفسه في زيارتين ليس تكراراً
رموز غير متسقة	<code>value_counts()</code>	نظافة <code>str</code> ثم <code>map</code>	الرموز غير المطابقة يجب أن تطفو بوصفها NaN
قيمة مستحيلة	بوابة معقولية <code>between()</code>	حذف الصف (قيد في السجل)	إبلاغ المختبر المصدر بالأخطاء المتكررة
متطرفة لكنها حقيقية	سياجا المدى الربيعي	راية للمراجعة	هذا هو المريض الذي وُجد النظام من أجله

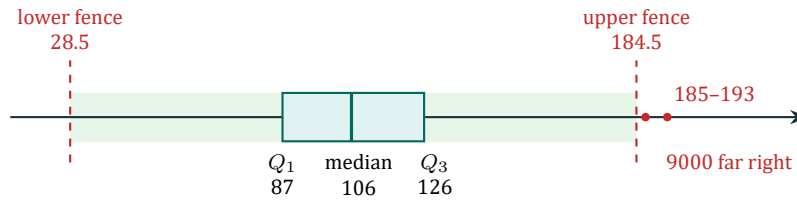
جدول 3.2: بوابات المعقولية الفسيولوجية مقابل المجالات المرجعية للبالغين. البوابات حدود تدقيق، لا مجالات طبيعية.

المتغير	بوابة التدقيق	المرجع للبالغين	مسوّغ البوابة
العمر (سنوات)	0–110	—	حد المعمرين الموثق
الغلوكوز (ملغ/دل)	20–1000	70–99 صائماً	ما دون 20 أو فوق 1000: إعادة فحص
الكرياتينين (ملغ/دل)	0.1–20	ذكور 0.7–1.3؛ إناث 0.6–1.1	القيمة السالبة مستحيلة
الهيموغلوبين (غ/دل)	3–25	ذكور 13–17؛ إناث 12–15	القيمة 99 انزياح فاصلة عشرية
الكريات البيض ($\times 10^9/L$)	0.5–100	4–11	القيمة 900 تتجاوز أي أبيضاض دم

يعرض الشكل 3.1 البروتوكول بوصفه مسار عمل: ست بوابات بين تصدير خام ومجموعة بيانات معتمدة، تترك كل بوابة قيداً في السجل. ويضع الشكل 3.2 السياجين الفعليين اللذين سنحسبهما اليوم على خط أعداد الغلوكوز.



شكل 3.1: مسار التنظيف في اليوم الثالث مطبقاً على `messy_lab_data.csv`. كل بوابة تدون ما غيّره وسبب التغيير؛ وأعداد الصفوف هي الأعداد الحقيقية من جلسة اليوم.



شكل 3.2: سياجا المدى الربيعي للغلوكوز في messy_lab_data.csv (قيم الجلسة الحقيقية): $Q_1 = 87$ و $Q_3 = 126$ و $IQR = 39$ ، والسيجان عند 28.5 و 184.5 ملغ/دل. أربع قيم شاذة (185-193 ملغ/دل) هي فرط سكر دم متطرف لكنه حقيقي؛ أما الخامسة، 9000 ملغ/دل، فمستحيلة.

3.2 التطبيق الموجّه (40 دقيقة)

■ عرض تطبيقي مباشر — اعتماد تصدير مختبري فاسد

نأخذ الملف messy_lab_data.csv وهو 600 سجل مختبري فيه نحو 8% قيم مفقودة، وعشرون إرسالاً مكرراً، وتسع كتابات لجنسين اثنين، وعدة نتائج مستحيلة — عبر المسار الكامل في الشكل 3.1. يكتب المدربون كل خطوة بأنفسهم. والهدف ليس مجرد إطار بيانات نظيف بل إطار خاضع للمساءلة: بحلول الدقيقة 60 يكون كل تصحيح معدوداً ومسوّغاً.

الخطوة 1: حقّل التصدير وانظر إليه مباشرة

لا ننظف أبداً ما لم نرّه. نحمل الملف، ونتأكد من أبعاده، ونتفحص حفنة من الصفوف مع الأعمدة الأهم في عمل اليوم.

```
1 import pandas as pd
2 import numpy as np
3
4 pd.set_option("display.width", 100)
5
6 lab = pd.read_csv("messy_lab_data.csv")
7 print(lab.shape)
8 print(lab[["patient_id", "age", "sex", "glucose_mgdl"]].head(6))
```

شيفرة 3.2: أول تماس مع التصدير الملوّث.

	patient_id	age	sex	glucose_mgdl
(600, 8)				
0	MOH-01132	62.0	M	88.0
1	MOH-00303	50.0	male	85.0
2	MOH-00673	68.0	female	136.0
3	MOH-00622	25.0	male	85.0
4	MOH-00921	10.0	M	142.0
5	MOH-00909	37.0	M	NaN

سته صفوف تكشف عييين بالفعل: عمود sex يخلط M و male و female، والصف 5 بلا غلوكوز. لاحظ أيضاً أن age يُطبع 62.0 لا 62: أُجبرت pandas على خزن العمود كفاصلة عائمة لأن أعمدة الأعداد الصحيحة لا تستطيع حمل NaN — فالفاصلة العشرية نفسها عرّض من أعراض فقدان بيانات في المنبع.

الخطوة 2: إحصاء القيم المفقودة

نعدّ الفراغات في كل عمود ونحول الأعداد إلى نسب مئوية؛ فالتعبير isna() يعطي قيماً منطقية، ومتوسط القيم المنطقية هو بالضبط كسر قيم True.

```
1 print(lab.isna().sum())
```

```

2
3 pct_missing = (lab.isna().mean() * 100).round(1)
4 print(pct_missing)

```

شيفرة 3.3: عدد القيم المفقودة ووزنها.

```

patient_id      0
age             52
sex             0
glucose_mgdl    55
hgb             51
creatinine_mgdl 47
wbc             35
sample_date     0
dtype: int64
patient_id      0.0
age             8.7
sex             0.0
glucose_mgdl    9.2
hgb             8.5
creatinine_mgdl 7.8
wbc             5.8
sample_date     0.0
dtype: float64

```

يقع العبء بين 5.8% و9.2% لكل تحليلية — وهو معدل نمطي لتصدير منسوخ يدوياً. والأهم أن patient_id و sample_date كاملاً، ولذلك يمكن على الأقل نسبة كل نتيجة في هذا الملف إلى صاحبها وتأريخها.

الخطوة 3: بوابة الهوية

البروتوكول أولاً والحظ ثانياً: قبل أي إصلاح نحذف الصفوف التي فُقد فيها patient_id، لأن نتيجة لا يمكن نسبتها إلى مريض لا يمكن أن تصل أبداً إلى ملفه. في هذا التصدير لا تجد البوابة شيئاً — وهذا يجد ذاته يستحق الطباعة، لأن تصدير الشهر القادم قد لا يكون بهذا اللطف.

```

1 print(lab["patient_id"].isna().sum())
2
3 lab = lab.dropna(subset=["patient_id"])
4 print(lab.shape)

```

شيفرة 3.4: حذف السجلات التي فقدت هويتها.

```

0
(600, 8)

```

الخطوة 4: التعويض بوسيط الغلوكوز — مع أثر ورقي

الغلوكوز هو تحليلتنا الحرجة للتحليل و9.2% منه مفقود. حذف 55 سجلاً كاملاً فيما عدا ذلك سيبدد نحو أربعمئة قياس صالح في أعمدها الأخرى، لذلك نعوّض — لكن بالوسيط، ولا نعوّض بصمت أبداً: عمود راية تدقيق يسجل بدقة أي القيم مصنوعة.

```

1 glucose_median = lab["glucose_mgdl"].median()
2 print(glucose_median)
3
4 lab["glucose_imputed"] = lab["glucose_mgdl"].isna()
5 lab["glucose_mgdl"] = lab["glucose_mgdl"].fillna(glucose_median)

```

```

6
7 print(lab["glucose_imputed"].sum())
8 print(lab["glucose_mgdl"].isna().sum())

```

شيفرة 3.5: التعويض بالوسيط مع راية تدقيق.

```

106.0
55
0

```

لماذا الوسيط لا الوسط الحسابي؟ ما زال العمود الخام يحوي خطأ 9000 ملغ/دل، والوسط الحسابي للعمود مسحوب صعوداً إلى 123.6 ملغ/دل بفعل تلك الخلية الواحدة — فلو استخدمناه لحققنا الخطأ في 55 سجلاً بريئاً. أما الوسيط، 106 ملغ/دل، فيتجاهله تماماً. هذه هي حجة الحصانة من الكتلة النظرية وقد صارت ملموسة.

✖ خطأ شائع — التعويض الصامت تلفيق للبيانات

القيمة المعوّضة التي لا تترك أثراً لا يمكن تمييزها من قياس حقيقي — لا للطبيب، ولا للمدقق، ولا للنموذج. أنشئ دائماً عمود الـ `fillna` قبل استدعاء `rtPCR_results.csv` تكون `ct_value` فارغة تحديداً عندما تكون النتيجة سلبية — وملؤها بالوسيط يعني إسناد حمل فيروسي إلى مريض خالٍ من الفيروس. ستثبت ذلك بنفسك في واجب الليلة.

الخطوة 5: التكرارات التامة

تزرع إعادة محاولات الواجهات وإعادة الرفع نسخاً طبق الأصل في التصدير. يؤشر `duplicated()` كل إعادة لصف سابق؛ ويبقى `drop_duplicates()` الظهور الأول لكل صف.

```

1 print(lab.duplicated().sum())
2
3 lab = lab.drop_duplicates()
4 print(lab.shape)

```

شيفرة 3.6: عدّ التكرارات التامة وإزالتها.

```

20
(580, 9)

```

عشرون صفّاً — أي تضخيم بنسبة 3.3% في كل إحصائية كان هذا الملف سيغذيها. لاحظ أن الإطار صار الآن بتسعة أعمدة: رايتنا `glucose_imputed` تسافر مع البيانات من هنا فصاعداً. وكنا دقيقاً في معنى التكرار هنا: المريض نفسه إذا سُحبت له عينة فعلاً مرتين في تاريخين مختلفين فهذه بيانات ترصد لا خطأ؛ لا تُزال إلا الصفوف المتطابقة بايتاً ببايت.

الخطوة 6: جنس واحد، رمز واحد

تدقيق `value_counts()` قبل الإصلاح إلزامي — فهو يظهر حجم فوضى الترميز ويعطينا المجموع الذي يجب أن نحافظ عليه.

```

1 print(lab["sex"].value_counts(dropna=False))

```

شيفرة 3.7: تدقيق عمود الجنس قبل الإصلاح.

```

sex
M      88
M      75
male   73
M      64

```

```
female    59
female    58
F         56
F         55
f         52
Name: count, dtype: int64
```

تسعة رموز لجنسين اثنين — وسطاران من سطور M يبدوان متطابقين لأن الفرق بينهما مسافة خفية في البداية أو النهاية. لهذا يأتي `sex.strip()` أولاً. نسلسل مسار النظافة ذا الحركات الثلاث ونختتم بقاموس `map` قياسي.

```
1 sex_clean = lab["sex"].str.strip().str.upper().str[0]
2 lab["sex"] = sex_clean.map({"M": "M", "F": "F"})
3
4 print(lab["sex"].value_counts(dropna=False))
```

شيفرة 3.8: التوحيد الموجه إلى الترميز القياسي للجنس.

```
sex
M    300
F    280
Name: count, dtype: int64
```

المجاميع محفوظة: $580 = 300 + 280$ ، وهو بالضبط عدد الصفوف بعد إزالة التكرارات، ولا يظهر أي صف NaN — إذن تعرّف القاموس على كل رمز أصلي. ولو وُجد رمز شارد مثل U لحوّله `map` إلى NaN، فيطوفو في التدقيق بدلاً من أن يلوّث فئة.

الخطوة 7: سياجا المدى الربيعي للغلوكوز

الآن مطاردة القيم الشاذة، تماماً كما في الشكل 3.2: الربعيات من `quantile()`، والسيجان عند $1.5 \times IQR$ ، وقناع منطقي (Boolean Mask) لعدّ ما يقع خارجهما وتفحصه.

```
1 q1 = lab["glucose_mgdl"].quantile(0.25)
2 q3 = lab["glucose_mgdl"].quantile(0.75)
3 iqr = q3 - q1
4 lower_fence = q1 - 1.5 * iqr
5 upper_fence = q3 + 1.5 * iqr
6 print("Q1 =", round(q1, 1), " Q3 =", round(q3, 1),
7       " IQR =", round(iqr, 1))
8 print("fences:", round(lower_fence, 1), "to", round(upper_fence, 1))
9
10 is_outlier = ~lab["glucose_mgdl"].between(lower_fence, upper_fence)
11 print(is_outlier.sum())
12 cols = ["patient_id", "age", "sex", "glucose_mgdl"]
13 print(lab.loc[is_outlier, cols].sort_values("glucose_mgdl"))
```

شيفرة 3.9: حساب سياجي توكي بمعالجة موجهة.

```
Q1 = 87.0  Q3 = 126.0  IQR = 39.0
fences: 28.5 to 184.5
5
   patient_id  age  sex  glucose_mgdl
165 MOH-00730  48.0   F         185.0
54  MOH-00565   NaN   F         187.0
563 MOH-00460  91.0   F         188.0
374 MOH-00369  73.0   M         193.0
318 MOH-00066  48.0   F        9000.0
```

هنا تثبت النظرية جدارتها. أربع من القيم الشاذة الخمس تقع عند 185-193 ملغ/دل: فوق السياج، نعم، لكنها في صميم الواقع السكري — مريض سكري بعد وجبة أو سيئ الضبط. إنها متطرفة لكنها حقيقية ويجب أن تبقى. أما الخامسة، 9000 ملغ/دل، فهي القيمة المستحيلة من قصتنا الافتتاحية. وبدلاً من حذف أي شيء الآن، نلحق راية مراجعة باستخدام np.where — استراتيجية الـراية لا الحذف في سطر موجه واحد.

```
1 flag = np.where(is_outlier, "review", "ok")
2 print(pd.Series(flag).value_counts())
```

شيفرة 3.10: وضع راية على القيم الخارجة عن السياجين للمراجعة البشرية.

```
ok      575
review    5
Name: count, dtype: int64
```

الخطوة 8: بوابات المعقولة الفسيولوجية

الحذف محجوز للمستحيل، والمستحيل تعرّفه الفسيولوجيا لا الإحصاء: البوابات السخية في الجدول 3.2. كل بوابة قناع between(): والأعمدة التي ما زالت تحوي قيم NaN صادقة تحصل على شرط إفلات صريح عبر isna() حتى لا يُعَدَم سجل ناقص بجريمة لم يرتكبها.

```
1 age = lab["age"]
2 glu = lab["glucose_mgdl"]
3 cre = lab["creatinine_mgdl"]
4 hgb = lab["hgb"]
5 wbc = lab["wbc"]
6
7 age_ok = age.between(0, 110) | age.isna()
8 glu_ok = glu.between(20, 1000)
9 cre_ok = cre.between(0.1, 20) | cre.isna()
10 hgb_ok = hgb.between(3, 25) | hgb.isna()
11 wbc_ok = wbc.between(0.5, 100) | wbc.isna()
12
13 plausible = age_ok & glu_ok & cre_ok & hgb_ok & wbc_ok
14 print("implausible rows:", (~plausible).sum())
15
16 bad_cols = ["patient_id", "age", "glucose_mgdl",
17            "creatinine_mgdl", "hgb", "wbc"]
18 print(lab.loc[~plausible, bad_cols])
19
20 lab_clean = lab[plausible].copy()
21 print(lab_clean.shape)
```

شيفرة 3.11: الإبقاء على السجلات الممكنة بشرياً فقط.

```
implausible rows: 5
  patient_id  age  glucose_mgdl  creatinine_mgdl  hgb  wbc
113 MOH-01167  58.0          82.0           1.20  10.1  900.0
240 MOH-00238  37.0         105.0           0.91  99.0   7.5
318 MOH-00066  48.0        9000.0           1.13  15.7  14.3
323 MOH-00640  42.0          65.0          -2.00  13.4   5.9
504 MOH-00227 250.0          61.0           1.37  NaN  13.3
```

اقرأ الضحايا الخمس كما يقرؤها الطبيب: كريات بيض 900 تتجاوز أسوأ ابيضاض دم مسجل في التاريخ؛ وهيموغلوبين 99 غ/دل انزراح فاصلة عشرية لقيمة معقولة هي 9.9؛ وغلوكوز 9000 نعرفه؛ وكراتينين 2.00 — ينهك الحساب قبل أن ينهك طب الكلى؛ وعمر 250 لا ينتهي إلى أي مريض في هذه الوزارة. ولاحظ ما أبقت البوابات: قيم الغلوكوز الأربع 185-193 مرت، لأن البوابات بُنيت من الفسيولوجيا لا من السياج. كان الكشف إحصائياً؛ وكان الحكم سريرياً.

✖ خطأ شائع — الدالة `between()` تعامل NaN معاملة False

تعيد `between()` القيمة False عند NaN — ولذلك فإن قناع معقولة بلا شرط إفلات `isna()` يستبعد بصمت كل سجل فيه أي تحليلية مفقودة. في هذا الملف يحذف القناع الصارم 169 صفاف بدلاً من 5: أي أنه كان سيعدم 164 سجلاً حقيقياً لكنه ناقص إلى جانب الخمسة المستحيلة. والفخ نفسه في زي آخر: `dropna()` بلا وسيط `subset` يحذف كل صف يحوي أي NaN — شغل بلا تدبر على الجدول كما هو في هذه الخطوة فيتبخر 165 من صفوفه الـ 580 (وعلى التصدير الخام كان سيتبخر 210 من 600). قرر دائماً لكل عمود على حدة، لا على مستوى الجدول كله.

الخطوة 9: دليل ما قبل التنظيف وما بعده

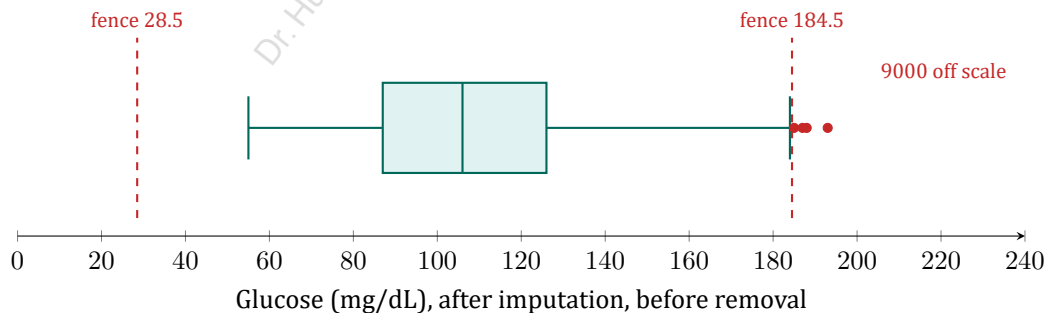
الفعل الأخير في أي بروتوكول تنظيف هو إثبات أنه نجح. نعيد تحميل الملف الخام للمقارنة ونضع جدولي `describe` جنباً إلى جنب، مبقين على أكثر خمسة صفوف تلخيصية دلالة.

```
1 raw = pd.read_csv("messy_lab_data.csv")
2 num = ["age", "glucose_mgdl", "creatinine_mgdl", "hgb", "wbc"]
3 summary_rows = ["count", "mean", "std", "min", "max"]
4 print(raw[num].describe().round(1).loc[summary_rows])
5 print(lab_clean[num].describe().round(1).loc[summary_rows])
```

شيفرة 3.12: مقارنة الملخص قبل التنظيف وبعده.

	age	glucose_mgdl	creatinine_mgdl	hgb	wbc
count	548.0	545.0	553.0	549.0	565.0
mean	45.5	123.6	0.9	13.8	9.2
std	21.9	382.1	0.3	4.0	37.6
min	0.0	55.0	-2.0	7.8	2.3
max	250.0	9000.0	1.5	99.0	900.0
	age	glucose_mgdl	creatinine_mgdl	hgb	wbc
count	524.0	575.0	528.0	525.0	543.0
mean	45.2	107.6	0.9	13.7	7.7
std	20.3	28.7	0.2	1.7	2.6
min	0.0	55.0	0.4	7.8	2.3
max	95.0	193.0	1.5	19.0	20.0

ينهار الانحراف المعياري للغلوكوز من 382.1 إلى 28.7 ملغ/دل وينخفض الوسط الحسابي من 123.6 إلى 107.6 — خمسة صفوف محذوفة وتعويض واحد نزيه كانت تشوه كل إحصائية لاحقة. وكل قيمة عظمى الآن قيمة يستطيع طبيب أن يوقع عليها. جدول ما قبل وما بعد هذا هو الملخص التنفيذي لبروتوكول التنظيف، والشكل 3.3 توأمه البياني.



شكل 3.3: مخطط صندوق للغلوكوز في `messy_lab_data.csv` مبني من ربعيات الجلسة الحقيقية ($Q_1 = 87$ ، الوسيط 106، $Q_3 = 126$ ؛ والشاربان عند 55 و184). أربع قيم شاذة متطرفة لكنها حقيقية تتجمع عند 185-193 ملغ/دل؛ والقيمة المستحيلة 9000 ملغ/دل تقع بعيداً خارج الإطار.

3.3 التحدي السريري (30 دقيقة)

التحدي السريري — اعتماد التصدير لصالح المديرية

انتقل مختبر مستشفى قضاء للتو من نظام LIS عمره خمس عشرة سنة، وأول تصدير موحد له — messy_lab_data.csv — مقرر أن يدخل إحصاءات ترصد السكري في المديرية، ثم يدخل في وقت لاحق من هذا الأسبوع بيانات تدريب نماذجنا التشخيصية. يشترط المدير العام مجموعة بيانات معتمدة: منظمة وفق بروتوكول، ومعها سجل يستطيع مدقق من الوزارة مطابقتها سطرًا بسطر. أنت ضابط جودة البيانات المناوب. اعملوا في ثنائيات:

1. شغّلوا إحصاء القيم المفقودة كاملاً (الأعداد والنسب المئوية) وسَمّوا العمود الحرج للتحليل الأشد تأثراً.
2. افترضوا بوابة الهوية: تحققوا من اكتمال patient_id واحذفوا أي صفوف يتعذر نسبتها إلى مريض.
3. عوّضوا الغلوكوز المفقود بوسيط العمود، منشئين راية التدقيق glucose_imputed أولاً.
4. عدّوا الصفوف المكررة تكراراً تاماً واحذفوها؛ ودوّنوا العدد في السجل.
5. وخذوا sex إلى الترميز القياسي M/F؛ وأثبتوا باستخدام value_counts() أنه لم يُفقد أي سجل ولم يبق أي رمز بلا مطابقة.
6. طبّقوا سياج المدى الربيعي على الغلوكوز وبوابات المعقولية في الجدول 3.2؛ ولا تحذفوا إلا الصفوف المستحيلة، ويَبَيّنوا لماذا تنجو القيم الشاذة المتطرفة لكنها حقيقية.
7. ابنوا سجل التنظيف بوصفه إطار بيانات بالأعمدة issue و count و justification، ثم احفظوا البيانات المعتمدة باسم messy_lab_data_clean.csv باستخدام to_csv مع index=False وتحققوا بإعادة قراءة الملف.
8. حضّروا دفاعاً سريرياً مدته 60 ثانية عن قراراتكم لجلسة الاستخلاص.

✓ حل المدرب — بروتوكول الاعتماد من البداية إلى النهاية

البروتوكول الكامل، بالترتيب الدقيق الذي تدربنا عليه في العرض.

```
1 import pandas as pd
2 import numpy as np
3
4 pd.set_option("display.width", 100)
5
6 raw = pd.read_csv("messy_lab_data.csv")
7 lab = raw.dropna(subset=["patient_id"]) # identity gate
8
9 glucose_median = lab["glucose_mgdl"].median() # 106.0 mg/dL
10 lab["glucose_imputed"] = lab["glucose_mgdl"].isna()
11 lab["glucose_mgdl"] = lab["glucose_mgdl"].fillna(glucose_median)
12
13 n_dupes = lab.duplicated().sum()
14 lab = lab.drop_duplicates()
15
16 sex_clean = lab["sex"].str.strip().str.upper().str[0]
17 lab["sex"] = sex_clean.map({"M": "M", "F": "F"})
18
19 age = lab["age"]
20 glu = lab["glucose_mgdl"]
21 cre = lab["creatinine_mgdl"]
22 hgb = lab["hgb"]
23 wbc = lab["wbc"]
24 plausible = ((age.between(0, 110) | age.isna())
25              & glu.between(20, 1000))
```

```

26         & (cre.between(0.1, 20) | cre.isna())
27         & (hgb.between(3, 25) | hgb.isna())
28         & (wbc.between(0.5, 100) | wbc.isna()))
29 lab_clean = lab[plausible].copy()
30
31 ledger = pd.DataFrame({
32     "issue": ["missing patient_id", "missing glucose",
33             "exact duplicate rows", "inconsistent sex codes",
34             "implausible values"],
35     "count": [int(row["patient_id"].isna().sum()),
36              int(row["glucose_mgdl"].isna().sum()),
37              int(n_dupes),
38              int((~row["sex"].isin(["M", "F"])).sum()),
39              int((~plausible).sum())],
40     "action": ["drop row", "median impute + flag", "drop_duplicates",
41              "strip/upper/map", "remove row"],
42     "justification": ["identity required", "9.2 pct MCAR",
43                     "double counting", "9 codes, 2 sexes",
44                     "outside human range"],
45 })
46 print(ledger)
47
48 lab_clean.to_csv("messy_lab_data_clean.csv", index=False)
49
50 check = pd.read_csv("messy_lab_data_clean.csv")
51 print("rows:", len(check), " duplicates:", check.duplicated().sum())
52 print("sex codes:", sorted(check["sex"].unique()))
53 g = check["glucose_mgdl"]
54 print("glucose range:", g.min(), "to", g.max())
55 a = check["age"]
56 print("age range:", a.min(), "to", a.max())

```

شيفرة 3.13: التنظيف المعتمد للتصدير المختبري.

	issue	count	action	justification
0	missing patient_id	0	drop row	identity required
1	missing glucose	55	median impute + flag	9.2 pct MCAR
2	exact duplicate rows	20	drop_duplicates	double counting
3	inconsistent sex codes	477	strip/upper/map	9 codes, 2 sexes
4	implausible values	5	remove row	outside human range
rows: 575 duplicates: 0				
sex codes: ['F', 'M']				
glucose range: 55.0 to 193.0				
age range: 0.0 to 95.0				

التفسير السريري والدفاع. يضم الملف المعتمد 575 سجلاً من أصل 600 (استبقاء بنسبة 95.8%)، كل واحد منها قابل للنسبة إلى صاحبه، وفريد، ومرمّز قياسياً، ويمكن بشرياً. الدفاع الأول — التعويض بوسيط الغلوكوز: حذف الصفوف الـ 55 الفارغة الغلوكوز كان سيبدد نحو 400 قياس صالح في أعمدها الأخرى؛ واختير الوسيط (106 ملغ/دل، قيمة صيام طبيعية) بدلاً من الوسط الحسابي (123.6 ملغ/دل) لأن الوسط كان ملوثاً بخطأ 9000 ملغ/دل — كما أن الرأية glucose_imputed تتيح لأي محلل مستقبلي استبعاد القيم المصنوعة بقناع واحد. الدفاع الثاني — الإبقاء على القيم الشاذة 185–193 ملغ/دل: أشرفها سياج المدى الربيعي، لكن 185–193 ملغ/دل هو سكري عادي غير مضبوط — أي بالضبط الفئة السكانية التي وُجد برنامج ترصد السكري للعثور عليها. حذفها كان سيحرف تقدير الانتشار نحو الانخفاض ويعلم نموذجاً مستقبلياً أن فرط سكر الدم الشديد لا يحدث. وأحد الناجين يستحق ملاحظة: سجل العمر 0 يجتاز بوابة (0, 110) between لأن الوليد مريض ممكن. وفي الممارسة يوضع على مثل هذا السجل مؤشر مراجعة يدوية على سجل القبول. كشف إحصائي، ثم حكم سريري — بهذا الترتيب دائماً.

أسئلة النقاش الجماعي.

1. يذكر السجل 477 إصلاحاً لرموز الجنس مقابل صفر سجل مفقود. يجادل زميل بأن الإصلاح كان بالتالي تجميلياً واختيارياً. ما الذي ينكسر لاحقاً — سريرياً وإحصائياً — لو بقيت الرموز التسعة؟
2. الصف MOH-00238 يحمل هيموغلوبين 99 غ/دل، وهو على الأرجح شبه المؤكد انزياح فاصلة عشرية عن 9.9. لقد حذفنا الصف؛ فهل كان ينبغي أن نصححه إلى 9.9 بدلاً من ذلك؟ ومن يملك صلاحية إجراء ذلك التصحيح — المحلل، أم المختبر المصدر؟
3. في الشهر القادم يرسل المختبر نفسه تصديراً جديداً. ما الذي ينبغي أن يتحول من بروتوكول اليوم إلى سكرت شجري تلقائي، وما الذي يجب أن يبقى قراراً بشرياً في كل مرة؟

3.4 الواجب المنزلي والتقييم الذاتي

تمارين الواجب المنزلي.

1. إحصاء القيم المفقودة. على الملفين `chemistry_panel.csv` و `cbc_results.csv` شغل تدقيق فقدان الكامل (الأعداد والنسب المئوية). أبلغ عن العمود الأشد تأثراً في كل ملف، أو صرح بأن الملف كامل.
2. الحصانة بيدك. على الملف الخام `messy_lab_data.csv` احسب كلاً من `mean()` و `median()` للعمود `glucose_mgdl`. اشرح في جملتين، مستشهداً بصف 9000 ملغ/دل، لماذا واحدة فقط من القيمتين قيمة تعويض آمنة.
3. فقدان ذو الدلالة. في الملف `rtPCR_results.csv` عدّ قيم `ct_value` المفقودة وقارنها مع `value_counts()` للعمود `result`. صرح بالنمط الذي تجده، وبنّ في جملة واحدة الضرر السريري للتعويض بوسيط ذلك العمود.
4. سياجان على الدهون. احسب سياجي المدى الربيعي للعمود `triglycerides_mgdl` في `chemistry_panel.csv`، وعدّ القيم الشاذة بقناع، وصنّفها: هل مستوى شحوم ثلاثية مقداره 900 ملغ/دل مستحيل أم متطرف لكنه حقيقي؟ سوّغ سريرياً.
5. من البروتوكول إلى السياسة (اختياري). صُغ إجراء عمل قياسياً من نصف صفحة، بلغة واضحة لمشرفي المختبرات، يصف البوابات الست في الشكل 3.1 وما تدوّنه كل بوابة في السجل.

اختبار التحقق الذاتي.

1. ما التعبير الواحد الذي يعدّ القيم المفقودة في كل عمود من إطار بيانات `df`؟
2. لعمود فيه $Q_1 = 80$ و $Q_3 = 120$ ملغ/دل، احسب المدى الربيعي وسياجي توكي كليهما.
3. مريض طوارئ فاقد الوعي سُجل له غلوكوز 550 ملغ/دل. مستحيل أم متطرف لكنه حقيقي — وماذا يفعل البروتوكول بالصف؟
4. لماذا يُفضّل الوسيط على الوسط الحسابي لتعويض التحليلات المختبرية؟
5. كُتب الإدخال `male` بمسافة ضالة على كل جانب. إلّا تحوّل `str.strip()` ثم `str.upper()` ثم `str[0]` مطبقاً بهذا الترتيب، وماذا تفعل `map` اللاحقة على القاموس القياسي `M/F` برمز لا تتعرف عليه؟

أجوبة الاختبار. (1) `df.isna().sum()`: $IQR = 40$ (2) $80 - 60 = 20$ ملغ/دل، والسياج الأعلى $120 + 60 = 180$ ملغ/دل. (3) متطرف لكنه حقيقي — متوافق مع الحمض الكيتوني السكري أو حالة فرط الأسمولية؛ يُستبقى الصف وتوضع عليه راية مراجعة، ولا يُحذف أبداً. (4) الوسيط حصين: قيمة مستحيلة واحدة (مثل 9000 ملغ/دل) تسحب الوسط الحسابي لكنها تترك الوسيط دون تغيير يذكر، ولذلك لا ينشر التعويض بالوسيط الخطأ. (5) يصبح `M` (إزالة المسافات، ثم الأحرف الكبيرة، ثم الحرف الأول)؛ وتحوّل `map` أي رمز غير معروف إلى `NaN`، فيطفو عمداً في تدقيق فقدان التالي بدلاً من الاختباء داخل فئة.

♣ القيمة لمديرية التشخيص — طبقة بيانات معتمدة لكل تقارير المديرية

يحوّل بروتوكول اليوم تنظيف البيانات من عادة شخصية غير ماثية إلى أداة مؤسسية. وبصورة ملموسة: التصديرات المختبرية الشهرية من أي محافظة يمكن تدقيقها واعتمادها في دقائق بسكرت بدلاً من أيام باليد؛ وكل رقم تنشره المديرية — انتشار، وعبء عمل، ومؤشرات جودة — يقف على سجلات منزوعة التكرار ومفحوصة المعقولة؛ وسجل التنظيف يمنح التدقيق الداخلي ومكتب الوزير حساباً سطرًا بسطر لكل تعديل، وهو الفرق بين ثقب بي ودقق علي؛ ونماذج الذكاء الاصطناعي في الأيام 6-10 ستُدرّب على مدخلات معتمدة، مما يخفف مباشرة خطر الخطأ التشخيصي فيها. كما أن أنماط السجل المتكررة (انزياحات الفاصلة العشرية المزمّنة لدى مختبر ما، وتكرارات واجهة ما) تصبح تغذية راجعة موجهة إلى المختبرات المصدر — فتُصلح جودة البيانات حيث تُنتج، لا حيث تُحلّل فحسب.

خلاصات اليوم

- بيانات الرعاية الصحية ملوثة لأسباب بنيوية — الإدخال اليدوي، وأعطال الأجهزة، والخلط بين الوحدات، ودمج الأنظمة القديمة — ولذلك فالتنظيف بروتوكول دائم، لا عملية إنقاذ لمرة واحدة.
- دقق قبل أن تصلح: `shape()` و `sum()` و `isna()` والنسب المئوية و `value_counts()` تخبرك بما تعالجه قبل أن تعالجه.
- لا تحذف صفاً إلا عندما يضيع معناه (فقدان `patient_id`)؛ وعوض التحلية بالوسيط، ولا نعوض أبداً بلا عمود راية وقيد في السجل.
- التكرارات التامة مرضى معدودون مرتين: `sum()` و `duplicated()` ثم `drop_duplicates()`.
- مفهوم واحد، رمز واحد: `str.strip()` ثم `str.upper()` ثم `str[0]` ثم `map` مع حفظ مجاميع `value_counts()` قبل العملية وبعدها.
- سياجا المدى الربيعي ($Q_3 + 1.5 \times IQR$ و $Q_1 - 1.5 \times IQR$) يكشفان القيم الشاذة بحصانة، لكن الحكم سريري: احذف المستحيل، وضع راية على المتطرف لكنه حقيقي، ولا تقص إلا فيما ندر.
- سجل تنظيف مع ملف معتمد محفوظ يحولان عملك إلى دليل: 600 صف خام صارت اليوم 575 صفاً خاضعاً للمساءلة.
- غداً، في الفصل 4، نضع البيانات المعتمدة موضع العمل: الفترة المنطقية المتقدمة وتحليلات العمليات السريرية — العثور على كل مدخن مصاب بالسكري تجاوز الستين في البصرة، بسطر موجه واحد في كل مرة.

باب 4

اليوم 4 — الفترة المنطقية المتقدمة وتحليلات العمليات السريية

في كل صباح، في مكان ما من مستشفيات المديرية، يُطرح على طبيب مقيم سؤال من قبيل: كم عدد مرضانا الراقدين ممن تجاوزوا الستين ويعانون سكرياً غير منضبط وارتفاعاً في كولسترول LDL ؟ الإجابة التقليدية تكلف عسراً كاملاً من التمرير عبر جداول البيانات، وحين تصل يكون المدير قد انتقل إلى السؤال التالي. اليوم نحول ذلك السؤال إلى سطر بايثون واحد. فالقناع المنطقي (Boolean Mask) — وهو عمود من قيم True و False ينتج عن مقارنة عمود كامل بعتبة سريية — هو الأداة المركزية في هذه الدورة كلها، وعليه يرتكز كل استعمال قادم لاصطياد الحالات وكل وسم قادم لتعلم الآلة. ثم نتدرج من إيجاد المرضى إلى تلخيص الخدمات: التجميع (groupby) للرقود حسب القسم، وحساب معدلات إعادة الرقود، وبناء الجداول التقاطعية لعوامل الخطورة، والدمج (merge) بين سجل البيانات الديموغرافية ولوحة الكيمياء السريية تماماً كما يربط نظام معلومات المختبر العينة بمرضاها طالب الفحص. وبنهاية الجلسة ستُجمع تقرير مؤشرات الأداء الرئيسة KPI لقسم الطوارئ الذي كان، حتى هذا اليوم، يستهلك من الموظف الإحصائي معظم أسبوع.

الدقائق	الجزء	المحتوى والأنشطة
00-05	مراجعة اليوم 3	البيانات التنظيمية من الفصل 3: لماذا لا تكون الفلاتر موثوقة إلا بعد التنظيف: وعد اليوم: صفر جمل شرطية
05-15	النظرية الأولى	الأقنعة المنطقية: معاملات المقارنة على أعمدة كاملة: الجمع بين المعايير بمعاملات AND و OR و NOT والأقواس الإلزامية: isin و between و str.contains
15-20	النظرية الثانية	التقسيم-التطبيق-الجمع وتشبيه حامل الأنابيب: agg: الفرق بين value_counts و groupby.size و crosstab و pivot_table: الدمج الداخلي مقابل الأيسر في merge: pd.to_datetime و dt
20-30	التطبيق الموجّه الأول	تحميل patients_demographics.csv و chemistry_panel.csv الدمج على patient_id: فحص الجدول المدمج
30-40	التطبيق الموجّه الثاني	اصطياد الحالات متعدد المعايير في سطر مقنّع واحد: isin على المحافظات: المدخنون ذوو الدهون الثلاثية المرتفعة: نطاق ما قبل السكري بالدالة between
40-50	التطبيق الموجّه الثالث	متوسط مدة الرقود بالدالة groupby: agg بالمتوسط والوسيط والعدد: معدل إعادة الرقود بوصفه متوسط عمود منطقي
50-60	التطبيق الموجّه الرابع	crosstab للجنس مع التدخين: pivot_table لقيم HbA1c حسب المحافظة والجنس: تحليل تواريخ الرقود: الرقود شهرياً
60-80	التحدي	يبي المتدربون تقرير مؤشرات الأداء الرئيسة لقسم الطوارئ المطلوب من المدير على الجدول المدمج
80-88	الحل	عرض المدرب خطوة بخطوة: الفئات العمرية بالدالة pd.cut، وترتيب معدلات إعادة الرقود، وجدول KPI الموحد، والتفسير السريي
88-90	الختام	الخلاصات الرئيسة، وشرح الواجب المنزلي، وتشويق لليوم 5 والتمثيل البصري

✓ الأهداف التعليمية

بنهاية هذه الجلسة ستكون قادراً على:

1. بناء أقنعة منطقية على أعمدة كاملة والجمع بين ثلاثة معايير سريرية أو أكثر باستخدام & و | و ~ مع الأقواس الصحيحة.
2. اختيار مجموعات الفئات بالدالة `isin`، والمديات العددية بالدالة `between`، والأنماط النصية بالدالة `str.contains`، مع بيان السؤال السريري الذي تجيب عنه كل منها.
3. دمج سجل البيانات الديموغرافية مع جدول مختبري على `patient_id` والتنبيؤ، قبل تشغيل الشيفرة، بالمرضى الذين يبقون بعد الدمج الداخلي مقابل الدمج الأسير.
4. تلخيص أي مؤشر حسب القسم أو المحافظة أو الفئة العمرية بالدالتين `agg` و `groupby`، وشرح مبدأ التقسيم-التطبيق-الجمع في جملة واحدة.
5. حساب معدل (إعادة الرقود، أو الإيجابية، أو الالتزام) بوصفه متوسط عمود منطقي — دون حلقات ودون جمل شرطية.
6. إنتاج الجداول التقاطعية بالدالتين `pd.crosstab` و `pivot_table`، وعدّ الرقود الشهرية بالدالة `pd.to_datetime` والموصّل `..dt`.
7. تجميع إطار بيانات KPI مضغوط يجيب عن خمسة أسئلة على مستوى المديرية، والدفاع عن نتائجه بلغة سريرية واضحة.

➔ حقبة أدوات الذكاء الاصطناعي والاعتماديات — مكتبات اليوم

يعمل اليوم بأكمله على مكتبة `pandas`؛ وتأتي معها `NumPy` فتعمل بصمت تحت كل قناع.

```
1 import pandas as pd
2
3 pd.set_option("display.width", 100)
```

شيفرة 4.1: استيرادات اليوم 4

```
pip install pandas
```

pandas هي محرك إطار البيانات (DataFrame) الذي نستخدمه منذ اليوم 2؛ واليوم نشغل أدواتها التحليلية الأربع — الفهرسة المنطقية، و `groupby`، و `merge`، و `pivot_table` — التي تحل مجتمعةً محل الفترة اليدوية في إكسل وصيغ `VLOOKUP` والجداول المحورية التي تستهلك ساعات من وقت الموظفين في كل شعبة إحصاء. أما **NumPy** (التي تستوردها `pandas` داخلياً) فتتفقد المقارنات في شفرة C مُجمّعة، ولهذا يعود قناع على 1,200 مريض — أو 1.2 مليون — خلال أجزاء من الألف من الثانية، ولا يخطئ أبداً في نسخ صف كما قد يحدث في الفترة اليدوية.

4.1 السياق السريري والإطار النظري (20 دقيقة)

خلف لنا الفصل 3 جداول نظيفة موثوقة — لكن البيانات النظيفة لا تجيب عن أي سؤال بذاتها. فأسئلة المديرية انتقائية (أي المرضى يستوفون هذه المعايير؟)، ومقارنة (أي قسم يُبقي المرضى أطول مدة؟)، وعلائقية (ماذا تقول البيانات الديموغرافية عن النتائج المختبرية؟). وكل نوع منها يقابل أداة واحدة بالضبط في `pandas`، ولا يحتاج أي منها إلى حلقة تكرارية.

🔍 القناع المنطقي

عندما تقارن عموداً كاملاً بعدد، لا تزور `pandas` كل صف على حدة: فالتعبير `merged["age"] > 60` يُقيّم لجميع المرضى الـ 1,200 دفعة واحدة ويعيد سلسلة (Series) بالطول نفسه من قيم `True/False` — أي قناعاً. ووضعهُ داخل `merged.loc[mask, cols]` يُبقي الصفوف ذات القيمة `True` فقط، تماماً كما يحتجز المنخل الجسيمات الأكبر من ثقوبه وحدها. وتترتب على ذلك نتيجتان مهمتان سريرياً. الأولى أن القناع نفسه بيانات: `mask.sum()` يعدّ الحالات (`True` تُحسب 1) و `mask.mean()` هو انتشارها. والثانية أن القناع يوثق تعريف الحالة في صيغة قابلة للتنفيذ — فالمدقق يقرأ العتبة مباشرة من

الشيفرة، وهو ما لا يتيح أي جدول بيانات مُفلتر يدوياً.

الجمع بين المعايير: AND و OR و NOT

تعريفات الحالات السريرية تجمع بين معايير: مسنّ ومصاب بالسكري ومختل شحوم الدم معاً. المعاملات الجامعة في pandas هي & (و)، | (أو)، ~ (نفي) — وليست الكلمات الإنكليزية أبداً، فهي تفشل على الأعمدة الكاملة لأن بايثون لا يستطيع اختزال 1,200 قيمة صحيحة إلى قيمة واحدة. وثمة قاعدة لا تفاوض عليها: كل مقارنة تحصل على قوسها الخاصين، لأن & و | أشد ارتباطاً من > و <، فيُعاد تفسير الصيغة غير المقوّسة بترتيب خاطئ. اكتب `(merged["hba1c_pct"] >= 6.4) & (merged["age"] > 60)` (6.5) فيقرأ التعبير كما يُقرأ تعريف الحالة الذي يرمّزه.

فلتر المجموعات والمديات والنصوص

ثلاث إضافات تكمل مفردات الفلتر. `isin` يختبر الانتماء إلى مجموعة فئات — فالتعبير `merged["governorate"].isin(["Basra", "Najaf", "Karbala"])` يجد جميع رقود الإقليم الجنوبي في نداء واحد، مستبدلاً ثلاثة شروط OR متسلسلة. و `between(low, high)` يختبر مجالاً مغلقاً، وهو مثالي للمجالات المرجعية: فناطق ما قبل السكري هو بالضبط `merged["hba1c_pct"].between(5.7, 6.4)`. أما `str.contains` فيبحث في النص الحر: `merged["department"].str.contains("Medicine")` يطابق كل اسم قسم يحتوي تلك الكلمة — وهي الأداة نفسها التي ستجد لاحقاً ذات الرئة داخل تشخيصات الخروج غير المهيكل، حيث تفشل المساواة التامة.

الجدول 4.1 هو قاموس الترجمة لهذا اليوم: وينبغي للمتدربين أن يتحركوا على طول أي صف منه في الاتجاهين كليهما.

جدول 4.1: من السؤال السريري إلى تعبير pandas

السؤال السريري	الأداة	صيغة التعبير
الغلوكوز 126 فأكثر؟	المقارنة	<code>df["glucose_mgd1"] >= 126</code>
سكري ومسّن معاً؟	&	<code>(m1) & (m2)</code>
العناية المركزة أو الطوارئ؟	isin	<code>df["department"].isin([...])</code>
غير مدخن؟	~	<code>~(df["smoker"] == "Yes")</code>
HbA1c ضمن 5.7–6.4؟	between	<code>df["hba1c_pct"].between(5.7, 6.4)</code>
اسم القسم يذكر Medicine؟	str.contains	<code>df["department"].str.contains("Medicine")</code>
كم عدد مستوفي التعريف؟	sum	<code>mask.sum()</code>
ما نسبة من يستوفونه؟	mean	<code>mask.mean()</code>

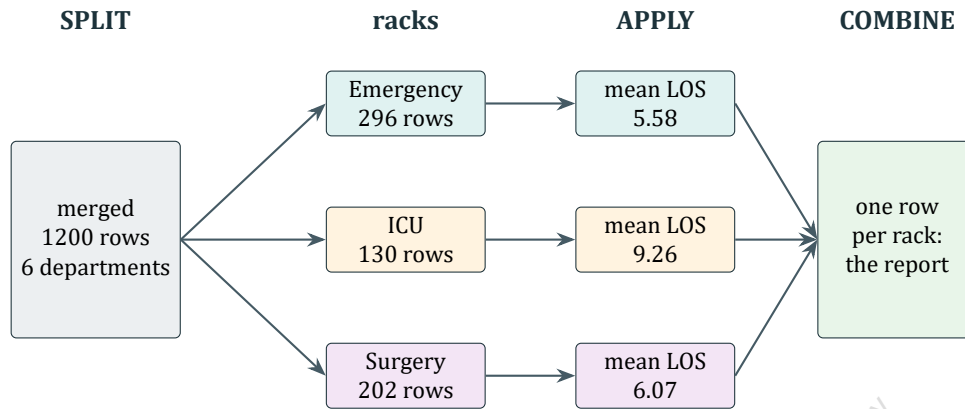
★ لؤلؤة سريرية — العتبات المستخدمة اليوم

جميع عتبات اليوم معايير قياسية للبالغين. الحالة السكرية (وفق الجمعية الأمريكية للسكري ADA): غلوكوز صائم ≥ 126 ملغ/دل أو سكر تراكمي $\geq 6.5\%$ HbA1c يعرف السكري؛ و $5.7\% - 6.4\%$ HbA1c بين $5.7\% - 6.4\%$ يعرف ما قبل السكري؛ ودون 5.7% طبيعي. الشحوم (وفق NCEP ATP III): كولسترول LDL بقيمة ≥ 160 ملغ/دل مرتفع ($130 - 159$ حدّي الارتفاع)؛ والدهون الثلاثية ≥ 200 ملغ/دل مرتفعة ($150 - 199$ حدّي). وتشغيلياً، فإن معدل إعادة الرقود خلال 30 يوماً في قسم ما إذا بقي فوق المتوسط المؤسسي باستمرار (نحو $12\% - 13\%$ في سجلنا) استدعى مراجعة للملفات. والفلتر لا يكون قابلاً للدفاع عنه إلا بقدر عتبهته — فاذكر الدليل الإرشادي إلى جانب القناع في كل نص تحليلي.

التقسيم-التطبيق-الجمع: تشبيه حامل الأنابيب

تخيّل 1,200 أنبوب دم تصل في صندوق واحد، كل منها موسوم بقسمه. لا يحلل أي تقني الصندوق كتلة واحدة. أولاً تقسم: تفرز الأنابيب في ستة حوامل، واحد لكل قسم. ثم تطبق: تجري القياس نفسه على كل حامل — وليكن متوسط مدة الرقود. وأخيراً تجمع: تكتب سطرًا واحداً لكل حامل في التقرير. هذا بالضبط ما تفعله `agg` والدالة `merged.groupby("department")["length_of_stay_days"].mean()` يرسمه. والشكل 4.1 يرسمه. `agg(["mean", "median", "count"])` يجري ثلاثة قياسات لكل حامل في تمريرة تعمّم خطوة التطبيق — فالنداء

واحدة. وللعَدَّ الخالص، يترتب `value_counts()` فئات عمود واحد حسب التكرار، بينما يعدّ `groupby(...).size()` الصفوف لكل مجموعة بترتيب المجموعات.



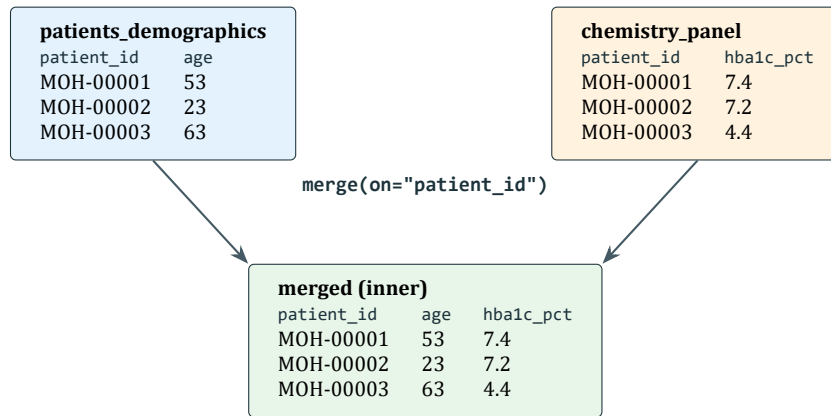
شكل 4.1: التقسيم-التطبيق-الجمع بتشبيه حامل الأنابيب: تُقسَّم السجلات إلى حوامل حسب القسم، ويُطبَّق إحصاء واحد على كل حامل، ثم تُجمع النتائج في تقرير واحد (ثلاثة من ستة حوامل معروضة: القيم الحقيقية).

الجدول التقاطعي والجدول المحوري

أداتان تحولان المجموعات إلى جداول ثنائية البعد. فالنداء `pd.crosstab(df["sex"], df["smoker"])` يعدّ المرضى في كل توليفة من متغيرين فئويين — وهو جدول التوافق الذي يقوم عليه كل اختبار مربع كاي قرأته يوماً في مجلة علمية. و `df.pivot_table(values="hba1c_pct", index="governorate", columns="sex", aggfunc="mean")` يعدّ pivot_table شقيقه العددي: فالنداء `df.pivot_table(values="hba1c_pct", index="governorate", columns="sex", aggfunc="mean")` يضع في كل خلية إحصاء تلخيصياً لا عدداً. إنه الجدول المحوري الذي تعرفه أصلاً في إكسل — لكنه قابل لإعادة الإنتاج في الشهر المقبل على بيانات جديدة، ومحصّن ضد سحبة فأرة شاردة.

ربط الجداول بالدالة merge

لا تخزن المديرية كل شيء في ملف واحد أبداً: فالبيانات الديموغرافية تسكن سجل الرقود، والكيمياء في نظام المختبر. والدالة `merge` تربطهما على المفتاح المشترك: `patients.merge(chem, on="patient_id", how="inner")`. والوسيط `how` يقرر أي المرضى يبقى. فالدمج الداخلي (`inner`) يُبقي فقط المرضى الموجودين في كلا الجدولين — الفوج الجاهز للتحليل ببيانات كاملة. والدمج الأيسر (`left`) يُبقي كل مريض من الجدول الأيسر (الديموغرافي) ويملأ القيم المختبرية الغائبة بالقيمة `NaN` — منظور التدقيق الذي يكشف من رقد ولم يُفحص قط. والاختيار سريري لا تقني: إسقاط غير المفحوصين صائب تماماً لتحليل خطورة كيميائي حيوي، وخاطئ تماماً لتدقيق تغطية الفحوصات. والشكل 4.2 يعرض الآلية.



inner: only IDs in both survive; left: all patients, untested get NaN

شكل 4.2: دمج سجل الرقود مع لوحة الكيمياء السريرية: يُطابق كل صف ديموغرافي مع الصف المختبري الذي يحمل المعرف نفسه.

التواريخ بوصفها بيانات من الدرجة الأولى

التاريخ المخزن نصاً يمكن طباعته لكن لا يمكن استجوابه. بعد النداء `pd.to_datetime(merged["admission_date"])` يفهم العمود التقويم، ويكشف الموصل `dt`. أجزاء: `dt.month`، `dt.year`، و `dt.day_name()`. وبالاقتراح مع `value_counts().sort_index()` يعطي سطر واحد عدد الرقود شهرياً — وهو العمود الفقري للترصد الموسمي. ولاحظ فرق الترتيب: `sort_index()` يرتب حسب الشهر التقويمي (سلسلة زمنية)، بينما `value_counts()` الافتراضي يرتب حسب التكرار (جداول ترتيبية).

✖ خطأ شائع — البتر الصامت في الدمج الداخلي

لا يبلغ الدمج الداخلي عن أي خطأ حين يُسقط مريض — إنه ببساطة يعيد صفوفاً أقل. في تجربة تمهيدية اقتصرت على أول 100 سجل كيمياء، أعاد الدمج الداخلي لسجلنا ذي الـ 1,200 مريض 100 صف فقط: اختفى 1,100 مريض راقد دون أي رسالة. اجعلها عادة انعكاسية أن تطبع `shape`. قبل كل دمج وبعده: وإذا كان عدد الصفوف يجب ألا يتقلص، فاستخدم `how="left"` وعُدّ قيم NaN بدلاً من ذلك — فأولئك المرضى غير المفحوصين قد يكونون هم النتيجة الحقيقية.

4.2 التطبيق الموجّه (40 دقيقة)

■ عرض تطبيقي مباشر — من ملفين إلى إحاطة للمديرية

الهدف: انطلاقاً من `patients_demographics.csv` و `chemistry_panel.csv`، بناء جدول فوج مدمج واحد والإجابة عن تسعة أسئلة متصاعدة — من قائمة حالات بثلاثة معايير إلى أعداد الرقود الشهرية — دون حلقة واحدة أو جملة شرطية واحدة. يكتب المتدربون مع المدرب: وتوقع كل ناتج قبل تشغيل الخلية.

الخطوة 1: تحميل الجدولين والدمج على `patient_id`

يحمل سجل الرقود هوية المريض؛ وتحمل لوحة الكيمياء ما قاسه المختبر. والمفتاح المشترك `patient_id` يتيح للدالة `merge` محاذاتهم صفاً بصف: 9 أعمدة ديموغرافية زائد 13 عمود كيمياء، مع احتساب المفتاح مرة واحدة، تعطي 21 عموداً.

```

1 import pandas as pd
2
3 pd.set_option("display.width", 100)
4
5 patients = pd.read_csv("patients_demographics.csv")

```

```

6 chem = pd.read_csv("chemistry_panel.csv")
7 print("patients:", patients.shape)
8 print("chemistry:", chem.shape)
9
10 merged = patients.merge(chem, on="patient_id", how="inner")
11 print("merged: ", merged.shape)
12
13 cols = ["patient_id", "age", "department", "hba1c_pct", "ldl_mgd1"]
14 print(merged[cols].head(3))

```

شيفرة 4.2: دمج البيانات الديموغرافية مع لوحة الكيمياء

```

patients: (1200, 9)
chemistry: (1200, 13)
merged: (1200, 21)
  patient_id  age  department  hba1c_pct  ldl_mgd1
0  MOH-00001   53  Pediatrics      7.4         82
1  MOH-00002   23  Emergency      7.2        153
2  MOH-00003   63  Pediatrics      4.4        147

```

لم يُفقد أي صف: فلكل مريض راقِد سجل كيمياء، ولذلك فالدمج الداخلي آمن هنا، ويصبح merged مصدر الحقيقة الوحيد لجلستنا.

الخطوة 2: اصطيات الحالات متعدد المعايير في سطر مقتنع واحد

سؤال المدير: المرضى فوق الستين ممن لديهم $HbA1c \geq 6.5\%$ (نطاق السكري) وكوليسترول $LDL \geq 160$ ملغ/دل (مرتفع). ثلاث مقارنات مقوّسة ومعاملات &؛ وعلامات الاستمرار بالشرطة المائلة العكسية تُبقي كل معيار في سطره المقروء الخاص.

```

1 high_risk = (merged["age"] > 60) \
2             & (merged["hba1c_pct"] >= 6.5) \
3             & (merged["ldl_mgd1"] >= 160)
4
5 print("Case count:", high_risk.sum())
6 cases = merged.loc[high_risk, ["patient_id", "age", "hba1c_pct", "ldl_mgd1"]]
7 print(cases.head(5))

```

شيفرة 4.3: مرضى مسنون مصابون بالسكري مع ارتفاع LDL

```

Case count: 14
  patient_id  age  hba1c_pct  ldl_mgd1
217  MOH-00218   75      7.0        178
278  MOH-00279   72      7.1        179
310  MOH-00311   72      7.0        164
312  MOH-00313   84      7.0        171
452  MOH-00453   89      7.5        175

```

أربعة عشر مريضاً من أصل 1,200 يستوفون التعريف الكامل — قائمة إحالة قابلة للتنفيذ نحو عيادة الخطورة القلبية الوعائية. ولاحظ أن `high_risk.sum()` عدّ الحالات قبل أن نعرضها أصلاً؛ فالقناع هو الإحصاء.

✖ خطأ شائع — الأقواس المنسية وأداة العطف الخاطئة

خطآن كلاسيكيان يظهران في الدقائق العشر الأولى من كل ورشة. الأول أن `merged["age"] > 60` و `merged["hba1c_pct"] >= 6.5` يفشل لأن & أشد ارتباطاً من >: إذ يحاول بايثون حساب `merged["hba1c_pct"] >= 6.5 & merged["age"] > 60` أولاً. والإصلاح آلي — أقواس حول كل مقارنة. والثاني أن الكلمة الإنكليزية and بين قناعتين ترفع الخطأ `ValueError: The truth value of a Series is ambiguous`، لأن and تطلب جواباً واحداً بنعم أو لا بينما يحمل قناعاك 1,200 جواباً. اقرأ ذلك الخطأ على أنه كنت تقصد &.

الخطوة 3: مجموعات الفئات والمديات والبحث النصي

أربعة فلاتر سريعة، كل منها مفردة مختلفة: `isin` يوفر علينا ثلاث جمل `OR`، ونطاق ما قبل السكري مثال مدرسي على `between`، `str.contains` يجد الأقسام التي يذكر اسمها كلمة `Medicine`.

```
1 south = merged["governorate"].isin(["Basra", "Najaf", "Karbala"])
2 print("Southern admissions:", south.sum())
3
4 smoker_tg = (merged["smoker"] == "Yes") \
5             & (merged["triglycerides_mgdl"] >= 200)
6 print("Smokers with high TG:", smoker_tg.sum())
7
8 prediab = merged["hba1c_pct"].between(5.7, 6.4)
9 print("Pre-diabetic band:", prediab.sum())
10
11 med_depts = merged["department"].str.contains("Medicine")
12 print("'Medicine' departments:", med_depts.sum())
```

شيفرة 4.4: فلاتر المجموعات والمديات والنصوص

```
Southern admissions: 354
Smokers with high TG: 39
Pre-diabetic band: 302
'Medicine' departments: 308
```

كل رقم هنا جملة إحاطة جاهزة: 354 رقوداً جاءت من الإقليم الجنوبي؛ و39 مدخناً حالياً يحملون أيضاً دهوناً ثلاثية مرتفعة — قائمة أهداف جاهزة لاستشارة مشتركة للإقلاع عن التدخين وضبط الشحوم؛ و302 مريض يقعون في نطاق ما قبل السكري حيث لا يزال تدخل نمط الحياة قادراً على منع المرض.

الخطوة 4: متوسط مدة الرقود لكل قسم

أول تقسيم-تطبيق-جمع لنا: سطر واحد يفرز 1,200 سجل في ستة حوامل أقسام ويحسب متوسط مدة الرقود في كل منها.

```
1 los = merged.groupby("department")["length_of_stay_days"].mean().round(2)
2 print(los)
```

شيفرة 4.5: متوسط مدة الرقود حسب القسم

```
department
Emergency      5.58
ICU             9.26
Internal Medicine 5.43
Oncology        5.38
Pediatrics      5.20
Surgery         6.07
Name: length_of_stay_days, dtype: float64
```

تبرز العناية المركزة فوراً بمعدل 9.26 يوماً مقابل نحو 5.5 في كل مكان آخر — أمر متوقع للعناية الحرجة، لكنه الآن مُكَمَّم وقابل للتتبع شهراً بعد شهر.

الخطوة 5: ثلاثة إحصاءات لكل حامل بالدالة `agg`

المتوسط من دون وسيطه وحجم عينته نصف قصة؛ والدالة `agg` تجري الثلاثة كلها في تمريرة واحدة وتعيد إطار بيانات جاهزاً للتقرير.

```
1 los_stats = (merged.groupby("department")["length_of_stay_days"]
2             .agg(["mean", "median", "count"]).round(2))
```

```
3 print(los_stats)
```

شيفرة 4.6: المتوسط والوسيط والعدد في تمريرة واحدة

	mean	median	count
department			
Emergency	5.58	5.0	296
ICU	9.26	9.0	130
Internal Medicine	5.43	5.0	308
Oncology	5.38	5.0	112
Pediatrics	5.20	4.0	152
Surgery	6.07	5.0	202

في كل ردهة يجلس المتوسط فوق الوسيط — وهي بصمة توزيع رقود ملتوٍ نحو اليمين: معظم المرضى يغادرون سريعاً، وقلة يبقون طويلاً جداً. وعمود count بقي من الإفراط في تفسير الحوامل الصغيرة (قسم الأورام فيه 112 رقوداً فقط).

الخطوة 6: معدل إعادة الرقود بوصفه متوسط عمود منطقي

أرشفُ اصطلاحات اليوم: تحويل readmitted_30d من نص نعم/لا إلى قيمة منطقية يجعل المعدل ببساطة متوسط العمود (True تدخل المتوسط بقيمة 1). وتجميع ذلك المتوسط حسب القسم يعطي جدول ترتيب جودة مرتباً في سطرين.

```
1 merged["readmit"] = merged["readmitted_30d"] == "Yes"
2
3 readmit_rate = (merged.groupby("department")["readmit"]
4                  .mean().round(3)
5                  .sort_values(ascending=False))
6 print(readmit_rate)
```

شيفرة 4.7: معدل إعادة الرقود خلال 30 يوماً حسب القسم

department	
Oncology	0.161
Emergency	0.149
Surgery	0.129
Internal Medicine	0.127
ICU	0.108
Pediatrics	0.099
Name: readmit, dtype: float64	

يعيد قسم الأورام 16.1% من مرضاه خلال 30 يوماً وطب الأطفال 9.9%، حول متوسط مؤسسي قدره 13.0%. وبعض إعادات رقود الأورام دورات علاج كيميائي مخططة — فالرقم يفتح مراجعة الملفات ولا يغلقها.

الخطوة 7: الجدول التقاطعي للجنس مع حالة التدخين

جدول توافق في نداء واحد — الجدول الذي كان الإحصائي الحيوي سيُدخله في اختبار مربع كاي.

```
1 print(pd.crosstab(merged["sex"], merged["smoker"]))
```

شيفرة 4.8: جدول التوافق بين الجنس والتدخين

smoker	No	Yes
sex		
F	505	52
M	454	189

انتشار التدخين يعتمد على الجنس اعتماداً صارخاً: 189 من 643 رجلاً (29%) مقابل 52 من 557 امرأة (9%) — بما يتسق مع الوبائيات الإقليمية ويتصل مباشرة بتوجيه عيادات الإقلاع عن التدخين.

الخطوة 8: الجدول المحوري — متوسط HbA1c حسب المحافظة والجنس

الدالة `pivot_table` تحل محل محور إكسل — المحافظات صفوفاً، والجنس أعمدة، ومتوسط HbA1c في الخلايا.

```
1 hba1c_pivot = merged.pivot_table(values="hba1c_pct",
2                                   index="governorate",
3                                   columns="sex",
4                                   aggfunc="mean").round(2)
5 print(hba1c_pivot)
```

شيفرة 4.9: متوسط HbA1c حسب المحافظة والجنس

sex	F	M
governorate		
Anbar	5.50	5.50
Baghdad	5.62	5.66
Basra	5.41	5.52
Diyala	5.74	5.67
Erbil	5.57	5.65
Karbala	5.34	5.39
Najaf	5.59	5.42
Nineveh	5.55	5.57

يحوط متوسط HbA1c حول 5.4-5.7% في كل مكان — وهو أمر مطمئن على مستوى السكان، وتذكّر بأن المتوسطات تخفي المرضى الـ 237 الذين هم فردياً في نطاق السكري ولا يزال بوسع قناع أن يجدهم.

الخطوة 9: تحليل التواريخ وعدّ الرقود شهرياً

أخيراً، نعلم تاريخ الرقود أن يسلك سلوك تاريخ حقيقي، ونستخرج الشهر، ونعدّ؛ و `sort_index()` يحافظ على الترتيب التقويمي.

```
1 merged["admission_date"] = pd.to_datetime(merged["admission_date"])
2 merged["month"] = merged["admission_date"].dt.month
3
4 monthly = merged["month"].value_counts().sort_index()
5 print(monthly)
```

شيفرة 4.10: الرقود شهرياً عبر الموصّل `dt`.

month	
1	93
2	88
3	107
4	110
5	110
6	94
7	88
8	131
9	88
10	100
11	92
12	99
Name: count, dtype: int64	

يبلغ آب الذروة بـ 131 رقوداً — أعلى بنحو الثلث من أهدأ الشهور. وسواء أكان ذلك انعكاساً لأمراض الصيف المرتبطة بالحر أم أثراً تسجيلياً، فهذا بالضبط السؤال الذي ينبغي للمديرة أن تطرحه الآن؛ والشكل 4.3 يجعل النمط مستحيل الإغفال.

4.3 التحدي السريري (30 دقيقة)

التحدي السريري — طلب المدير الصباحي لمؤشرات الأداء

- في الساعة 8:15 من صباح هذا اليوم طلب مدير مديرية التشخيص لقطة تشغيلية من صفحة واحدة لسجل رقود عام 2025، مدمجاً مع لوحة الكيمياء، قبل اجتماع التخطيط الظهري. باستخدام جدول merged من التطبيق الموجّه، أنتج ما يلي:
- النمط الشهري: الرقود لكل شهر تقويمي، وشهر الذروة مع عدده.
 - الرقود حسب الفئة العمرية: متوسط ووسيط length_of_stay_days للفئات 0-17 و 18-39 و 40-64 و 65+، مبنيةً بالدالة pd.cut (دون جمل شرطية!) مع عدد المرضى في كل فئة.
 - ترتيب الجودة: معدل إعادة الرقود خلال 30 يوماً لكل قسم، مرتباً من الأسوأ إلى الأفضل.
 - الجغرافيا: المحافظات الثلاث الأولى بعدد الرقود.
 - سجل الخطورة الحرجة: عدد المرضى الذين لديهم $HbA1c \geq 6.5\%$ وعمر ≥ 65 — المجموعة التي تحتاج متابعة منسقة لسكري المسنين.
 - الصفحة الواحدة: جَمع الأرقام الرئيسة في إطار بيانات KPI صغير واحد (value, indicator) واكتب تفسيراً من خمس جمل للمدير.
- اعملوا في أزواج؛ أنهما المهام 1-5 في الدقائق الخمس عشرة الأولى، واحفظوا الدقائق الخمس الأخيرة لجدول KPI والتفسير.

✓ حل المدرب — تقرير مؤشرات الأداء العاجل، سطرًا سطرًا

التقرير كله ثماني عبارات خالية من الحلقات. المهمة 1 تعيد استخدام سلسلة monthly من الخطوة 9؛ و idxmax() يعيد وسم القيمة العظمى (أي الشهر)، و max() يعيد قيمتها.

```
1 # Task 1 --- monthly pattern and peak
2 monthly = merged["month"].value_counts().sort_index()
3 print("Peak month:", monthly.idxmax(), "with", monthly.max())
4
5 # Task 2 --- LOS by age band
6 age_band = pd.cut(merged["age"],
7                   bins=[0, 17, 39, 64, 120],
8                   labels=["0-17", "18-39", "40-64", "65+"])
9 los_by_band = (merged.groupby(age_band, observed=True)
10               ["length_of_stay_days"]
11               .agg(["mean", "median", "count"]).round(2))
12 print(los_by_band)
13
14 # Task 3 --- readmission ranking (worst first)
15 readmit_rate = (merged.groupby("department")["readmit"]
16               .mean().round(3)
17               .sort_values(ascending=False))
18
19 # Task 4 --- top-3 governorates
20 top3 = merged["governorate"].value_counts().head(3)
21 print(top3)
22
23 # Task 5 --- critical-risk register
24 critical = (merged["hba1c_pct"] >= 6.5) & (merged["age"] >= 65)
```

```

25 print("Critical-risk patients:", critical.sum())
26
27 # Task 6 --- the one-pager
28 kpi = pd.DataFrame({
29     "indicator": ["peak_month", "peak_admissions", "top_governorate",
30                 "worst_readmit_dept", "worst_readmit_rate",
31                 "critical_risk_n"],
32     "value": [monthly.idxmax(), monthly.max(), top3.index[0],
33              readmit_rate.index[0], readmit_rate.iloc[0],
34              critical.sum()])
35 print(kpi)

```

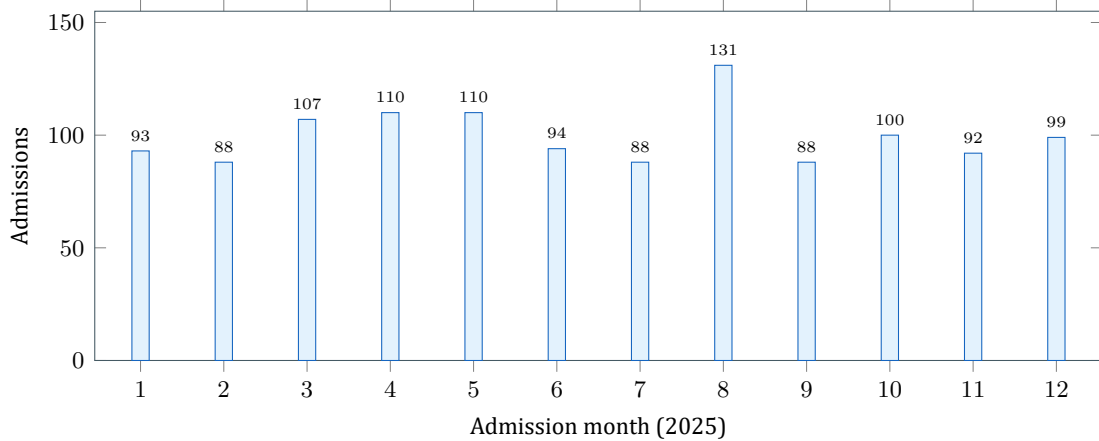
شيفرة 4.11: تقرير مؤشرات الأداء للمدير — الحل الكامل

```

Peak month: 8 with 131
      mean  median  count
age
0-17   6.00     5.0    117
18-39   5.91     5.0    338
40-64   5.87     5.0    530
65+     6.23     5.0    215
governorate
Baghdad   336
Basra     154
Nineveh   153
Name: count, dtype: int64
Critical-risk patients: 39
      indicator  value
0      peak_month      8
1  peak_admissions    131
2   top_governorate  Baghdad
3  worst_readmit_dept  Oncology
4  worst_readmit_rate    0.161
5   critical_risk_n     39

```

التفسير للمدير (خمس جمل). بلغت الرقود ذروتها في آب بـ 131 — أعلى بنحو الثلث من أهدأ الشهور — ولذلك ينبغي مراجعة ملاكات أواخر الصيف وتدبير المستهلكات قبل العام المقبل. مدة الرقود مستقرة عبر الفئات العمرية (الوسيطات كلها 5 أيام)، لكن أعلى متوسط لدى فئة 65+ البالغ 6.23 يوماً يكشف ذبلاً من رقود المسنين المطولة يستحق عناية بتخطيط الخروج. ويحمل قسم الأورام أعلى معدل إعادة رقود خلال 30 يوماً بنسبة 16.1% — جزء منه عودات علاج كيميائي مخططة، لكنه يظل جديراً بمراجعة ملفات مقابل المتوسط المؤسسي البالغ 13.0%. وتسهم بغداد وحدها بـ 336 من أصل 1,200 رقود، أي أكثر من ضعف البصرة (154) أو نينوى (153)، فحمل مسارات الإحالة شديد المركزية. وأخيراً، هناك 39 مريضاً هم في آن واحد بعمر 65 فأكثر وفي نطاق HbA1c السكري — فوج مسعى قابل للعدّ هو قائمة التسجيل الأولى الطبيعية لعيادة متابعة منسقة لسكري المسنين.



شكل 4.3: الرقود لكل شهر تقويمي في سجل عام 2025 (أعداد حقيقية)؛ ذروة آب البالغة 131 هي التي تقود توصية الملاكات في تقرير مؤشرات الأداء.

أسئلة النقاش الجماعي.

1. يتصدر قسم الأورام ترتيب إعادة الرقود، غير أن بعض إعادات رقوده دورات علاج كيميائي مخططة. ما العمود الإضافي في السجل الذي يجعل هذا المؤشر عادلاً، وكيف تفلتره بقناع؟
2. اذكر سؤالين من أسئلة المديرية يكون فيهما الدمج الأيسر للبيانات الديموغرافية على البيانات المختبرية هو الخيار الوحيد القابل للدفاع عنه.
3. فئة +65 لديها أطول متوسط رقود لكن وسيطها هو نفسه في كل فئة أخرى. أي الإحصاءين ينتمي إلى الصفحة الواحدة، وماذا تخبر الفجوة بينهما مخطط الخروج؟

4.4 الواجب المنزلي والتقييم الذاتي

تمارين الواجب المنزلي. التمارين الأربعة كلها تستخدم مجموعات بيانات الدورة؛ لا حلقات ولا جمل شرطية.

1. اصطياد حالات فقر الدم. ادمج سجل البيانات الديموغرافية (patients_demographics.csv) مع جدول صورة الدم الكاملة (cbc_results.csv) على patient_id. وباستخدام عتبات خاصة بكل جنس (هيموغلوبين دون 13 غ/دل للرجال ودون 12 للنساء)، ابن قناعاً مركباً واحداً بالمعاملين & و |، وعُدّ المرضى المصابين بفقر الدم، وقَدِّم معدل فقر الدم لكل قسم بوصفه متوسط القناع.
2. جدول ترتيب المختبرات. في rtPCR_results.csv، قنّع العمود result على القيمة Positive، واحتفظ بالجدول المفلتر باسم pos، وشغّل جدول الترتيب ذا السطر الواحد:

```
1 pos.groupby("lab_name")["ct_value"].agg(["mean", "count"])
```

ثم جملة واحدة: لماذا يُقلق متوسط Ct الأدنى منهجياً في أحد المختبرات ضابط الجودة؟

3. المحور الاستقلالي. على الجدول المدمج، ابن pivot_table لمتوسط glucose_mgdl يجعل department صفوفاً و smoker أعمدة. أي خلية هي الأعلى، وهل الفرق ذو دلالة سريرية؟
4. الجودة الموسمية. جَمِّع العمود المنطقي readmit حسب month: هل يتجاوز معدل إعادة الرقود في أي شهر المتوسط السنوي بأكثر من ثلاث نقاط مئوية؟

اختبار التحقق الذاتي.

1. لماذا يفشل merged["age"] > 60 & merged["hba1c_pct"] >= 6.5، وما التعبير المصحح؟
2. تحتاج إلى عدّ الرقود لكل قسم بترتيب الأقسام الثابت: value_counts() أم groupby(...).size()، ولماذا؟
3. بعد دمج البيانات الديموغرافية (1,200 صف) مع جدول مختبري، جاءت نتيجتك بـ 1,050 صفاً. ماذا حدث، وأي تغيير في الوسائط يستعيد الصفوف الـ 1,200 كلها؟

4. يعيد `merged["readmit"].mean()` القيمة 0.13. ترجم ذلك إلى جملة واحدة لمدير مستشفى.

5. ماذا تعيد `pd.cut`، ولماذا هي البديل الخالي من الحلقات عن سَلَم `if/elif` للفئات العمرية؟

أجوبة الاختبار. (1) & أشد ارتباطاً من المقارنات، فيقيّم بايثون `merged["hba1c_pct"]` & 60 أولاً؛ اكتب & `merged["age"] > 60` (2) `merged["hba1c_pct"] >= 6.5` (3) أسقط الدمج الداخلي 150 مريضاً بلا صف مختبري؛ و `how="left"` يبقّي الصفوف 1,200 كلها مع `NaN` للنتائج الغائبة. (4) ثلاثة عشر بالمئة من المرضى الراقدين عادوا خلال 30 يوماً من الخروج. (5) سلسلة فئوية مرتبة تسند كل مريض إلى فئة واحدة؛ فحدود الفئات تحل محل سَلَم `if/elif` وتطبق على العمود كله دفعة واحدة.

♣ القيمة لمديرية التشخيص — من تقارير الأسبوع إلى أجوبة الصباح نفسه

تحول عُدة اليوم أكثر الطلبات الإحصائية شيوعاً في المديرية — قوائم الحالات، وجداول ترتيب الأقسام، ومعدلات إعادة الرقود، والأنماط الموسمية — من مشاريع كتابية تستغرق أياماً إلى نصوص برمجية تعمل في ثوانٍ وتُعاد على سجل الشهر المقبل دون تعديل. والمكاسب ثلاثية: السرعة (صفحة KPI الواحدة تجيب عن طلب في الصباح نفسه)، وقابلية التدقيق (كل عتبة تجلس في الشيفرة إلى جانب دليلها الإرشادي، فيتحقق المراجع من تعريف الحالة بدل الوثوق بجدول مفلتر)، والاتساق (مكتبا بغداد والبصرة اللذان يشغلان النص نفسه يتجان مؤشرات قابلة للمقارنة — وهو الشرط المسبق للمقارنة المعيارية الوطنية).

خلاصات اليوم

- القناع المنطقي عمود كامل الطول من `False/True`: `sum()` يعدّ الحالات و `mean()` هو معدلها — القناع هو الإحصاء.
- اجمع المعايير `&` و `|` و فقط، وكل مقارنة بين قوسها الخاصين — الكلمتان الإنكليزيتان `or/and` لا تعملان على الأعمدة.
- `isin` يجيب عن أسئلة المجموعات، و `between` عن أسئلة المجالات المرجعية، و `str.contains` عن أسئلة النص الحر.
- `groupby` هو التقسيم-التطبيق-الجمع: افرز الأنابيب في حوامل، وقس كل حامل، واكتب سطرًا لكل حامل؛ و `agg` يجري عدة قياسات في تمريرة واحدة.
- المعدل هو متوسط عمود منطقي — إعادة الرقود والإيجابية والالتزام كلها تؤول إلى الاصطلاح ذي السطرين نفسه.
- `crosstab` و `pivot_table` محاور إكسل قابلة لإعادة الإنتاج ومحضنة ضد التلف.
- `merge` يدمج على `patient_id`: الداخلي يبقّي الفوج كامل التوثيق، والأيسر يبقّي كل مريض — وفحص `shape`. بعد كل دمج أمر غير قابل للتفاوض.
- `pd.to_datetime` مع `dt` يحول التواريخ إلى استخبارات موسمية؛ و `pd.cut` يحول الأعمار إلى فئات سريرية دون جملة شرطية واحدة.

غداً، في الفصل 5، تتعلم هذه الجداول الكلام: سنحول نتائج `groupby` اليوم والأعداد الشهرية إلى رسوم بمستوى النشر، ونُدع العلامات الحيوية تروي قصتها بصرياً.

باب 5

اليوم 5 — التصوير البصري للبيانات وقصة العلامات الحيوية

يستطيع نظام معلومات المختبر أن يطبع عشرة آلاف قيمة هييموغلوبين في أعمدة مرتبة، فتومئ اللجنة برؤوسها بأدب ولا تتعلم شيئاً. ارسم القيم العشرة آلاف نفسها منحنى واحداً، وسيرى كل طبيب في القاعة، خلال ثانيتين، أين يقع مريض فقر الدم، وهل ينفصل الرجال والنساء إلى مجموعتين سكانيتين، وهل تستحق القيم الشاذة (Outliers) اتصالاً هاتفياً. إن شبكية العين أسرع أداة تشخيصية تملكها المديرية؛ والتصوير البصري للبيانات (Data Visualization) هو الطريقة التي نغذيها بها. بالأمس، في الفصل 4، تعلمت أن تعزل بدقة المرضى الذين يهتمون باستخدام الأقنعة المنطقية (Boolean Masks) وأن تلخصهم بالتجميع (groupby). واليوم تتحول تلك الجداول المفلترتة إلى صور: مدرجات تكرارية (Histograms) تكشف كيف تولد المجالات المرجعية، ومخططات صندوقية (Boxplots) تضع غلوكوز مريض السكري وغير المصابين به جنباً إلى جنب، ومنحنى وبائي مُجمّع من الطوايح الزمنية الخام لفحوص RT-PCR وخريطة حرارية للارتباط (Correlation Heatmap) تكثف لوحة كيمياء سريرية من اثني عشر محللاً في لحظة واحدة. وبنهاية الجلسة ستبني لوحة الترخيد ذات اللوحات الأربع التي يُفتتح بها الإيجاز الأسبوعي للمديرية.

الدقائق	الجزء	المحتوى والأنشطة
00-05	الاسترجاع والتأطير	مراجعة اليوم الرابع: الأقنعة و groupby؛ سؤال اليوم — كيف تتحول الأرقام إلى قرارات؟
05-15	النظرية الأولى	لماذا نصور البيانات؛ تشرح الشكل في seaborn: matplotlib بوصفها الطبقة الإحصائية؛ جودة النشر
15-20	النظرية الثانية	دليل اختيار المخطط: المدرج التكراري، المخطط الصندوقي، الخطي، الانتشار، الخريطة الحرارية؛ المحاور الأمنية
20-35	التطبيق الموجّه الأول	دمج صورة الدم الكاملة مع البيانات الديموغرافية؛ أول مدرج تكراري؛ تراكب الهيموغلوبين حسب الجنس
35-45	التطبيق الموجّه الثاني	مخطط صندوقي للغلوكوز حسب نتيجة السكري؛ مدرج تكراري لقيم Ct للعينات الإيجابية
45-60	التطبيق الموجّه الثالث	المنحنى الوبائي الأسبوعي؛ انتشار مؤشر كتلة الجسم مقابل الغلوكوز؛ الخريطة الحرارية للكيمياء؛ savefig بدقة 300 dpi
60-80	التحدي السريري	بناء لوحة الترخيد 2 × 2 للإيجاز الأسبوعي
80-90	النتائج والنقاش الختامي	كتابة النتائج الرئيسية الثلاث؛ نقاش؛ الواجب المنزلي ولمحة عن اليوم السادس

✓ الأهداف التعليمية

- بنهاية هذه الجلسة ستكون قادراً على:
- اختيار نوع المخطط الصحيح (المدرج التكراري، المخطط الصندوقي، الخطي، الانتشار، الخريطة الحرارية) لسؤال سريري معيّن، وتسويق هذا الاختيار.
 - بناء شكل matplotlib بصورة صريحة من أجزائه — Axes و Figure وتسميات المحاور مع الوحدات والعنوان ومفتاح الرموز

— باستخدام `plt.subplots()`.

3. تركيب توزيعين متراكبين بالشفافية لمقارنة الهيموغلوبين بين الجنسين، مع تعليم الحدود بواسطة `axvline`.
4. بناء منحني وبائي أسبوعي من تواريخ جمع عينات RT-PCR الخام باستخدام `to_datetime()` و `dt.isocalendar()` ومن دون حلقات تكرارية.
5. رسم خريطة حرارية لارتباطات لوحة الكيمياء السريرية بواسطة `seaborn` وقراءتها بحثاً عن الأزواج ذات الدلالة السريرية.
6. تصدير أشكال بجودة النشر بواسطة `savefig(..., dpi=300)` لتقارير الوزارة.
7. تجميع لوحة متابعة متعددة اللوحات بواسطة `plt.subplots(2, 2)` وتلخيصها في ثلاث نتائج رئيسية موجهة إلى الإدارة.

➤ حقبة أدوات الذكاء الاصطناعي والاعتماديات — حزمة الرسم البياني لليوم الخامس

يضيف اليوم مكتبة الرسم القياسيتين في بايثون إلى أساس `pandas` و `NumPy` الذي بنته الأيام من الثاني إلى الرابع.

```
1 import pandas as pd
2 import numpy as np
3 import matplotlib.pyplot as plt
4 import seaborn as sns
```

شيفرة 5.1: الاستيرادات المستخدمة في جميع أنحاء اليوم الخامس

```
pip install matplotlib seaborn
Successfully installed matplotlib-3.10.8 seaborn-0.13.2
```

matplotlib هي محرك الرسم: كل شكل ومحور وعمود وخط هو كائن تتحكم به أنت — أشكال بأسلوب مؤسسي موحد، ووحدات على كل محور، وتصدير بدقة الطباعة للتقارير الرسمية بدلاً من لقطات شاشة الجداول الإلكترونية. أما **seaborn** فتجلس فوق **matplotlib** وتفهم إطارات البيانات (`DataFrame`) والإحصاء؛ إذ تحول دالتها `heatmap` مصفوفة الارتباط إلى ملخص مرمز بالألوان في سطر واحد، فيتمكن ضابط الترصد من مسح لوحة كيمياء سريرية كاملة في ثوانٍ. وتواصل **pandas** و **numpy** (المثبتتان سلفاً) القيام بالتحضير: فالدمج والتجميع ومعالجة التواريخ تغذي كل مخطط.

5.1 السياق السريري والإطار النظري (20 دقيقة)

الأرقام تقنع المتخصصين؛ والصور تقنع المؤسسات. عندما تطلب المديرية موازنة لشرائط أجهزة قياس الغلوكوز، فإن الشريحة الرابعة ليست جدولاً يضم 1200 قيمة غلوكوز — بل مدرج تكراري بخط أحمر متقطع عند 126 ملغ/دل وذيل مظلل كُتب عليه 15.9% من المرضى المفحوصين. يخدم التصوير البصري سيدتين: فهو بالنسبة إلى المحلل أداة اكتشاف (التوزيعات تفضح أخطاء إدخال البيانات، والمجموعات ثنائية القمة، والأجهزة المنجرفة تدريباً التي تنعمها الملخصات)؛ وهو بالنسبة إلى الإدارة أداة تواصل.

لماذا تتفوق العين على الجدول

الإحصاءات الموجزة تضغط البيانات؛ والضغط يدمر. بنى الإحصائي أنسكومب (`Anscombe`) أربع مجموعات بيانات متطابقة في المتوسطات والتباينات والارتباطات، لكن مخططات انتشارها لا يشبه أحدها الآخر في شيء — واحدة خطية، وواحدة منحنية، وواحدة تهيمن عليها قيمة شاذة واحدة. والترجمة السريرية: يمكن لجناحين أن يبلغا عن متوسط الكرياتينين نفسه بينما يخفي أحدهما عنقوداً من مرشحي الغسيل الكلوي. يحافظ مخطط التوزيع على ما يتخلص منه المتوسط بالضبط — الشكل والانتشار والالتواء والفجوات والقيم الشاذة — والقيم الشاذة هي الموطن الذي يعيش فيه التشخيص. قاعدة المديرية: لا تبليغ عن متوسط قط من دون أن تكون قد نظرت إلى التوزيع الذي أنتجه.

دليل اختيار المخطط

يُختار المخطط بحسب السؤال، لا بحسب العادة أبداً. خمسة أسئلة تغطي جلّ العمل التشخيصي: (1) كيف تتوزع قيم محلّ واحد؟ — المدرج التكراري: قسّم المحور إلى فئات وعدّ المرضى في كل فئة. (2) هل تختلف قيمة بين المجموعات، مع وجود قيم شاذة؟ — المخطط الصندوقي: الوسيط وصندوق المدى الربيعي (IQR) والشاربان والقيم المتطرفة المعلّمة. (3) كيف تتحرك كمية عبر الزمن؟ — المخطط الخطي: العلامات الحيوية، وانجراف ضبط الجودة، والمنحنيات الوبائية. (4) هل يرتبط محلّان مريضاً بمريض؟ — مخطط الانتشار: نقطة واحدة لكل مريض، والشفافية في مواجهة التراكب المفرط للنقاط. (5) أي المتغيرات الكثيرة تتحرك معاً؟ — الخريطة الحرارية لمصفوفة الارتباط. يربط الجدول 5.1 كل سؤال بمثال اليوم؛ والشكل 5.1 هو خريطة القرار التي تُعلّق فوق مكتب المحلل.

جدول 5.1: دليل اختيار المخطط مع أمثلة اليوم السريرية.

السؤال السريري	المخطط	مثال اليوم
كيف يتوزع محلّ واحد؟ هل تختلف قيمة بين المجموعات؟	المدرج التكراري المخطط الصندوقي	الهيموغلوبين عبر 1200 حالة دخول الغلوكوز لدى مرضى السكري مقابل غيرهم
كيف تتغير عبر الزمن؟	المخطط الخطي	الإيجابيات الأسبوعية لفحص RT-PCR لعام 2025
هل يرتبط محلّان؟ أي المتغيرات تتحرك معاً؟	مخطط الانتشار الخريطة الحرارية	مؤشر كتلة الجسم مقابل الغلوكوز ارتباطات لوحة الكيمياء السريرية

Start from the clinical question — not from the chart

How are values of *one* analyte distributed?
(Hgb, Ct)

Histogram

Does a value differ between *groups*? Any
outliers?

Boxplot

How does a quantity move over *time*?

Line chart

Are *two* analytes related, patient by pa-
tient?

Scatter plot

Which of *many* variables move together?

Heatmap

شكل 5.1: خريطة القرار لاختيار نوع المخطط: خمسة أسئلة سريرية والمخطط الذي يجيب عن كل منها. كل شكل في جلسة اليوم هو حالة من حالات هذه الصفوف الخمسة.

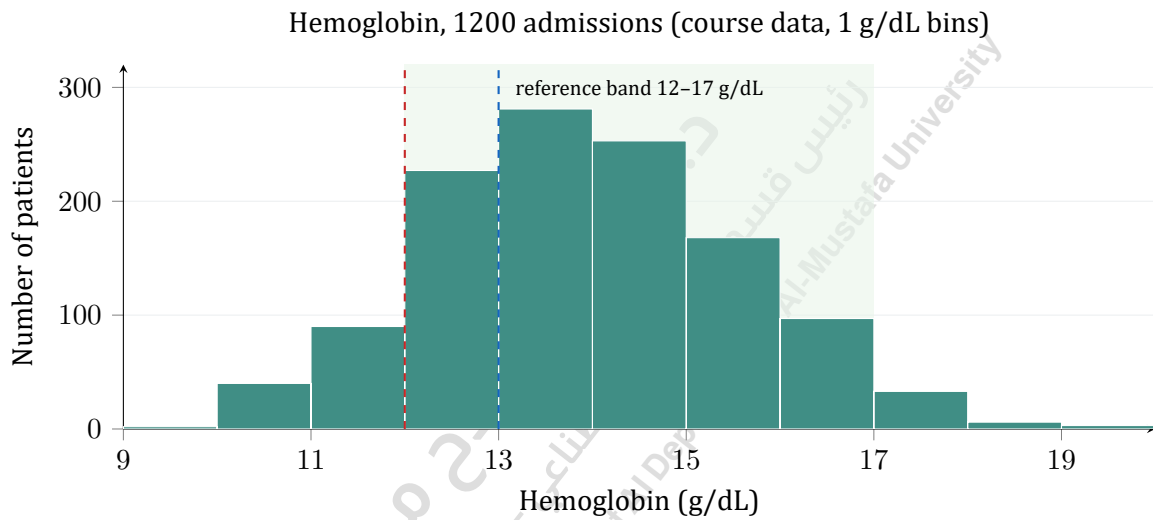
تشرح الشكل في matplotlib

تميّز matplotlib بين الشكل Figure — اللوحة القماشية بأكملها — والمحاور Axes — نظام إحداثيات واحد مرسوم عليها. والصيغة الاحترافية صريحة: `fig, ax = plt.subplots()` تنشئ الاثنين معاً وتسلمك مقبضين مسميين. وكل ما عدا ذلك هو توابع على هذين المقبضين: `ax.hist` و `ax.plot` و `ax.scatter` و `ax.boxplot` ترسم؛ و `ax.set_xlabel` و `ax.set_ylabel` و `ax.set_title` و `ax.legend` تضيف الشروح؛ و `ax.axhline` و `ax.axvline` تعلّمان قيم العتبة السريرية؛ و `fig.savefig` تصدّر. ويمكن لشكل واحد أن يحمل محاور عدة — إذ تعيد `plt.subplots(2, 2)` الشبكة 2 × 2 التي تُبنى منها لوحات المتابعة. ولا تستبدل هذا `seaborn` هذا التشرح؛ بل ترسم مخططات واعية إحصائياً على كائن Axes من matplotlib (لاحظ الوسيط `ax=ax`). وجودة النشر قائمة تدقيق لا موهبة: تسميات محاور مع الوحدات، وعنوان معبّر، ومفتاح

رموز حين تتشارك سلسلتان مخططات واحداً، وfig.savefig مع dpi=300 — معيار الطباعة: فالإعداد الافتراضي dpi 100 يبدو ضبابياً في وثيقة وزارية.

المجالات المرجعية تولد من التوزيعات

من أين يأتي الهيموغلوبين الطبيعي 13-17 غ/دل؟ ليس من النظرية — بل من مدرج تكراري. يقيس المختبر عينة مرجعية كبيرة من الأصحاء، ويرسم التوزيع، ويأخذ الـ 95% المركزية (من المئين 2.5 إلى المئين 97.5، أو المتوسط ± 1.96 انحرافاً معيارياً عندما يكون التوزيع غاوسياً) مجالاً مرجعياً. ولهذا نتيجتان يوميتان: بحكم البناء نفسه، يقع 5% من الأشخاص الأصحاء خارج المجال، ولذلك فإن نتيجة واحدة شاذة شذوذاً طفيفاً في لوحة من عشرين محلاً أمر متوقع لا مقلق؛ كما أن المجالات خاصة بكل فئة سكانية — فتوزيعا الهيموغلوبين لدى الذكور والإناث لهما مركزان متميزان، وهذا هو سبب حملهما حدوداً منفصلة. يعرض الشكل 5.2 حالات الدخول الـ 1200 الخاصة بنا مع تظليل النطاق المرجعي: المجال وهو ينبثق من البيانات.



شكل 5.2: الأعداد الحقيقية المجمعة في فئات للهيموغلوبين من cbc_results.csv. يمتد الشريط المظلل على المجال المرجعي المشترك للبالغين؛ ويعلم الخطان المتقطعان الحد الأدنى للإناث (12 غ/دل، أحمر) والحد الأدنى للذكور (13 غ/دل، أزرق). الذيل الأيسر هو عبء فقر الدم الذي تتحمله المديرية.

★ لؤلؤة سريرية — المجالات المرجعية وقيم العتبة المستخدمة اليوم

الهيموغلوبين: الذكور 13-17 غ/دل، والإناث 12-15 غ/دل؛ ويقع فقر الدم تحت الحد الأدنى الخاص بكل جنس. غلوكوز البلازما الصائم: طبيعي دون 100 ملغ/دل، واختلال تحمّل 100-125 ملغ/دل، والسكري عند ≥ 126 ملغ/دل في مناسبتين (معايير ADA). السكر التراكمي: HbA1c طبيعي دون 5.7%، وما قبل السكري 5.7-6.4%، والسكري عند $\geq 6.5\%$. قيمة عتبة الدورة (Ct) لفحص: RT-PCR هي الدورة التي يعبر عندها التآلق خط الكشف؛ وانخفاض Ct يعني فيروساً أكثر — فدون 25 حمل فيروسي مرتفع، و25-30 متوسط، و30-35 منخفض، و35-40 منخفض جداً قرب حد الكشف حيث ينبغي التفكير في إعادة الفحص — مع التذكر دائماً أن مقاييس Ct تعتمد على العدة التشخيصية والمنصة. هذه هي بالضبط الخطوط التي نرسمها على أشكال اليوم بواسطة axhline وaxvline.

✖ خطأ شائع — المحور المبتور: كيف تكذب بيانات صادقة

أسرع طريقة لتضليل إيجاز ما من دون تزيف رقم واحد هي بتر المحور الصادي. في بيانات اليوم، متوسط السكر التراكمي HbA1c هو 5.56% للإناث و5.57% للذكور — متطابقان سريرياً. أرسم هذين العمودين على محور يمتد من 5.50 إلى 5.60 فيشمل عمود الذكور بضعف الارتفاع البصري، موحياً بفرق بين الجنسين لا وجود له. قواعد الدار: محاور العد تبدأ من الصفر؛ والتكبير المطلوب فعلاً (قيم Ct مثلاً) يُصرّح به في التسمية التوضيحية؛ ولا يُكَبَّر محور أبداً بعد رؤية أي قصة يروها التكبير.

5.2 التطبيق الموجّه (40 دقيقة)

■ عرض تطبيقي مباشر — من ملفات CSV الخام إلى خمسة مخططات بجودة النشر

الهدف: إنتاج مثال عملي واحد لكل مخطط في دليل الاختيار انطلاقاً من ملفات الدورة الخام — مدرجات تكرارية متراكبة للهيموغلوبين حسب الجنس، ومخطط صندوق للغلوكوز حسب نتيجة السكري، ومدرج تكراري لقيم Ct، ومنحنى وبائي أسبوعي لفحص RT-PCR، ومخطط انتشار لمؤشر كتلة الجسم مقابل الغلوكوز، وخريطة حرارية لارتباطات الكيمياء السريرية — كل منها موسوم بالكامل ومُصدّر بدقة 300 dpi. والشيفرة كلها مكتوبة بأسلوب المعالجة الموجهة (Vectorization): عمليات دمج وأقنعة و groupby ومن دون حلقات تكرارية. يكتب المتدربون مع المدرب؛ وكل استدعاء رسم يسبقه فحص سلامة مطبوع.

الخطوة 1: دمج نتائج صورة الدم مع البيانات الديموغرافية

يحمل ملف صورة الدم الكاملة (CBC) النتائج لكن من دون الجنس والعمر اللذين يقيمان في الملف الديموغرافي. وكما في الفصل 4، يقوم الدمج (merge) على patient_id بوصلهما. اطبع الأبعاد دائماً بعد كل دمج — فانفجار الصفوف أو فقدانها بصمت هو أكثر حوادث البيانات شيوعاً.

```
1 import pandas as pd
2 import numpy as np
3 import matplotlib.pyplot as plt
4 import seaborn as sns
5
6 pd.set_option("display.width", 100)
7
8 cbc = pd.read_csv("cbc_results.csv")
9 demo = pd.read_csv("patients_demographics.csv")
10 blood = cbc.merge(demo[["patient_id", "age", "sex"]], on="patient_id")
11 print(blood.shape)
12 print(blood[["patient_id", "hgb", "wbc", "age", "sex"]].head(4))
```

شيفرة 5.2: دمج نتائج صورة الدم الكاملة مع البيانات الديموغرافية

```
(1200, 11)
  patient_id  hgb  wbc  age sex
0 MOH-00001  16.3   6.7   53  M
1 MOH-00002  13.6  12.7   23  M
2 MOH-00003  13.9   8.2   63  M
3 MOH-00004  12.7   7.7   67  M
```

ألف ومئتا صف دخلت، وألف ومئتا صف خرجت — دمجٌ واحد لواحد كما هو متوقع، وصار الجنس الآن ملحفاً بكل قيمة هيموغلوبين.

الخطوة 2: أول مدرج تكراري — التشرّيح عملياً

محللٌ واحد وسؤال واحد: كيف يتوزع الهيموغلوبين؟ نبنى الشكل من أجزاء مسمّاة كي يتبع كل مخطط لاحق الهيكل نفسه. الملخص العددي أولاً، ثم الصورة التي تفسّره.

```
1 print(blood["hgb"].describe().round(1))
2
3 fig, ax = plt.subplots(figsize=(8, 5))
4 ax.hist(blood["hgb"], bins=24, color="teal", edgecolor="white")
5 ax.set_xlabel("Hemoglobin (g/dL)")
6 ax.set_ylabel("Number of patients")
7 ax.set_title("Hemoglobin distribution, all admissions 2025")
8 fig.savefig("hgb_hist.png", dpi=300)
```

شيفرة 5.3: مدرج تكراري موسوم بالكامل للهيموغلوبين

```
count    1200.0
mean      13.9
std        1.6
min        9.8
25%       12.8
50%       13.8
75%       15.0
max       19.5
Name: hgb, dtype: float64
```

المتوسط البالغ 13.9 غ/دل يبدو مطمئناً، لكن التوزيع يمتد من 9.8 إلى 19.5، والكشف الأيسر تحت 12 هو بالضبط عبء حالات فقر الدم (قارن الشكل 5.2).

الخطوة 3: مجموعتان سكانيتان في إطار واحد — الهيموغلوبين حسب الجنس

هل تلك القاعدة العريضة هي حقاً مجموعة واحدة؟ افصل حسب الجنس بقناعين منطقيين، وركّب المدرجين التكرارين أحدهما فوق الآخر بالشفافية (alpha)، وعلم الحد المرجعي الأدنى لكل جنس. ويؤكد ناتج groupby المطبوع ما ستراه العين.

```
1 print(blood.groupby("sex")["hgb"].agg(["count", "mean", "std"]).round(2))
2
3 male = blood[blood["sex"] == "M"]
4 female = blood[blood["sex"] == "F"]
5
6 fig, ax = plt.subplots(figsize=(8, 5))
7 ax.hist(male["hgb"], bins=24, alpha=0.6, label="Male", color="steelblue")
8 ax.hist(female["hgb"], bins=24, alpha=0.6, label="Female", color="firebrick")
9 ax.axvline(13.0, linestyle="--", color="steelblue")
10 ax.axvline(12.0, linestyle="--", color="firebrick")
11 ax.set_xlabel("Hemoglobin (g/dL)")
12 ax.set_ylabel("Number of patients")
13 ax.set_title("Hemoglobin by sex with lower reference limits")
14 ax.legend()
15 fig.savefig("hgb_by_sex.png", dpi=300)
```

شيفرة 5.4: مدرجات تكرارية متراكبة للهيموغلوبين حسب الجنس

```
count    mean    std
sex
F        557   13.10  1.38
M        643   14.58  1.53
```

جرسان متداخلاً يفصل بين مركزيهما 1.5 غ/دل — البرهان البصري على سبب حمل الهيموغلوبين حدوداً خاصة بكل جنس. ومن دون alpha كان المدرج التكراري الثاني سيغطي الأول ببساطة؛ فالشفافية تحول مخططين إلى مقارنة.

الخطوة 4: المخطط الصندوقي — الغلوكوز حسب نتيجة السكري

سؤال جديد، مخطط جديد: هل يختلف الغلوكوز بين مجموعتي النتيجة، ومن هم أصحاب القيم الشاذة؟ قناعان يبينان المجموعتين من عمود النتيجة 0/1؛ وترسم ax.boxplot الوسيط وصندوق المدى الربيعي والشاربين ونقاط القيم الشاذة، بينما تضيف axhline عتبة التشخيص عند 126 ملغ/دل.

```
1 dm = pd.read_csv("diabetes_dataset.csv")
2 print(dm.groupby("outcome")["glucose"].describe().round(1))
```

```

3     ["count", "25%", "50%", "75%"]])
4
5 no_dm = dm[dm["outcome"] == 0]["glucose"]
6 yes_dm = dm[dm["outcome"] == 1]["glucose"]
7
8 fig, ax = plt.subplots(figsize=(7, 5))
9 ax.boxplot([no_dm, yes_dm], tick_labels=["No diabetes", "Diabetes"])
10 ax.axhline(126, linestyle="--", color="firebrick")
11 ax.set_ylabel("Plasma glucose (mg/dL)")
12 ax.set_title("Glucose by diabetes outcome (n=768)")
13 fig.savefig("glucose_box.png", dpi=300)

```

شيفرة 5.5: مخطط صندوقي للغلوكوز حسب النتيجة

	count	25%	50%	75%
outcome				
0	485.0	88.0	104.0	121.0
1	283.0	120.0	136.0	151.0

يجلس وسيط مرضى السكري (136) فوق خط التشخيص، ووسيط غير المصابين (104) تحته — ومع ذلك يتداخل الصندوقان حول 120، وهذا بالضبط هو السبب في أن قيمة غلوكوز واحدة لا تشخص السكري. ومنطقة التداخل تلك هي حيث ستثبت هندسة الميزات في الفصل 6 جدارتها.

الخطوة 5: قيم Ct للعينات الإيجابية في فحص RT-PCR

الصف الأول من الدليل مطبقاً على التشخيص الجزيئي: توزيع قيم Ct بين العينات الإيجابية يخبر المختبر هل يلتقط العدوى مبكراً (Ct منخفضة، حمل فيروسي مرتفع) أم متأخراً. عمود ct_value فارغ للعينات السلبية، ولذلك نفلتر إلى الإيجابيات أولاً — تقني من صميم اليوم الرابع.

```

1 pcr = pd.read_csv("rtPCR_results.csv", parse_dates=["collection_date"])
2 print(pcr["result"].value_counts())
3
4 pos = pcr[pcr["result"] == "Positive"]
5 print(pos["ct_value"].describe().round(1))
6
7 fig, ax = plt.subplots(figsize=(8, 5))
8 ax.hist(pos["ct_value"], bins=20, color="purple", edgecolor="white")
9 ax.axvline(30, linestyle="--", color="firebrick")
10 ax.set_xlabel("Ct value (cycles)")
11 ax.set_ylabel("Number of positive samples")
12 ax.set_title("Ct distribution of positives (dashed: moderate/low boundary)")
13 fig.savefig("ct_hist.png", dpi=300)

```

شيفرة 5.6: مدرج تكراري لقيم عتبة الدورة بين العينات الإيجابية

```

result
Negative      371
Positive      358
Inconclusive    71
Name: count, dtype: int64
count      358.0
mean        24.8
std          6.6
min         14.1
25%         19.1
50%         24.8

```

```
75%      31.0
max      36.0
Name: ct_value, dtype: float64
```

وسيط Ct البالغ 24.8 يعني أن نصف الإيجابيات جميعها تحمل حملاً فيروسياً مرتفعاً عند جمع العينة — أي أن الحالات تُكتشف وهي لا تزال معدية. أما الربع الذي يتجاوز Ct قيمة 31 فهو اكتشافات متأخرة أو عدوى في طور الانحسار.

الخطوة 6: المنحنى الوبائي الأسبوعي

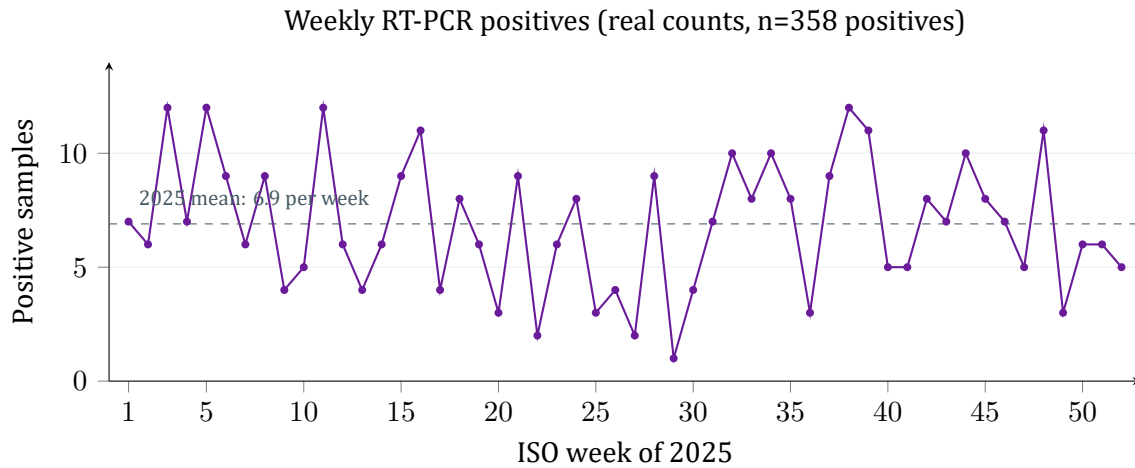
الشكل الأكثر طلباً في أي فاشية: الإيجابيات في كل أسبوع. يحوي الملف الخام صفّاً واحداً لكل عينة مع طابع زمني؛ والتجميع `groupby` على رقم أسبوع ISO المستخرج من `collection_date` يبني المنحنى كله في سطرين — من دون حلقات تكرارية ومن دون حساب يدوي للتواريخ.

```
1 week = pos["collection_date"].dt.isocalendar().week
2 weekly = pos.groupby(week)["sample_id"].count()
3 print(weekly.head(6))
4 print("Weeks:", weekly.shape[0], "| peak week:", weekly.idxmax(),
5       "| cases:", weekly.max())
6
7 fig, ax = plt.subplots(figsize=(9, 4.5))
8 ax.plot(weekly.index, weekly.values, marker="o", color="teal")
9 ax.set_xlabel("ISO week of 2025")
10 ax.set_ylabel("Positive samples")
11 ax.set_title("Weekly RT-PCR positives, all laboratories")
12 fig.savefig("epi_curve.png", dpi=300)
```

شيفرة 5.7: الإيجابيات الأسبوعية لفحص تفاعل البوليميراز المتسلسل كمخطط خطي

```
week
1      7
2      6
3     12
4      7
5     12
6      9
Name: sample_id, dtype: int64
Weeks: 52 | peak week: 3 | cases: 12
```

يعرض الشكل 5.3 المنحنى الكامل مرسوماً من هذه الأعداد بعينها: من حالة واحدة إلى 12 حالة في الأسبوع، ومن دون موجة مستمرة — نمط متوطن. القيمة هي خط الأساس ذاته؛ ففي الموسم المقبل، ثلاثة أسابيع متتالية فوق هذا النطاق هي الإنذار.



شكل 5.3: المنحنى الوبائي لعام 2025 المبني في القائمة 5.7 من `rtPCR_results.csv`: الإيجابيات الأسبوعية تتذبذب حول متوسط قدره 6.9 من دون موجة مستمرة. هذا الخط الأساسي يحدد عتبة الإنذار للموسم المقبل.

✖ خطأ شائع — أسابيع ISO تلتف حول رأس السنة

يعيد `dt.isocalendar().week` أسبوع سنة ISO، لا أسبوع السنة التقويمية. ففي ملفنا نحن، تنتهي العينات الأربع المجموعة في 29-31 كانون الأول 2025 إلى الأسبوع 1 من سنة 2026، ولذلك يضيفها `groupby(week)` بصمت إلى الأسبوع 1 الخاص بكانون الثاني — تشويه صغير هنا، لكنه في مجموعة بيانات تمتد عبر شتاءين سيخيظ فاشيتين مختلفتين في نقطة واحدة. الصيغة الآمنة عبر السنوات تجمع حسب فترة أسبوعية بدلاً من ذلك: استدع `dt.to_period` بتردد أسبوعي على `collection_date` وجمع حسب الناتج. وكلما عبرت سلسلة زمنية الأول من كانون الثاني، فضل `dt.to_period`.

الخطوة 7: مخطط الانتشار — مؤشر كتلة الجسم مقابل الغلوكوز

السؤال الرابع من الدليل: هل يرتبط محلّان؟ نقطة واحدة لكل مريض، مؤشر كتلة الجسم على المحور السيني والغلوكوز على الصادي. ومع 768 نقطة متراكبة، يسمح $\alpha=0.4$ للكثافة بالظهور، وتلون `np.where` مرضى السكري بلون مختلف — معالجة موجهة، من دون تلوين بحلقة وجملية شرطية.

```
1 print("Correlation bmi vs glucose:",
2       dm["bmi"].corr(dm["glucose"]).round(2))
3
4 colors = np.where(dm["outcome"] == 1, "firebrick", "teal")
5 fig, ax = plt.subplots(figsize=(7, 5.5))
6 ax.scatter(dm["bmi"], dm["glucose"], c=colors, alpha=0.4, s=18)
7 ax.axhline(126, linestyle="--", color="dimgray")
8 ax.set_xlabel("Body mass index (kg/m2)")
9 ax.set_ylabel("Plasma glucose (mg/dL)")
10 ax.set_title("BMI vs glucose, red = diabetic outcome (n=768)")
11 fig.savefig("bmi_glucose.png", dpi=300)
```

شيفرة 5.8: مخطط انتشار لمؤشر كتلة الجسم مقابل الغلوكوز ملوّن حسب النتيجة

Correlation bmi vs glucose: 0.76

ارتباط قدره 0.76 قوي بمقاييس البيانات السريرية: يرتفع الغلوكوز مع مؤشر كتلة الجسم، وتتركز النقاط الحمراء في الربع العلوي الأيمن — إنها الصورة الكامنة خلف ميزات التنبؤ التي سنبنها غداً.

الخطوة 8: الخريطة الحرارية — لوحة الكيمياء السريرية في لمحة

اننا عشر محلاً يعطون 66 ارتباطاً ثنائياً؛ ولا أحد يقرأ 66 مخطط انتشار. تحسب `DataFrame.corr()` المصفوفة وترسمها `sns.heatmap` طابعة كل معامل في خليته. نختار ستة محلات كي تبقى المصفوفة المطبوعة قابلة للقراءة.

```
1 chem = pd.read_csv("chemistry_panel.csv")
2 cols = ["glucose_mgdl", "hba1c_pct", "creatinine_mgdl", "urea_mgdl",
3         "total_chol_mgdl", "ldl_mgdl"]
4 corr = chem[cols].corr()
5
6 short = corr.round(2)
7 short.index = ["glucose", "hba1c", "creat", "urea", "chol", "ldl"]
8 short.columns = short.index
9 print(short)
10
11 fig, ax = plt.subplots(figsize=(8, 6.5))
12 sns.heatmap(corr, annot=True, fmt=".2f", cmap="vlag",
13             vmin=-1, vmax=1, ax=ax)
14 ax.set_title("Chemistry panel correlation matrix (n=1200)")
15 fig.tight_layout()
16 fig.savefig("chem_heatmap.png", dpi=300)
```

شيفرة 5.9: الخريطة الحرارية لارتباطات لوحة الكيمياء السريرية

	glucose	hba1c	creat	urea	chol	ldl
glucose	1.00	0.85	-0.02	-0.04	0.21	0.14
hba1c	0.85	1.00	-0.03	-0.03	0.21	0.16
creat	-0.02	-0.03	1.00	0.65	0.02	0.06
urea	-0.04	-0.03	0.65	1.00	-0.01	0.02
chol	0.21	0.21	0.02	-0.01	1.00	0.90
ldl	0.14	0.16	0.06	0.02	0.90	1.00

تبرز ثلاث كتل دافئة، وكل واحدة منها فيزيولوجيا خالصة: الغلوكوز مع السكر التراكمي (HbA1c 0.85) — متوسط السكر لثلاثة أشهر يتعقب سكر الصيام)، والكرياتينين مع اليوريا (0.65) — كلاهما مؤشر تصفية كلوية)، والكوليسترول الكلي مع LDL (0.90) — إذ LDL هو الجزء الغالب). أما الخريطة الحرارية التي تخالف الفيزيولوجيا فهي أول علامة على عمود موسوم خطأ أو مشكلة في جهاز — فالخريطة الحرارية أداة ضبط جودة بقدر ما هي أداة بحث.

5.3 التحدي السريري (30 دقيقة)

التحدي السريري — لوحة الترصد الأسبوعية

يريد المدير العام صفحة واحدة تُفتح بها إجازات يوم الاثنين: العبء الاستقبالي والوضع الوبائي، أربع لوحات، شكل واحد، جودة طباعة. انطلاقاً من مجموعات بيانات الدورة، أمامك 20 دقيقة لبناء الشكل؛ أما النتائج الرئيسة الثلاث فتكتب أثناء النقاش الختامي الذي يليه.

1. حمل `chemistry_panel.csv` و `patients_demographics.csv` و `rtPCR_results.csv` وادمج الكيمياء مع الديموغرافيا على `patient_id`.
2. أنشئ شكلاً 2×2 بواسطة `plt.subplots(2, 2, figsize=(12, 9))`.
3. اللوحة A: مدرج تكراري لعمود `glucose_mgdl` مع خط `axvline` أحمر متقطع عند 126 ملغ/دل؛ واطبع عدد ونسبة القيم عند الخط أو فوقه.
4. اللوحة B: مخطط صندوق لعمود `hba1c_pct` حسب الجنس مع `axhline` متقطع عند 6.5%؛ واطبع وسيط السكر التراكمي لكل جنس.

5. اللوحة C: منحني الإيجابية الأسبوعية لفحص RT-PCR من الخطوة 6؛ واطبع أسبوع الذروة وعدد حالاته.
6. اللوحة D: مخطط انتشار للعمر مقابل glucose_mgdl مع alpha وخط 126 ملغ/دل؛ واطبع ارتباط العمر بالغلوكوز.
7. أضف fig.suptitle() وfig.tight_layout() وصدر إلى dashboard.png بواسطة fig.savefig مع dpi=300.
8. اكتب النتائج الرئيسة الثلاث كما لو كنت ستقولها شفاهاً للمدير العام — الأرقام أولاً، والمصطلحات أخيراً.

✓ حل المدرب — البناء الكامل للوحة الترصّد

كل لوحة تعيد استخدام نمط من التطبيق الموجّه؛ والمهارة الجديدة الوحيدة هي فهرسة شبكة المحاور التي تعيدها plt.subplots(2, 2)

```

1 import pandas as pd
2 import matplotlib.pyplot as plt
3
4 chem = pd.read_csv("chemistry_panel.csv")
5 demo = pd.read_csv("patients_demographics.csv")
6 pcr = pd.read_csv("rtPCR_results.csv", parse_dates=["collection_date"])
7
8 labs = chem.merge(demo[["patient_id", "age", "sex"]], on="patient_id")
9 pos = pcr[pcr["result"] == "Positive"]
10 week = pos["collection_date"].dt.isocalendar().week
11 weekly = pos.groupby(week)["sample_id"].count()
12
13 n_high = (labs["glucose_mgdl"] >= 126).sum()
14 print("High glucose (>=126):", n_high, "of 1200 =",
15       round(100 * n_high / 1200, 1), "%")
16 print(labs.groupby("sex")["hba1c_pct"].median())
17 print("Peak week:", weekly.idxmax(), "with", weekly.max(), "positives")
18 print("Corr age vs glucose:",
19       labs["age"].corr(labs["glucose_mgdl"]).round(2))
20
21 m_h = labs[labs["sex"] == "M"]["hba1c_pct"]
22 f_h = labs[labs["sex"] == "F"]["hba1c_pct"]
23
24 fig, axes = plt.subplots(2, 2, figsize=(12, 9))
25
26 axes[0, 0].hist(labs["glucose_mgdl"], bins=30, color="teal",
27                edgecolor="white")
28 axes[0, 0].axvline(126, linestyle="--", color="firebrick", linewidth=2)
29 axes[0, 0].set_title("A. Fasting glucose (n=1200)")
30 axes[0, 0].set_xlabel("Glucose (mg/dL)")
31 axes[0, 0].set_ylabel("Patients")
32
33 axes[0, 1].boxplot([m_h, f_h], tick_labels=["Male", "Female"])
34 axes[0, 1].axhline(6.5, linestyle="--", color="firebrick")
35 axes[0, 1].set_title("B. HbA1c by sex")
36 axes[0, 1].set_ylabel("HbA1c (%)")
37
38 axes[1, 0].plot(weekly.index, weekly.values, marker="o",
39               color="purple")
40 axes[1, 0].set_title("C. Weekly RT-PCR positives")
41 axes[1, 0].set_xlabel("ISO week 2025")
42 axes[1, 0].set_ylabel("Positive samples")
43
44 axes[1, 1].scatter(labs["age"], labs["glucose_mgdl"], alpha=0.35,
```

```

45         s=14, color="steelblue")
46 axes[1, 1].axhline(126, linestyle="--", color="firebrick")
47 axes[1, 1].set_title("D. Age vs glucose")
48 axes[1, 1].set_xlabel("Age (years)")
49 axes[1, 1].set_ylabel("Glucose (mg/dL)")
50
51 fig.suptitle("Diagnostics Directorate weekly surveillance dashboard",
52             fontsize=14)
53 fig.tight_layout()
54 fig.savefig("dashboard.png", dpi=300)

```

شيفرة 5.10: لوحة المديرية ذات اللوحات الأربع

```

High glucose (>=126): 191 of 1200 = 15.9 %
sex
F      5.5
M      5.5
Name: hba1c_pct, dtype: float64
Peak week: 3 with 12 positives
Corr age vs glucose: 0.01

```

التفسير السريري — النتائج الرئيسية الثلاث. (1) العبء الاستقلابي: لدى 191 من أصل 1200 مريض مفحوص (15.9%) غلوكوز صائم في النطاق السكري (≥ 126 ملغ/دل)؛ فخطّط لطاقة الفحص التأكيدي والإحالة لمريض واحد من كل ستة. (2) لا فجوة بين الجنسين في الضبط السكري: وسيط السكر التراكمي HbA1c هو 5.5% لدى الجنسين كليهما — فوزّع الفحص حسب عوامل الخطورة لا حسب الجنس؛ والمحور الأيمن في اللوحة B يمنع القصة الزائفة التي كان سيرونها مخطط أعمدة مبتور. (3) الوضع الوبائي مستقر: بلغت الإيجابيات الأسبوعية ذروتها عند 12 (أسبوع ISO الثالث) مقابل متوسط قدره 6.9 من دون ارتفاع مستمر؛ فالطاقة الاستيعابية كافية، وبصبح هذا النطاق عتبة الإنذار للموسم المقبل. وتضيف اللوحة D نتيجة سلبية جديرة بالإبلاغ: العمر والغلوكوز غير مرتبطين هنا ($r = 0.01$)، ولذلك لا سند في بياناتنا نحن لفحص الغلوكوز الموجه حسب العمر — فلوحات المتابعة وُجدت لتقتل الافتراضات بقدر ما وُجدت لتأكيدهما.

نقاش جماعي.

1. تعرض اللوحة A نسبة 15.9% عند 126 ملغ/دل أو فوقه — لكن هذه حالات دخول مستشفى لا عينة سكانية. كيف تكتب تسميتها التوضيحية كي لا تقرأها الإدارة على أنها معدل انتشار وطني؟
2. المدرج التكراري لقيم Ct والمنحنى الوبائي يأتیان من الملف نفسه. أي قرار يقود كل منهما، وأي زميل يحتاج إلى أيهما؟
3. سنُستنسخ لوحة المتابعة بتصوير رمادي التدرج. أي اللوحات تنجو، وما الذي يصلح البقية (أنماط الخطوط، التهشير، العلامات)؟

5.4 الواجب المنزلي والتقييم الذاتي

تمارين الواجب المنزلي.

1. التدخين وكريات الدم البيضاء. ادمج cbc_results.csv مع patients_demographics.csv وركّب مدرجين تكراريين متراكبين لعدد كريات الدم البيضاء — المدخنون مقابل غير المدخنين — مع $\alpha=0.6$ ومفتاح رموز ومحاور موسومة. هل يميل المدخنون إلى قيم أعلى؟
2. ألم الصدر والكوليسترول. من heart_risk.csv ارسم مخططات صندوقية لعمود cholesterol لكل نوع من chest_pain_type باستدعاء واحد لـ df.boxplot() بوسيطه column و by. أي نوع يحمل أعلى وسيط؟
3. المنحنى الشهري. أعد بناء المنحنى الوبائي بفئات شهرية: طبق dt.to_period() بتردد شهري على collection_date وجمّع حسب الناتج. ما الذي يظهره المنحنى الأسبوعي ويخفيه الشهري؟
4. مطاردة المتنبئات. احسب مصفوفة الارتباط الكاملة لملف diabetes_dataset.csv وارسمها خريطة حرارية بمعاملات مكتوبة. أي متغير منفرد يرتبط أقوى ارتباطاً بعمود outcome؟ احتفظ بالشكل — فستندس منه ميزات غداً.

اختبار التحقق الذاتي.

1. أي مخطط يجب عن سؤال كيف يتوزع محلل واحد، وأياها يجب عن سؤال هل تختلف مجموعتان؟
2. في المخطط الصندوقي، ماذا تمثل حافتا الصندوق والخط الأوسط والنقاط المعزولة؟
3. لماذا يحتاج مخطط انتشار ل768 مريضاً إلى α ، وما العيب الذي يظهر من دونه؟
4. مدرج تكراري لعدّ الغلوكوز يبدأ محوره الصادي من 40 بدلاً من 0. ما الانطباع الذي يخلقه هذا، وما قاعدة الدار؟
5. منحناك الوبائي يمتد من كانون الأول 2025 إلى شباط 2026. لماذا يكون `dt.isocalendar().week` وحده غير آمن، وما البديل الآمن؟

إجابات الاختبار. (1) المدرج التكراري لتوزيع محلل واحد؛ والمخطط الصندوقي لمقارنة المجموعات (وكشف القيم الشاذة). (2) حافتا الصندوق هما المئينان 25 و 75 (المدى الربيعي)، والخط الأوسط هو الوسيط، والنقاط المعزولة هي القيم الشاذة الواقعة خلف الشاربين. (3) النقاط تتراكب (التراكب المفرط)؛ ومن دون الشفافية تبدو المناطق الكثيفة والمتفرقة متطابقتين. (4) البتر يضخم الفروق بين الأعمدة؛ فمحاور العدّ تبدأ من الصفر وأي تكبير يُصرّح به في التسمية التوضيحية. (5) أرقام أسابيع ISO تُستأنف مع السنة الجديدة، فيندمج الأسبوع 1 من 2026 مع الأسبوع 1 من 2025؛ جُمع حسب فترات أسبوعية من `dt.to_period` (أو حسب سنة ISO والأسبوع معاً) بدلاً من ذلك.

♣ القيمة لمديرية التشخيص — من غرفة البيانات إلى قاعة الإيجاز

اليوم تكتسب المديرية طبعها التواصلية. أنتجت الأيام 4-1 جداول لا يقرؤها إلا المحللون؛ أما أشكال اليوم فيقرؤها المدبرون والأطباء والوزراء. وبصورة ملموسة: المنحنى الوبائي الأسبوعي الذي كان يستهلك يوماً من عمل الجداول الإلكترونية صار سطرين من الشيفرة على التصدير المختبري الخام، قابلين لإعادة الإنتاج كل يوم اثنين بمحل واحد؛ وخطوط قيم العتبة تُرسم برمجياً، فتحمّل أشكال كل مستشفى العتبات السريرية نفسها؛ والتصدير بدقة 300 dpi يفي بمعايير النشر من دون قسم رسوم. وقواعد المحاور الأمنية تحمي المصادقية: فالشكل الوزاري الذي لا يستطيع أن يضلّل لا يمكن أن يُرفض.

خلاصات اليوم

- السؤال السريري هو الذي يختار المخطط: المدرج التكراري لتوزيع واحد، والصندوق للمجموعات، والخطي للزمن، والانتشار لمحلّين، والخريطة الحرارية لارتباطات كثيرة.
 - `fig, ax = plt.subplots()` هو الهيكل العظمي الشامل؛ والتسميات مع الوحدات والعنوان ومفتاح الرموز و `dpi=300` هي قائمة تدقيق النشر غير القابلة للتفاوض.
 - المجالات المرجعية هي 95% المركزية من توزيع أصحاء — فالمدرج التكراري هو مصدر الطبيعى، و5% من نتائج الأصحاء تقع خارجه بحكم التصميم.
 - المدرجات التكرارية المتراكبة مع α كشفت مجموعتي هيموغلوبين (متوسط الإناث 13.1 والذكور 14.6 غ/دل)؛ والمخططات الصندوقية وضعت وسيط مرضى السكري (136 ملغ/دل) فوق خط 126 لكنها أظهرت التداخل الذي يهزم التشخيص بقيمة واحدة.
 - المنحنى الوبائي هو `groupby` واحد على موصّل تواريخ — والتجميع حسب فترات أسبوعية من `dt.to_period` هو الصيغة الآمنة عبر السنوات.
 - استعادت الخريطة الحرارية للكيمياء فيزيولوجيا معروفة (الغلوكوز مع HbA1c 0.85، واليوريا مع الكرياتينين 0.65، والكوليسترول مع LDL 0.90)؛ وعندما تناقض خريطة حرارية الفيزيولوجيا، دقّق البيانات.
 - المحاور الأمنية واجب مؤسسي — والنتائج السلبية (ارتباط العمر بالغلوكوز $r = 0.01$) هي نتائج أيضاً.
 - `plt.subplots(2, 2)` تحوّل أربعة تحليلات إلى صفحة إيجاز واحدة — فلوحة المتابعة هي صوت المديرية الأسبوعي.
- غداً، في الفصل 6، تتحول الأنماط التي وجدتها عيناك اليوم — تدرّج الغلوكوز مع مؤشر كتلة الجسم، والصناديق المتداخلة — إلى ميزات مهندسة: أعمدة خام تُقسّم طبقيّاً وتُركّب وتُرمّز في متغيرات يستطيع نموذج تنبؤي أن يتعلم منها فعلاً.

باب 6

اليوم 6 — هندسة الميزات للطب التنبؤي

يُكتب التقرير المختبري للطبيب لا للخوارزمية. فالطبيب السريري الذي يقرأ «سكر تراكمي HbA1c بنسبة 7.4%، مدخن، بعمر 71 عاماً» يدرك في الحال أن هذا المريض ينتهي إلى رأس قائمة متابعة مرضى السكري. أما نموذج تعلم الآلة فلا يرى شيئاً من ذلك: إنه يستقبل أعمدة من الأرقام، ولا يتعلم إلا من شكل تلك الأرقام. وحرفة هندسة الميزات (Feature Engineering) بأكملها هي خطوة الترجمة الواقعة بينهما — أخذ الاستدلال التشخيصي الذي يسكن رأس الاختصاصي وترميزه، سطرًا بسطر، في أعمدة جديدة تستطيع الخوارزمية أن تنجزها. اليوم نبني هذا الجسر على بيانات مديرتنا نفسها: سندمج البيانات الديموغرافية مع لوحة الكيمياء السريرية، ونشتق نسبة الكوليسترول التي يستعملها أطباء القلب فعلاً، ونحوّل قيم سكر الدم إلى فئات التشخيص المعتمدة من الجمعية الأمريكية للسكري (ADA) بدقة تامة، ونحوّل الفئات النصية إلى رموز رقمية، ونختتم بدرجة خطورة شفاقة قابلة للتدقيق تُنتج قائمة عمل مرتّبة للمتابعة. وغداً، في الفصل 7، تصبح هذه الأعمدة المهندسة المدخل المباشر لأولى نماذج تعلم الآلة لدينا — فعمل اليوم ليس تمريناً، بل هو الأساس.

الدقائق	الجزء	المحتوى والأنشطة
00-05	المراجعة والتأطير	مراجعة اليوم الخامس (التمثيل البياني للعلامات الحيوية): ولماذا تحتاج النماذج إلى إشارات رقمية مهندسة
05-12	الإطار النظري الأول	ما هي الميزة: الأعمدة المشتقة: النسب السريرية بوصفها عمليات حسابية موجهة
12-20	الإطار النظري الثاني	العلامات باستخدام np.where: الفئات باستخدام pd.cut: الترميز؛ درجات الخطورة المركبة
20-28	التطبيق الموجه الأول	دمج البيانات الديموغرافية مع الكيمياء: بناء نسبة الكوليسترول TC/HDL
28-38	التطبيق الموجه الثاني	علامة فرط سكر الدم: فئات HbA1c التشخيصية: فحص القيم الحدية
38-48	التطبيق الموجه الثالث	الترميز بالقواميس مع map: الفئات العمرية: ميزات مشتقة من التواريخ
48-60	التطبيق الموجه الرابع	الترميز الأحادي الفعال مع pd.get_dummies: درجة الخطورة: قائمة أعلى 20 مريضاً
60-80	التحدي السريري	هندسة heart_risk.csv: الترميز، ودرجة الأولوية القلبية، وفئات الفرز
80-90	الخلاصة والواجب المنزلي	عرض الحل خطوة بخطوة: أسئلة النقاش: لمحة عن نمذجة اليوم السابع

✓ الأهداف التعليمية

بنهاية هذه الجلسة ستكون قادراً على:

1. أن تشرح، بلغة سريرية، لماذا تتطلب نماذج تعلم الآلة أعمدة إدخال رقمية حسنة البناء، وما الذي تضيفه هندسة الميزات إلى الدقة التشخيصية.
2. أن تشتق متغيرات سريرية جديدة من المُحلّلات الخام بعمليات حسابية موجهة (نسب مثل الكوليسترول الكلي إلى HDL) — دون حلقة تكرارية واحدة.
3. أن تنشئ علامات تشخيصية ثنائية باستخدام np.where وبحُدود فاصلة صحيحة (الغلوكوز الصائم ≥ 126 ملغ/دل).
4. أن تقسم مُحلّلاً متصلاً إلى فئات سريرية مرتبة باستخدام pd.cut، مستنسخاً فئات HbA1c المعتمدة من ADA بدقة، بما في ذلك القيمة الحدية 6.5.

5. أن ترمز الفئات الثنائية/الترتيبية باستخدام map والفئات الاسمية باستخدام pd.get_dummies، وأن تبرر أي أداة تناسب أي متغير.
6. أن تجمع العلامات المهندسة في درجة خطورة مركبة موزونة قابلة للتدقيق بالكامل، وأن تُنتج قائمة عمل مرتبة للمرضى.
7. أن تسلم نسخة مهندسة جاهزة للنمذجة من ملف heart_risk.csv لجلسة تعلم الآلة غداً.

➤ حقبة أدوات الذكاء الاصطناعي والاعتماديات — مكتبات هندسة الميزات

لا يحتاج اليوم سوى الأدوات الأساسيتين في الدورة — لا تثبيتات جديدة.

```
1 import pandas as pd
2 import numpy as np
3
4 pd.set_option("display.width", 100)
```

شيفرة 6.1: استيرادات اليوم السادس

```
pip install pandas numpy
```

توفر مكتبة **pandas** الدوال merge وmap وpd.cut وpd.get_dummies — الأدوات الأربع التي تحوّل التصديرات المخترية الخام إلى جداول جاهزة للنمذجة. وهذا يعني للمديرة أن نصاً برمجياً واحداً قابلاً لإعادة الإنتاج يحل محل ساعات من إعادة الترميز اليدوي في «إكسل» في كل مرة يسلم فيها مختبر إحدى المحافظات دفعة جديدة. وتوفر مكتبة **NumPy** الدالة np.where، وهي الشرط الموجه الذي يختم علامة تشخيصية على 1,200 مريض في سطر واحد. ولأنها شفرة C مجمعة في العمق، فإن الشيفرة نفسها ستتوسع دون تغيير إلى السجل الوطني بمئات الآلاف من الصفوف.

6.1 السياق السريري والإطار النظري (20 دقيقة)

كل نموذج تنبؤي ستشره الوزارة يوماً — خطورة إعادة القبول، وإنذارات الإلتان، وتحري مضاعفات السكري — يستهلك جدولاً مستطيلاً من الأرقام. النموذج لا يعرف الطب؛ إنه يعرف الأعمدة. فإذا كانت المعلومة الحاسمة سريرياً مدفونة داخل عنوان نصي، أو سلسلة تاريخ، أو زوج من المُحلّلات تهم نسبتهما أكثر من أي من القيمتين منفردة، فإن النموذج ببساطة لا يستطيع رؤيتها. هندسة الميزات هي الطريقة التي نجعل بها غير المرئي مرئياً.

الميزات معرفة سريرية مرمّزة في أعمدة

الميزة هي عمود إدخال واحد يُقدّم إلى النموذج. الأعمدة المخترية الخام قياسات؛ أما الميزات المهندسة فهي أحكام مُعبّر عنها رقمياً. فالبارة «قيمة HbA1c البالغة 6.5% فأكثر تشخيصية للسكري» تصبح عموداً من 0/1. والعبارة «تزداد الخطورة القلبية الوعائية مع نسبة الكوليسترول الكلي إلى HDL» تصبح عموداً ناتجاً قسمةً مشتقاً. النموذج المدرب على ميزات حسنة الهندسة يبدأ وفي مدخلاته عقود من العلم السريري مبنية سلفاً، ولهذا فإن جودة الميزات تفوق في أهميتها عادةً اختيار الخوارزمية. والعكس صحيح أيضاً: لا تستطيع أي خوارزمية استرجاع إشارة لم يحتوها جدول الميزات أصلاً.

الاشتقاق الموجه — العمليات الحسابية والعلامات والفئات

جميع تحويلات اليوم موجهة (Vectorized): تعبير واحد يعمل على عمود كامل دفعة واحدة، دون حلقات تكرارية ودون شروط صفاً بصف. ثلاثة أنماط تغطي كل اشتقاق سريري تقريباً: العمليات الحسابية بين الأعمدة تنشئ النسب والمؤشرات (/ total_chol، hdl، ومؤشر كتلة الجسم من الوزن والطول)؛ و np.where(condition, 1, 0) يحوّل عتبة واحدة إلى علامة ثنائية؛ و pd.cut يحوّل مُحلّلاً متصلاً إلى مجموعة مرتبة من الفئات السريرية عبر التقسيم الطيفي (Binning) بتحديد حواف الفئات وتسمياتها. ومع الخيار right=False تكون كل فترة مغلقة من جهة اليسار — [0, 5.7) و(5.7, 6.5] و(6.5, 20) — بحيث تقع القيمة 6.5 بالضبط في فئة السكري، تماماً كما يقتضي معيار ADA.

لماذا تحتاج النماذج إلى أرقام — قاعدتا الترميز

تعمل مقدرات scikit-learn على مصفوفات رقمية؛ والسلسلة النصية «Internal Medicine» لا معنى لها عندها. قاعدتان تغطيان التحويل عبر الترميز (Encoding). القاعدة 1 — الفئات الثنائية أو الترتيبية (الجنس، التدخين، نعم/لا، خفيف/متوسط/شديد) تُحوّل إلى أعداد صحيحة باستخدام `map({dictionary})`، لأن ترتيبها أو طبيعتها ذات المستويين يجعلان عموداً رقمياً واحداً أميناً لها. القاعدة 2 — الفئات الاسمية (القسم، المحافظة، نمط ألم الصدر) لا ترتب لها: فترميز `Emergency=1` و `ICU=2` و `Surgery=3` سيخبر النموذج زوراً بأن الجراحة تساوي «ثلاثة أضعاف» الطوارئ. وبدلاً من ذلك يوسّع `pd.get_dummies` العمود إلى مؤشر 0/1 واحد لكل فئة (الترميز الأحادي الفعّال One-Hot)، فيتيح للنموذج وزن كل قسم على حدة.

درجات الخطورة المركبة — الأداة الشفافة السابقة للذكاء الاصطناعي

المجموع الموزون للعلامات المهندسة، $score = w_1f_1 + w_2f_2 + \dots + w_kf_k$ ، هو أقدم أدوات دعم القرار في الطب — فمقاييس APACHE و CHA₂DS₂-VASc و Wells كلها هذا بالضبط. ويستحق الأمر بناء واحدة منها قبل أي تعلم آلي لسببين. الأول أنها قابلة للتدقيق بالكامل: يستطيع أي طبيب إعادة حساب درجة المريض يدوياً من جدول الأوزان المنشور، وهو أمر بالغ الأهمية لحكومة الوزارة والمراجعة الطبية القانونية. والثاني أنها السلف المفاهيمي للانحدار اللوجستي (Logistic Regression) الذي نلقاه غداً: فالنموذج المدرب لا يفعل أكثر من تعلّم الأوزان من البيانات بدلاً من أخذها من لجنة. وفهم الدرجة المبينة يدوياً اليوم يجعل النموذج المتعلّم شفافاً غداً.

★ لؤلؤة سريرية — حدود سكر الدم والشحوم المستخدمة اليوم

معايير التشخيص لدى الجمعية الأمريكية للسكري (البالغون): غلوكوز البلازما الصائم ≥ 126 ملغ/دل يدل على السكري، و ≥ 100 ملغ/دل على اختلال الغلوكوز الصائم؛ وقيمة HbA1c الأدنى من 5.7% طبيعية، و 5.7–6.4% ما قبل السكري، و $\geq 6.5\%$ سكري. وفي الشحوم، تُعد نسبة الكوليسترول الكلي إلى HDL (TC/HDL) منبئاً قلوبياً وعائياً أقوى من أي من المُحلّلين منفرداً: فما دون 3.5 أمثل، و ≥ 5 يشير إلى خطورة مرتفعة تستدعي الإحالة إلى عيادة الشحوم. والكوليسترول الكلي المرغوب أدنى من 200 ملغ/دل؛ و ≥ 240 ملغ/دل مرتفع صراحة. هذه هي العتبات الدقيقة التي سترمزها شيفرتنا — فالميزة المبينة على حد فاصل خاطئ هي خطأ تشخيصي منهجي يُطبّق على كل صف.

يلخّص الجدول 6.1 غُدّة اليوم في بطاقة مرجعية واحدة: كل تحويل نعلّمه، والأداة الموجهة التي تنفذه، ومثال من واقع المديرية.

جدول 6.1: غُدّة هندسة الميزات لليوم السادس

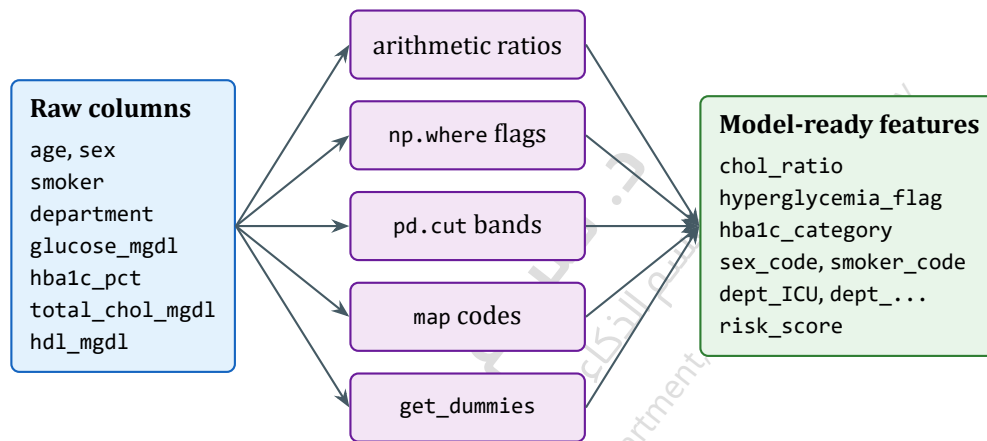
العملية	الأداة	مثال سريري
نسبة مشتقة	العمليات الحسابية بين الأعمدة	TC/HDL، اليوريا إلى الكرياتينين
علامة ثنائية	<code>np.where</code>	الغلوكوز ≥ 126 ملغ/دل
فئات مرتبة	<code>pd.cut</code>	Normal/Pre-diab./Diab. HbA1c
ترميز ثنائي/ترتيبي	<code>map + قاموس</code>	الجنس، التدخين، إعادة القبول
توسيع اسمي	<code>pd.get_dummies</code>	القسم، المحافظة
ميزات التواريخ	الموصل <code>.dt</code>	شهر القبول، يوم الأسبوع
درجة مركبة	مجموع موزون للعلامات	قائمة عمل الخطورة الاستقلالية

تستخدم الدرجة الاستقلالية التي سنبنها في التطبيق الموجه الأوزان الواردة في الجدول 6.2. لاحظ أن الأوزان أرقام مدوّرة بسيطة عمداً وقابلة للدفاع عنها سريرياً، اتّفق عليها قبل النظر في النتائج — وهذا ما يجعل الأداة قابلة للتدقيق.

جدول 6.2: درجة الخطورة الاستقلابية للمديرية — جدول الأوزان

المعيار	التعريف	الوزن
HbA1c في المدى السكري	$\geq 6.5\%$ HbA1c	+2.0
مدخن حالياً	smoker = Yes	+1.5
عمر متقدم	العمر ≥ 65 عاماً	+1.0
شحوم مُعَصَّدة	نسبة TC/HDL ≥ 5	+1.0
الحد الأقصى الممكن		5.5

يعرض الشكل 6.1 الجلسة كلها في مخطط واحد: أعمدة التصدير الخام على اليسار، والعمليات الموجهة الخمس في الوسط، وكتلة الميزات الجاهزة للنمذجة التي سيستهلكها اليوم السابع على اليمين.



شكل 6.1: مسار هندسة الميزات: تُحوّل أعمدة التصدير المختبري الخام بخمس عمليات موجهة إلى كتلة الميزات الرقمية التي تستهلكها نماذج تعلم الآلة.

6.2 التطبيق الموجّه (40 دقيقة)

■ عرض تطبيقي مباشر — من تصدير الكيمياء الخام إلى قائمة عمل استقلابية مرتبة

الهدف: انطلاقاً من patients_demographics.csv و chemistry_panel.csv، نهندس — دون أي حلقات إطلاقاً — نسبة الكوليسترول، وعلامة فرط سكر الدم، وفئات HbA1c وفق ADA، وترميّز الجنس والتدخين، والفئات العمرية، وميزات التواريخ، والأقسام بالترميز الأحادي الفعال، ودرجة الخطورة الاستقلابية الموزونة من الجدول 6.2، وننتهي بقائمة عمل متابعة لأعلى 20 مريضاً بمعرفات المرضى الوطنية. يكتب المتدربون الشيفرة تزامناً مع المدرب؛ وكل مخرج أدناه مخرج حقيقي على بيانات الدورة.

الخطوة 1: دمج البيانات الديموغرافية مع لوحة الكيمياء

كثيراً ما تقيم الميزات في ملفات مختلفة: العمر وحالة التدخين في سجل القبول، والغلوكوز والشحوم في التصدير المختبري. والنموذج يحتاجها جنباً إلى جنب، لذا ننفذ الدمج (merge) على المفتاح المشترك patient_id تماماً كما في اليوم الرابع. يحتفظ الدمج الداخلي بالمرضى الـ 1,200 الموجودين في الجدولين معاً ويوسّع الصف إلى 21 عموداً.

```

1 import pandas as pd
2 import numpy as np
3
4 pd.set_option("display.width", 100)
5
6 demo = pd.read_csv("patients_demographics.csv")
7 chem = pd.read_csv("chemistry_panel.csv")
  
```

```

8 df = demo.merge(chem, on="patient_id", how="inner")
9 print(df.shape)
10 print(df[["patient_id", "age", "sex", "glucose_mgdl", "hba1c_pct"]].head())

```

شيفرة 6.2: دمج البيانات الديموغرافية مع لوحة الكيمياء

```

(1200, 21)
  patient_id  age sex  glucose_mgdl  hba1c_pct
0  MOH-00001   53  M         147         7.4
1  MOH-00002   23  M         124         7.2
2  MOH-00003   63  M          55         4.4
3  MOH-00004   67  M         168         6.4
4  MOH-00005    3  F         122         6.5

```

الخطوة 2: نسبة مشتقة — الكوليسترول الكلي إلى HDL

أولى ميزتنا المهندسة عملية حسابية خالصة. قسمه عمود على آخر تقسمهما صفافاً بصف عبر المرضى الـ 1,200 دفعة واحدة — هذه هي المعالجة الموجهة (Vectorization) التي تدربنا عليها في اليوم الثالث، وقد صارت الآن في خدمة مؤشر قلبي وعائي معترف به. وسيط النسبة البالغ 3.92 يقع في المدى المقبول، لكن القيمة القصوى 11.10 تعود إلى مريض ينبغي أن تلتقيه عيادة الشحوم.

```

1 df["chol_ratio"] = (df["total_chol_mgdl"] / df["hdl_mgdl"]).round(2)
2 print(df[["patient_id", "total_chol_mgdl", "hdl_mgdl", "chol_ratio"]].head())
3 print(df["chol_ratio"].describe().round(2))

```

شيفرة 6.3: نسبة الكوليسترول TC/HDL

```

  patient_id  total_chol_mgdl  hdl_mgdl  chol_ratio
0  MOH-00001         161         45         3.58
1  MOH-00002         214         40         5.35
2  MOH-00003         234         44         5.32
3  MOH-00004         165         55         3.00
4  MOH-00005         158         46         3.43
count      1200.00
mean         4.07
std          1.28
min          1.39
25%          3.15
50%          3.92
75%          4.79
max          11.10
Name: chol_ratio, dtype: float64

```

الخطوة 3: علامة تشخيصية ثنائية باستخدام np.where

عتبة واحدة، سطر واحد. تقيّم `np.where(condition, 1, 0)` الشرط لكل الصفوف في آن واحد وتعيد 1 حيثما تحقق. نطبق معيار الغلوكوز الصائم لدى ADA: يحمل 191 مريضاً من أصل 1,200 (15.9%) غلوكوزاً في المدى السكري، وتؤكد متوسطات المجموعتين أن العلامة تفصل المجتمع فصلاً نظيفاً (140.7 مقابل 89.0 ملغ/دل).

```

1 df["hyperglycemia_flag"] = np.where(df["glucose_mgdl"] >= 126, 1, 0)
2 print(df["hyperglycemia_flag"].value_counts())
3 print(df.groupby("hyperglycemia_flag")["glucose_mgdl"].mean().round(1))

```

شيفرة 6.4: علامة فرط سكر الدم عند الحد الفاصل 126 ملغ/دل

```

hyperglycemia_flag
0    1009
1     191
Name: count, dtype: int64
hyperglycemia_flag
0     89.0
1    140.7
Name: glucose_mgdl, dtype: float64

```

الخطوة 4: فئات سريرية مرتبة باستخدام pd.cut

حين يعرف الدليل الإرشادي ثلاث فئات، لا تكفي علامة ثنائية. تأخذ `pd.cut` عمود `HbA1c` المتصل وقائمة حواف الفئات وقائمة التسميات، وتُسند كل مريض إلى فئته. ويجعل الخيار `right=False` كل فترة مغلقة من جهة يسارها، فتقع 6.5 في (6.5, 20] — فئة السكري — تماماً كما يقتضي معيار `HbA1c ADA` ($\geq 6.5\%$). وفحص القيم الحدية يثبت ذلك على صفوف حقيقية.

```

1 bands = [0, 5.7, 6.5, 20]
2 names = ["Normal", "Pre-diabetic", "Diabetic"]
3 df["hba1c_category"] = pd.cut(df["hba1c_pct"], bins=bands,
4                               labels=names, right=False)
5 print(df["hba1c_category"].value_counts())
6 mask = df["hba1c_pct"].between(6.4, 6.6)
7 print(df.loc[mask, ["hba1c_pct", "hba1c_category"]].head(3))

```

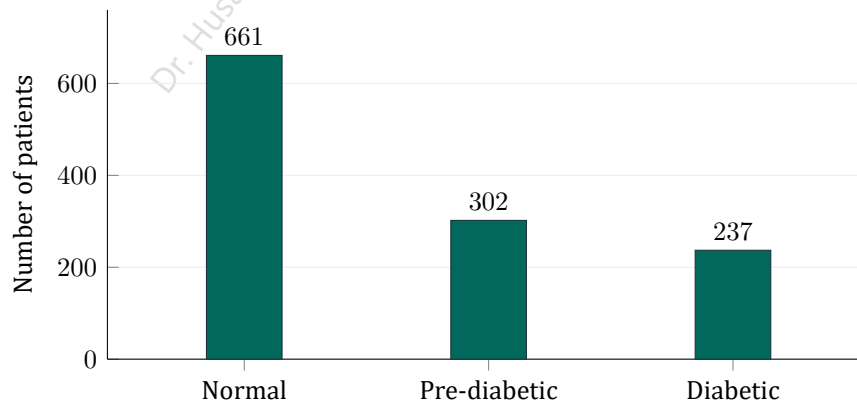
شيفرة 6.5: فئات `HbA1c` وفق `ADA` باستخدام `pd.cut`

```

hba1c_category
Normal        661
Pre-diabetic  302
Diabetic      237
Name: count, dtype: int64
   hba1c_pct hba1c_category
3         6.4   Pre-diabetic
4         6.5     Diabetic
26        6.5     Diabetic

```

هذا التوزيع يستحق رسماً بيانياً تضعه المديرية في إحاطة رسمية: يرسم الشكل 6.2 الفئات الثلاث بأعدادها الحقيقية. فقرابة مريض من كل خمسة في هذه المجموعة يقع في فئة السكري، وربع آخر في فئة ما قبل السكري — حجة لبرنامج تحري في صورة واحدة.



شكل 6.2: فئات `HbA1c` التشخيصية في المجموعة المدموجة (أعداد حقيقية من `pd.cut` على بيانات الدورة): 661 طبيعي، و302 ما قبل سكري، و237 سكري من أصل 1,200 مريض.

الخطوة 5: الترميز بالقاموس باستخدام map

الجنس وحالة التدخين فئتان ذواتا مستويين، لذا يمثل كلاهما عمود رقمي واحد تمثيلاً أميناً. تبحث map عن كل قيمة في قاموس؛ والتحقق من أن أي رمز لم يُعد مفقوداً ليس مجاملة اختيارية — إنه شبكة الأمان التي تلتقط الهجاءات غير المتوقعة.

```
1 df["sex_code"] = df["sex"].map({"M": 1, "F": 0})
2 df["smoker_code"] = df["smoker"].map({"Yes": 1, "No": 0})
3 print(df[["sex", "sex_code", "smoker", "smoker_code"]].head(4))
4 print("Unmapped sex codes:", df["sex_code"].isna().sum())
5 print("Unmapped smoker codes:", df["smoker_code"].isna().sum())
```

شفرة 6.6: الترميز الثنائي للجنس وحالة التدخين

```
sex  sex_code  smoker  smoker_code
0    M         1      No             0
1    M         1      No             0
2    M         1     Yes             1
3    M         1     Yes             1
Unmapped sex codes: 0
Unmapped smoker codes: 0
```

✱ خطأ شائع — الدالة map تفشل بصمت — عُدَّ قيم NaN دائماً

لا تطلق map أي خطأ عند قيمة غائبة عن القاموس — بل تُدرج NaN بهدوء. فعلى التصدير الفوضوي من اليوم الرابع، كان القاموس {M: 1, F: 0} سيترك كل صف يحمل male أو female أو f بمسافة زائدة قيمة NaN؛ وأي نموذج لاحق سيُسقط أولئك المرضى بصمت. القاعدة في نصوص المديرية البرمجية: كل map يتبعها مباشرة فحص .isna().sum()، والجواب المتوقع صفر. فإن لم يكن صفرًا، وخد الرموز أولاً (مهارات اليوم الرابع) — ولا توسع القاموس أبداً على عمياء دون معاينة القيم المخالفة.

الخطوة 6: الفئات العمرية والميزات المشتقة من التواريخ

تخدم pd.cut الديموغرافيا كما تخدم المُحلّلات: فئات الأطفال والشباب ومتوسطي العمر والمسنين تسلك سلوكاً مختلفاً في كل مرض تقريباً، والعمود المُفأى يتيح للتقارير والنماذج معاً استخدام هذه البنية. والتواريخ تخفي ميزات أيضاً — يستخرج الموصل dt. شهر القبول ويوم الأسبوع بسطر واحد لكل منهما، وهما المادة الخام لتحليلات الموسمية والملاكات.

```
1 df["age_band"] = pd.cut(df["age"], bins=[0, 18, 40, 65, 120],
2                          labels=["Child", "Young-adult",
3                                  "Middle-aged", "Elderly"],
4                          right=False)
5 print(df["age_band"].value_counts())
6 adm = pd.to_datetime(df["admission_date"])
7 df["admission_month"] = adm.dt.month
8 df["admission_dow"] = adm.dt.day_name()
9 print(df["admission_dow"].value_counts().head(3))
```

شفرة 6.7: الفئات العمرية وميزات تاريخ القبول

```
age_band
Middle-aged    530
Young-adult    338
Elderly        215
Child          117
Name: count, dtype: int64
admission_dow
Friday         190
```

```
Sunday    181
Monday    181
Name: count, dtype: int64
```

الخطوة 7: الترميز الأحادي الفعّال باستخدام pd.get_dummies

القسم متغير اسمي — لا يوجد ترتيب سريري بين الطوارئ والعناية المركزة والجراحة — لذا نوسّعه إلى ستة أعمدة 0/1 مستقلة. يُبقي الخيار dtype=int المؤشرات رقمية، ويزيل سطر str.replace المسافة التي كان قسم «Internal Medicine» سيتركها داخل اسم عمود. وكل صف يحمل واحداً وحيداً عبر المؤشرات الستة.

```
1 dept = pd.get_dummies(df["department"], prefix="dept", dtype=int)
2 dept.columns = dept.columns.str.replace(" ", "_")
3 df = pd.concat([df, dept], axis=1)
4 print(dept.sum())
5 print(df[["department", "dept_ICU", "dept_Emergency"]].head(4))
```

شيفرة 6.8: مؤشرات أحادية فعّالة للقسم

```
dept_Emergency    296
dept_ICU          130
dept_Internal_Medicine  308
dept_Oncology      112
dept_Pediatrics    152
dept_Surgery       202
dtype: int64
   department  dept_ICU  dept_Emergency
0  Pediatrics         0                0
1   Emergency         0                1
2  Pediatrics         0                0
3         ICU         1                0
```

✖ خطأ شائع — الترتيب الزائف أسوأ من غياب الترتيب

الاختصار المغربي بتطبيق map بقاموس {Emergency: 1, ICU: 2, ...} على عمود اسمي يزرع ادعاءً كميّاً زائفاً في البيانات: فهو يخبر كل نموذج لاحق بأن العناية المركزة ضعف الطوارئ وبأن الجراحة تقع بين ردهتين لا صلة بينهما. وستجهد النماذج الخطية بإخلاص في مواءمة ميل لهذه الخرافة. احتفظ بـ map للفئات الثنائية أو المرتبة حقاً؛ وكل ما عداها يمر عبر pd.get_dummies. (في النماذج الخطية الصّرفة يمكن إسقاط مؤشر واحد بالخيار drop_first=True لتجنب الفائض — ونماذج الأشجار التي نلقاها غداً لا تحتاج ذلك.)

الخطوة 8: درجة الخطورة المركبة وقائمة العمل

والآن الثمرة: المعايير الموزونة الأربعة من الجدول 6.2 في تعبير موجه واحد. يبلغ ثلاثة مرضى الحد الأقصى 5.5 — مسنون سكريون مدخنون بشحوم مُعَصَّدة — والترتيب التنازلي يحوّل الدرجة إلى قائمة عمل متابعة قابلة للتنفيذ بمعرفات وطنية، جاهزة لتتصل بها عيادة السكري.

```
1 df["diabetic_flag"] = np.where(df["hba1c_pct"] >= 6.5, 1, 0)
2 df["senior_flag"] = np.where(df["age"] >= 65, 1, 0)
3 df["chol_flag"] = np.where(df["chol_ratio"] >= 5, 1, 0)
4 df["risk_score"] = (2.0 * df["diabetic_flag"]
5                   + 1.5 * df["smoker_code"]
6                   + 1.0 * df["senior_flag"]
7                   + 1.0 * df["chol_flag"])
8 print(df["risk_score"].value_counts().sort_index().tail(4))
9 worklist = df.sort_values("risk_score", ascending=False)
```

```
10 cols = ["patient_id", "age", "hba1c_pct", "chol_ratio", "risk_score"]
11 print(worklist[cols].head(8).to_string(index=False))
```

شيفرة 6.9: درجة الخطورة الاستقلالية الموزونة وقائمة أعلى 20 مريضاً

```
risk_score
3.5    34
4.0     9
4.5    20
5.5     3
Name: count, dtype: int64
patient_id  age  hba1c_pct  chol_ratio  risk_score
MOH-01054    95      8.1      7.11      5.5
MOH-00313    84      7.0      5.90      5.5
MOH-00339    71      6.9      5.77      5.5
MOH-01069    78      7.0      4.68      4.5
MOH-01106    61      7.7      5.09      4.5
MOH-00031    93      6.7      3.94      4.5
MOH-00881    51      6.8      8.33      4.5
MOH-00081    56      7.1      5.43      4.5
```

في الصف نطبع (20) head لعرض قائمة العمل كاملة؛ والصفوف أعلاه تروي القصة السريرية بالفعل: كل مريض مدرج يقع في مدى HbA1c السكري، ويجمع معظمهم إليه التدخين أو العمر أو نسبة شحوم مُعَصَّدة. ثلاثة وعشرون مريضاً حازوا درجة 4.5 فأكثر — عبء عمل واقعي لأسبوع أول للمرضى متابعة واحدة.

6.3 التحدي السريري (30 دقيقة)

التحدي السريري — قائمة أولويات طب القلب الوقائي

تجري المديرية تجربة رائدة لعيادة طب قلب وقائي، وقد طلبت من وحدتك تهيئة ملف heart_risk.csv (900 إحالة لاختبار الجهد) لغرضين: (أ) غداً يصبح جدول التدريب لأول مصيِّف تعلم آلي في الدورة، لذا يجب أن يكون كل عمود رقمياً؛ (ب) واليوم تحتاج العيادة قائمة فرز مؤقتة مبنية على الأوزان المنشورة من لجنة طب القلب (الجدول 6.3). اعمل على المهام — ويجب أن تكون كل خطوة موجهة، دون حلقات ودون جمل if.

1. حمّل heart_risk.csv: واذكر أبعاده وعاین الأعمدة النصية الثلاثة (sex و chest_pain_type و exercise_angina).
2. رَمِّز sex (M=1, F=0) و exercise_angina (Y=1, N=0) باستخدام map، وتحقق من أن القيم غير المرمَّزة صفر.
3. طبّق الترميز الأحادي الفعّال على chest_pain_type باستخدام pd.get_dummies (بإدئة cp ونوع عدد صحيح)، واذكر عدد الحالات في كل نمط.
4. أنشئ age_band باستخدام pd.cut: دون 45، و 45-59، و 60 فأكثر.
5. ابن priority_score من الجدول 6.3 كتعبير موجه واحد.
6. صنّف المرضى باستخدام pd.cut إلى أولوية منخفضة (0-2) ومتوسطة (3-4) وعالية (5 فأكثر)؛ واذكر الأعداد.
7. اطبع مرضى الأولوية العشرة الأوائل واحفظ الجدول المهندس كاملاً باسم heart_risk_engineered.csv لليوم السابع.

جدول 6.3: أوزان لجنة طب القلب لدرجة الأولوية

المعيار	التعريف	النقاط
عمر متقدم	العمر ≥ 60 عاماً	+2
تظاهر صامت	نمط ألم الصدر ASY	+2
تغير ST إقفاري	انخفاض ST بمقدار ≥ 2 ملم	+2
كوليسترول مرتفع	الكوليسترول الكلي ≥ 240 ملغ/دل	+1
استجابة قلبية ضعيفة	أقصى معدل قلب < 120 نبضة في الدقيقة	+1
فرط سكر دم صائم	fasting_bs = 1	+1
الحد الأقصى الممكن		9

★ لأولوة سريرية — لماذا يسجل ألم الصدر اللاعراضي أعلى النقاط

في سجلات اختبار الجهد يشير الرمز ASY (لاعراضي) إلى مريض أُحيلوا دون ذبحة نمطية — وعلى نحو مفارقٍ يحمل هذا النمط أعلى انتشار لمرض الشرايين التاجية المؤكد، لأن الإقفار الصامت يتظاهر متأخراً وهو ممثلٌ تمثيلاً زائداً لدى السكرين. كما أن انخفاض القطعة ST المُحدث بالجهد بمقدار ≥ 2 ملم وأقصى معدل قلب متدني (< 120 نبضة في الدقيقة، قصور الاستجابة الزمنية القلبية) علامتان تقليديتان في اختبار الجهد على إقفار مهم. وأوزان اللجنة ترمز هذا الاستدلال بعينه — ولهذا فإن الدرجة المهندسة، المبنية دون أي تعلم آلي على الإطلاق، تفصل النتائج فصلاً حاداً منذ الآن، كما يُظهر الحل.

✓ حل المدرب — هندسة جدول طب القلب

الحل الكامل بترتيب المهام؛ والمخرجات نتائج حقيقية على ملف الدورة.

```

1 hr = pd.read_csv("heart_risk.csv")
2 print(hr.shape)
3
4 hr["sex_code"] = hr["sex"].map({"M": 1, "F": 0})
5 hr["angina_code"] = hr["exercise_angina"].map({"Y": 1, "N": 0})
6 print("Unmapped:", hr["sex_code"].isna().sum(),
7       hr["angina_code"].isna().sum())
8
9 cp = pd.get_dummies(hr["chest_pain_type"], prefix="cp", dtype=int)
10 hr = pd.concat([hr, cp], axis=1)
11 print(cp.sum())
12
13 hr["age_band"] = pd.cut(hr["age"], bins=[0, 45, 60, 120],
14                          labels=["Under-45", "45-59", "60-plus"],
15                          right=False)
16
17 hr["priority_score"] = (2 * np.where(hr["age"] >= 60, 1, 0)
18                        + 2 * hr["cp_ASY"]
19                        + 2 * np.where(hr["st_depression"] >= 2, 1, 0)
20                        + 1 * np.where(hr["cholesterol"] >= 240, 1, 0)
21                        + 1 * np.where(hr["max_hr"] < 120, 1, 0)
22                        + 1 * hr["fasting_bs"])
23
24 hr["priority_class"] = pd.cut(hr["priority_score"], bins=[0, 3, 5, 10],
25                               labels=["Low", "Moderate", "High"],
26                               right=False)
27 print(hr["priority_class"].value_counts())
28 print(hr.groupby("priority_class", observed=True)["target"].mean().round(2))
29
30 top10 = hr.sort_values("priority_score", ascending=False).head(10)

```

```

31 cols = ["age", "sex", "chest_pain_type", "st_depression",
32         "cholesterol", "priority_score"]
33 print(top10[cols].to_string())
34
35 hr.to_csv("heart_risk_engineered.csv", index=False)

```

شيفرة 6.10: حل التحدي — الترميز ودرجة الأولوية

```

(900, 10)
Unmapped: 0 0
cp_ASY      421
cp_ATA      150
cp_NAP      253
cp_TA        76
dtype: int64
priority_class
Low          455
Moderate     338
High         107
Name: count, dtype: int64
priority_class
Low          0.33
Moderate     0.47
High         0.73
Name: target, dtype: float64

```

	age	sex	chest_pain_type	st_depression	cholesterol	priority_score
833	77	F	ASY	3.9	241	7
288	72	M	ASY	2.1	271	7
437	61	M	ASY	2.5	263	7
173	77	M	ASY	2.0	222	7
182	60	M	ASY	3.4	175	7
841	66	M	ASY	0.6	291	6
826	63	M	NAP	3.8	248	6
698	60	M	NAP	2.6	288	6
421	70	M	ASY	2.9	237	6
46	60	M	ASY	0.2	247	6

التفسير السريري. يقسم الفرز 900 إحالة إلى 455 أولوية منخفضة و338 متوسطة و107 عالية — وهي طبقة عالية صغيرة بما يكفي لتستوعبها العيادة الرائدة في شهرها الأول. وسطر التحقق هو أهم مخرجات اليوم: يرتفع معدل المرض الملاحظ (target) من 33% في الفئة المنخفضة إلى 47% في المتوسطة و73% في العالية. أداة موزونة يدوياً وقابلة للتدقيق بالكامل بُنيت في خمسة عشر سطرًا تركّز المرض الحقيقي بأكثر من ضعفين منذ الآن — وكل معيار فيها يمكن الدفاع عنه سطرًا بسطر في تدقيق وزاري. غداً نسأل: هل يستطيع نموذج يتعلم أوزانه من البيانات أن يحقق أفضل من ذلك؟ والملف المهندس heart_risk_engineered.csv هو مدخله. لاحظ أن الأيام 7-10 تعيد اشتقاق ميزاتها من الملف الخام heart_risk.csv داخل كل فصل حتى تكون كل جلسة مكتبية بذاتها — فالملف الذي حفظته هنا نسختك المرجعية الخاصة من عمل اليوم.

أسئلة النقاش الجماعي.

1. تلتقط الفئة العالية 73% من المرض الحقيقي، لكن الفئة المنخفضة ما تزال تخفي 33%. ما العواقب السريرية والمتعلقة بالموارد لخفض عتبة الفئة العالية من 5 إلى 4؟
2. منحت اللجنة ألم الصدر ASY وزناً قدره 2+. هل ينبغي لمعيار إحالة قائم على الأعراض أن يحمل وزناً أكبر من معيار مُقاس مثل انخفاض القطعة ST؟ ومن المسؤول عن الأوزان — المحلل أم اللجنة؟
3. تعامل درجتنا المعايير على أنها مستقلة. أي أزواج في الجدول 6.3 مترابطة سريرياً، وكيف قد يضاعف ذلك احتساب الخطورة مرتين؟

6.4 الواجب المنزلي والتقييم الذاتي

تمارين الواجب المنزلي (أحضر دفترك غداً — التمرين الأول يغذي مباشرة تمهيد اليوم السابع):

1. النسبة الكلوية. على ملف `chemistry_panel.csv`، اهندس `urea_cr_ratio` = اليوريا مقسومة على الكرياتينين، واطبع ناتج `describe()` لها، وعلم المريض ذوي النسبة الأعلى من 40 (الموحية بنمط ما قبل كلوي) باستخدام `np.where`. كم مريضاً عَلم؟
2. ترميز إعادة القبول. رمز `readmitted_30d` إلى 0/1 واحسب معدل إعادة القبول لكل فئة من `age_band` (من الخطوة 6) باستخدام التجميع (`groupby`). أي فئة هي الأعلى؟
3. مؤشرات المحافظات. طبق الترميز الأحادي الفعال على `governorate` باستخدام `pd.get_dummies`. كم عموداً جديداً يظهر، وأي محافظة تسهم بأكبر عدد من المرضى؟
4. النسبة الكبدية. اهندس نسبة `AST/ALT` (نسبة `De Ritis`) وعلم النسب ≥ 2 ، وهو نمط يثير القلق في سياق تاريخ كحولي أو تشمع. كم مريضاً عَلم؟

اختبار التقييم الذاتي.

1. في تقسيم `HbA1c` إلى فئات، لماذا مررنا `right=False` إلى `pd.cut`؟
 2. رمز زميل الأقسام هكذا: `Emergency=1`، `ICU=2`، `Surgery=3`، ثم درّب نموذجاً خطياً. ما الخطأ؟
 3. ماذا تعيد `map` لقيمة غائبة عن قاموسها، ولماذا يشكل ذلك خطراً؟
 4. أي أداة تنشئ عمود 0/1 من الشرط الوحيد «الغلوكوز ≥ 126 »، وفي كم سطر؟
 5. اذكر سببين لبناء درجة مركبة موزونة يدوياً قبل تدريب أي نموذج تعلم آلي.
- الأجوبة. الواجب المنزلي: (1) متوسط النسبة 30.8، و178 مريضاً يتجاوزون 40؛ (2) المعدلات: `Child = 0.154` و `Middle-aged = 0.142` و `Elderly = 0.101` و `Young-adult = 0.101` — فئة الأطفال هي الأعلى في هذه المجموعة الاصطناعية؛ (3) ثمانية أعمدة `gov` جديدة؛ وبغداد الأكبر بواقع 336 مريضاً؛ (4) 16 مريضاً لديهم $AST/ALT \geq 2$. الاختبار: (1) لتكون الفترات مغلقة من جهة اليسار فتقع القيمة 6.5% بالضبط في (6.5, 20]، فئة السكري، بما يطابق معيار `ADA` البالغ 6.5%؛ (2) إنه يخترع ترتيباً وتباعداً زائفاً بين ردهات اسمية — فيوائم النموذج ميلاً لخرافة؛ والأداة الصحيحة هي `pd.get_dummies`؛ (3) تعيد `NaN` بصمت، فيختفي المرضى من التحليلات اللاحقة دون أي خطأ — تحقق دائماً من `np.where` مطبقة على عمود الغلوكوز عند عتبة 126 ملغ/دل — سطر واحد؛ (5) لأنها قابلة للتدقيق الكامل من الأطباء والإدارة، ولأنها الجسر المفاهيمي إلى الانحدار اللوجستي الذي وظيفته تعلم مثل هذه الأوزان من البيانات.

♣ القيمة لمديرية التشخيص — من التصديرات الخام إلى سجلات جاهزة للقرار

يحوّل مسار ميزات مبرمج تصدير المختبر كل شهر إلى السجل الجاهز للنمذجة نفسه دون أي إعادة ترميز يدوية، فتصبح مؤشرات التحري (أعداد فئات `HbA1c` وتوزيعات `TC/HDL`) قابلة لإعادة الإنتاج عبر المحافظات وأمنة عند التدقيق. ويمتص نمط الدرجة المركبة كلّ برنامج أداة فرز مؤقتة شفافة — قابلة للنشر هذا الأسبوع، ومدافعاً عنها في المراجعة، ومعاداة الحساب يدوياً — مع إنتاج الجداول الرقمية ذاتها التي تتطلبها نماذج تعلم الآلة في الأيام 7-10 بالضبط. وتتوقف المؤسسة عن الدفع مرتين لتهينة البيانات: فجدول التقرير وجدول النموذج هما المنتج نفسه.

خلاصات اليوم

- لا تتعلم النماذج إلا من الأعمدة الرقمية؛ وهندسة الميزات هي معرفة سريرية مترجمة إلى أعمدة تستطيع الخوارزمية وزنها.
- تبني العمليات الحسابية الموجهة المؤشرات المشتقة — `TC/HDL`، واليوريا إلى الكرياتينين — في سطر واحد عبر المجموعة كلها.
- تحوّل `np.where` عتبة دليل إرشادي واحدة إلى علامة 0/1؛ وتستنسج `pd.cut` مع `right=False` المعايير متعددة الفئات بدقة، بما فيها القيم الجديدة.
- `map` مع قاموس يرمز الفئات الثنائية والترتيبية؛ و `pd.get_dummies` يوسّع الفئات الاسمية — لا تفرض أبداً ترتيباً زائفاً على الردهات أو المحافظات.
- كل `map` يتبعه فحص `np.isnan()`؛ فقيم `NaN` الصامتة هي الطريقة التي يختفي بها المرضى من التحليلات — والتواريخ تحمل ميزات مجانية عبر الموصل `..dt`.

• الدرجة المركبة الموزونة أداة شفافة قابلة للتدقيق سابقة للذكاء الاصطناعي — درجتنا ضاعفت تركيز المرض القلبي الحقيقي بأكثر من مثلين من الفئة المنخفضة إلى العالية دون أي تعلم.

• الجدول المهندس heart_risk_engineered.csv هو بيانات تدريب الغد: ميزات تدخل، وتنبؤات تخرج.

غداً، في الفصل 7، نسلم هذه الميزات إلى مستهلكها الحقيقي الأول: خوارزميات تعلم الآلة — ونشاهد نموذجاً يتعلم من البيانات الأوزان التي عيّناها اليوم بقرار لجنة.

د. حسام صلاح مهدي
رئيس قسم الذكاء الاصطناعي — جامعة المصطفى
Dr. Husam Salah Mahdi — Head of AI Department, Al-Mustafa University

باب 7

اليوم 7 — الأسس الخوارزمية لتعلم الآلة

على مدى ستة أيام قمنا بتنظيف البيانات المختبرية ودمجها وتلخيصها وهندسة ميزاتها. واليوم نتجلى الغاية من كل ذلك الانضباط: نسلم البيانات إلى خوارزمية ونطلب منها أن تتعلم الأنماط التشخيصية. غير أنه قبل أن يلمس أي نموذج سجل مريض، علينا أن نتقن التصميم التجريبي الذي يفصل الذكاء الاصطناعي الطبي الجدير بالثقة عن الأمان الإحصائية. يُقِيم نموذج تعلم الآلة تماماً كما نقيّم الطبيب: ليس بمدى إجادته سرد الحالات التي درسها، بل بأدائه مع مريض لم يرههم قط. وكل خطوة في جلسة اليوم — تقسيم التدريب/الاختبار، والتقسيم الطبقي، والمعايرة القياسية (StandardScaler)، والحجر الصارم على مجموعة الاختبار — إنما وُجدت لحماية هذا المبدأ الواحد. تجاوزاً أياً منها يصبح رقم الدقة الذي تبلغه إلى المديرية محض خيال؛ فالنموذج الذي يبدو باهراً على الورق سيخطئ بصمت في تصنيف مريض حقيقيين في البصرة ونيوى. اليوم نبني البروتوكول؛ وغداً (الفصل 8) نبني المصنّفات التشخيصية نفسها.

الدقائق	الجزء	المحتوى والأنشطة
00-05	مراجعة اليوم 6	الميزات المرمّزة (Encoding) من اليوم 6 (الربط map، والرميز الأحادي one-hot، و pd.cut) هي بالضبط ما يستهلكه النموذج؛ واليوم نضيف البروتوكول.
05-10	ما هو تعلم الآلة	التعلم الخاضع للإشراف (Supervised Learning): الميزات X (اللوحة المختبرية) والهدف y (التشخيص المؤكّد)؛ التعلم بوصفه استخلاصاً للأنماط.
10-15	تشبيه امتحان البورد	مجموعة التدريب بوصفها حالات الإقامة، ومجموعة الاختبار بوصفها امتحان البورد؛ فرط التخصص (Overfitting) بوصفه حفظاً؛ ولماذا لا نقيّم أبداً على مجموعة التدريب.
15-20	المعايرة والتسرب	مقاييس الميزات: StandardScaler بوصفه الدرجة المعيارية السريرية؛ تسرب البيانات (Data Leakage) وقاعدة الملاءمة على التدريب فقط.
20-28	الخطوتان الموجّهتان 1-2	تحميل diabetes_dataset.csv؛ وتدقيق معدل الانتشار؛ وبناء X و y .
28-38	الخطوتان الموجّهتان 3-4	تقسيم طبقي بنسبة 75/25 مع random_state=42؛ والتحقق من الأبعاد ومن الحفاظ على معدل الانتشار.
38-50	الخطوتان الموجّهتان 5-6	ملاءمة StandardScaler على مجموعة التدريب فقط؛ وتحويل المجموعتين؛ وفحص مريض واحد قبل المعايرة وبعدها.
50-60	الخطوة الموجّهة 7	قواعد المقدير: fit و predict و score مع مصنّف بديل مؤقت؛ ومزالق التسرب.
60-85	التحدي السريري	تحضير heart_risk.csv من البداية إلى النهاية وصياغة فقرة المنهجية لبروتوكول دراسة وازارية.
85-90	الختام والواجب	عرض الحل، وأسئلة النقاش، واختبار التقييم الذاتي، ولمحة عن اليوم 8.

✓ الأهداف التعليمية

بنهاية هذه الجلسة ستكون قادراً على:

- تعريف تعلم الآلة الخاضع للإشراف بمصطلحات سريرية، وتحديد مصفوفة الميزات X ومتجه الهدف y في أي مجموعة بيانات تشخيصية.
- شرح فرط التخصص باستخدام تشبيه الإقامة/امتحان البورد، وبيان لماذا لا يُبلّغ عن الأداء إلا على مريض محجوزين لم يرههم

النموذج.

3. إجراء تقسيم طبقي للتدريب/الاختبار باستخدام `train_test_split(test_size=0.25, stratify=y, random_state=42)` وتبرير كل وسيطة تبريراً علمياً.
4. التحقق من الحفاظ على معدل انتشار الفئات في كلا القسمين باستخدام `value_counts(normalize=True)`.
5. تطبيق المعايير القياسية على الميزات باستخدام `StandardScaler` بملاءمته على مجموعة التدريب فقط، وتفسير القيمة المعايير بوصفها درجة معيارية سريرية.
6. التعرف على تسرب البيانات في مسار المعالجة المسبقة، وشرح لماذا يُبطل الأداء المبلّغ عنه في الدراسة.
7. تطبيق القواعد الموحدة لمكتبة `scikit-learn` — المقدر، `fit`، `transform`، `predict` — على أي نموذج مستقبلي.

→ حقيبة أدوات الذكاء الاصطناعي والاعتماديات — مكتبات اليوم

نضيف اليوم مكتبة `scikit-learn`، المكتبة القياسية لتعلّم الآلة في منظومة بايثون العلمية، إلى عدّتنا المألوفة.

```
1 import pandas as pd
2 import numpy as np
3 from sklearn.model_selection import train_test_split
4 from sklearn.preprocessing import StandardScaler
5 from sklearn.dummy import DummyClassifier
```

شيفرة 7.1: الاستيرادات المستخدمة على امتداد اليوم 7

```
pip install pandas numpy scikit-learn
```

تبقى **pandas** محرك الجداول لدينا: فهي تحمّل ملفات CSV الخاصة بالسجلات وتُبقي أسماء الأعمدة ملازمة لكل رقم، بحيث يعرف محلل المديرية دائماً أي قيمة تخص الغلوكوز وأنها تخص الكرياتينين. وتوفر **numpy** المصفوفات العددية السريعة التي تستهلكها **scikit-learn** داخلياً؛ وكل عملية معالجة موجهة (Vectorization) تدربنا عليها منذ اليوم 2 تعمل فوقها. أما **scikit-learn** فتقدم سير عمل لتعلّم الآلة بأكمله — التقسيم والمعايرة (وبدءاً من الغد) المصنّفات — خلف قواعد موحدة واحدة، ما يعني أن البروتوكول المُتحقّق منه على بيانات مستشفى واحد يمكن إعادة تشغيله دون تغيير على بيانات محافظة أخرى.

7.1 السياق السريري والإطار النظري (20 دقيقة)

قد يبدو تعلّم الآلة أمراً غريباً، لكن فكرته الجوهرية فكرة يعيشها كل طبيب سريري أصلاً: التعرف على الأنماط المتعلّم من حالات موثّقة. فالطبيب المبتدئ يصبح مشخّصاً كفواً برؤية آلاف المرضى الذين تأكدت مآلاتهم في نهاية المطاف. وتُفعل خوارزمية التعلم الخاضع للإشراف الشيء نفسه على نطاق واسع — إذ تقرأ سجلات تاريخية عُرفت فيها الصورة المخبرية والتشخيص النهائي المؤكّد معاً، وتستخلص الانتظامات الإحصائية التي تربط أحدهما بالآخر.

التعلم الخاضع للإشراف: الميزات X والهدف y

يتطلب التعلم الخاضع للإشراف كائنين متحاذيين. تحمل مصفوفة الميزات X صفّاً واحداً لكل مريض وعموداً واحداً لكل قياس — وفي حالتنا اللوحة المخبرية: الغلوكوز، ومؤشر كتلة الجسم، وضغط الدم، والعمر، وهكذا. ويحمل متجه الهدف y ، للمرضى أنفسهم وبالترتيب نفسه، المآل المؤكّد الذي نريد التنبؤ به: مصاب بالسكري أم لا، حدث قلبي أم لا. مهمة الخوارزمية أن تتعلم دالة f بحيث يكون $y \approx f(X)$ على مرضى لم نرهم. كل شيء في هذه الدورة خاضع للإشراف: فنحن نتعلم دائماً من سجلات وثّقت فيها الحقيقة المرجعية سلفاً — تشخيص مؤكّد أو مآل مفصول فيه. لذا فإن جودة y مسألة حوكمة سريرية لا مسألة برمجة: النموذج المدرب على تشخيصات متهاونة يتعلم تشخيصاً متهاوناً.

أخطر وهم في الذكاء الاصطناعي الطبي هو نموذج يبدو ممتازاً لأنه قُيّم على الحالات ذاتها التي درسها. وقد حلّ الطب هذه المشكلة للبشر منذ زمن بعيد: يتدرب الأطباء المقيمون على مجموعة من المرضى، ثم يقدّم لهم امتحان البورد مرضى مختلفين. ونحن ننسخ هذا التصميم

بهدفه.

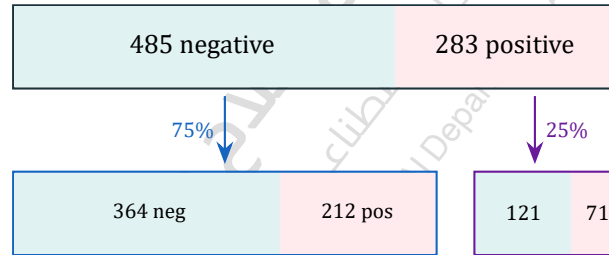
تقسيم التدريب/الاختبار: الإقامة وامتحان البورد

قبل أن يبدأ أي تعلم، نقسم السجل الموسوم إلى قسمين منفصلين لا يتقاطعان. مجموعة التدريب (هنا 75% من المرضى) هي الإقامة: يجوز للنموذج أن يدرس هذه الحالات ما شاء. ومجموعة الاختبار (25%) هي امتحان البورد: مرضى لم يصادفهم النموذج قط، لا تُفتح إلا مرة واحدة، في النهاية تماماً، لقياس المهارة التشخيصية الحقيقية. وسيطتان تجعلان التقسيم علمياً. stratify=y تُلزم كلا القسمين بالحفاظ على معدل انتشار المرض نفسه الموجود في السجل الكامل، فيكون للامتحان مزيج الحالات نفسه الذي كان للإقامة — ومن دونها يمكن للصدفة وحدها أن تمنح مجموعة الاختبار عينة غير ممثلة. أما random_state=42 فتثبت الخلط العشوائي بحيث يحصل أي زميل أو مدقق أو جهة رقابية تعيد تشغيل البروتوكول على التقسيم المطابق: قابلية إعادة الإنتاج هي المنهج العلمي، لا مجرد وسيلة راحة.

فرط التخصيص: حفظ يتكرر في هيئة مهارة

فرط التخصيص هو الطالب الذي حفظ أوراق امتحانات العام الماضي: أداء لا تشوبه شائبة على أسئلة رآها من قبل، وضياح تام حين يدخل مريض جديد. تقنياً، يكون النموذج قد لاءم الضجيج العشوائي لحالات تدريبه — المصادفات العرضية لـ 576 مريضاً بأعيانهم — بدلاً من البيولوجيا القابلة للتكرار التي تعمم على الألف مريض التالين. إنه الخطر الأكبر منفرداً في الذكاء الاصطناعي الطبي، لأن النموذج المفرط التخصيص لا يفشل بصخب؛ بل يفشل بصمت، وعلى المرضى الأهم بالضبط — مرضى المستقبل. وتقسيم التدريب/الاختبار هو أداتنا لكشفه: فالفجوة الكبيرة بين أداء التدريب وأداء الاختبار هي بصمة الحفظ (الشكل 7.2).

Full registry: 768 patients (36.8% positive)



Training set: 576 patients (36.8% positive) Test set: 192 (37.0%)

the residency — the model studies these cases the board exam — sealed until the end

شكل 7.1: التقسيم الطبقي بنسبة 75/25 لسجل السكري. لأننا نمرر stratify=y، يُحفظ مزيج الحالات (السموي = سلبى، الأحمر = إيجابى) بدقة شبه تامة في كلا القسمين — فيقدم الامتحان معدل انتشار المرض نفسه الذي قدمته الإقامة.

يتعلق الركن الثاني من بروتوكول اليوم بالمقاييس التي تعيش عليها قياساتنا. يعرض الجدول 7.1 قيماً تلخيصية حقيقية من لوحة الكيمياء السريرية الخاصة بنا (chemistry_panel.csv، n = 1200): يُقاس الكوليسترول بالملئات، والصوديوم حول 140، والبوتاسيوم حول 4، والكرياتينين دون 2. كثير من الخوارزميات تحسب مسافات أو مجاميع موزونة عبر الميزات؛ وعلى المقاييس الخام، فإن تذبذباً تافهاً سريرياً مقداره 10 ملغ/دل في الكوليسترول يطغى حسابياً على انزياح مهذّب للحياة مقداره 1 ملي مول/لتر في البوتاسيوم. الأرقام الكبيرة تُسكت الأرقام الصغيرة.

جدول 7.1: تفاوت المقاييس عبر لوحة كيمياء روتينية (قيم حقيقية من chemistry_panel.csv، 1200 مريض). من دون المعايرة القياسية، تهيمن الميزات في الصفوف العليا عددياً على الميزات في الصفوف السفلى.

المادة المُحلّلة	الحد الأدنى	الحد الأقصى	المتوسط
الكوليسترول الكلي (ملغ/دل)	90.0	309.0	184.95
الصوديوم (ملي مول/لتر)	130.0	152.0	139.45
الغلوكوز (ملغ/دل)	55.0	186.0	97.22
البوتاسيوم (ملي مول/لتر)	2.9	5.6	4.20
الكرياتينين (ملغ/دل)	0.4	1.61	0.89

المعايرة القياسية: الدرجة المعيارية التي تستخدمها أصلاً

يستبدل StandardScaler بكل قيمة خام القيمة

$$z = \frac{x - \bar{x}_{\text{train}}}{s_{\text{train}}},$$

أي عدد الانحرافات المعيارية التي يبعدها المريض عن متوسط مجموعة التدريب. بعد المعايرة، يصبح لكل ميزة متوسط $0 \approx$ وانحراف معياري $1 \approx$ ، فيتكلم البوتاسيوم والكوليسترول بصوت إحصائي متساوٍ. وقد استخدم الأطباء السريريون هذا التحويل بعينه لعقود: فدرجة معيارية للوزن مقابل الطول وفق منظمة الصحة العالمية قدرها 3- تُعرّف سوء التغذية الحاد الوخيم، ودرجة T لكثافة العظم بفحص DXA بقيمة $-2.5 \leq$ تُعرّف هشاشة العظام. لم يخترع تعلم الآلة الدرجة المعيارية؛ بل استعارها من مخطط النمو.

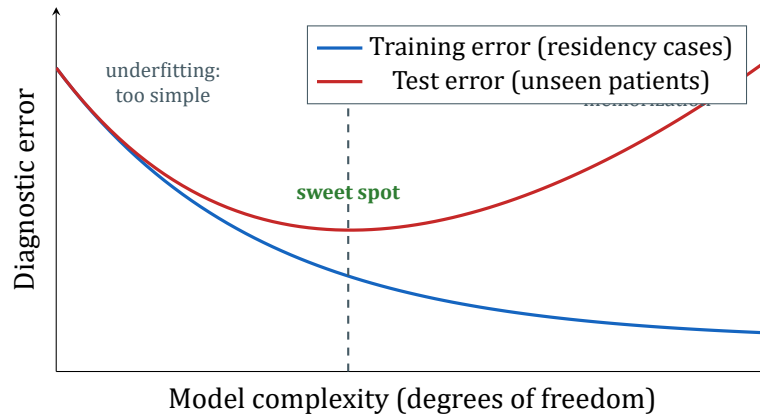
تسرب البيانات: حين يتسرب الامتحان إلى الإقامة

إن المعايير نفسه يتعلم رقمين لكل ميزة — المتوسط والانحراف المعياري. فإذا لاءمناه على السجل الكامل قبل التقسيم، احتوى هذان الرقمان أصلاً معلومات من مرضى الاختبار المستقبليين؛ لقد تسرب الامتحان إلى الإقامة، ولم تعد مجموعة الاختبار غير مرئية حقاً. القاعدة مطلقة: لائم المعايير على مجموعة التدريب فقط، ثم حوّل المجموعتين كليهما بمعاملات مجموعة التدريب. وتنطبق القاعدة نفسها على كل خطوة معالجة مسبقة متعلّمة سنستخدمها مستقبلاً — قيم التعويض، ومخططات الترميز، واختيارات الميزات. لا يُسقط التسرب البرنامج؛ بل يضخم بصمت الأداء المبلغ عنه، وهذا بالضبط ما يجعل الجهات الرقابية ومراجعي المجالات العلمية يبحثون عنه أولاً.

★ لؤلؤة سريرية — العتبات والدرجات المعيارية في الممارسة التشخيصية اليومية

يستند هدف اليوم y في نهاية المطاف إلى معايير تشخيصية قياسية، وفكرة الدرجة المعيارية تحكم عدداً منها أصلاً. معايير الجمعية الأمريكية للسكري (ADA): غلوكوز البلازما الصيامي $126 \geq$ ملغ/دل في مناسبتين، أو سكر تراكمي $\text{HbA1c} \geq 6.5\%$ ، يؤكد داء السكري؛ وغلوكوز صيامي 100-125 ملغ/دل أو $\text{HbA1c} 5.7-6.4\%$ يحدد نطاق ما قبل السكري الذي تستهدفه برامج التحري. وعلى جانب الدرجة المعيارية: نسم معايير النمو لمنظمة الصحة العالمية وزناً مقابل الطول دون 2- انحراف معياري بسوء تغذية حاد متوسط ودون 3- بسوء تغذية حاد وخيم، وتُعرّف منظمة الصحة العالمية هشاشة العظام بدرجة T بفحص DXA بقيمة $-2.5 \leq$. حين يُبلغ StandardScaler عن غلوكوز مريض بأنه $+1.08$ ، أقرأه تماماً كما تقرأ تلك المخططات: انحراف معياري واحد فوق المجموعة المرجعية — وهي هنا مجموعة التدريب.

أخيراً، السبب الذي يجعل قواعد اليوم جديرة بالحفظ: تتيح scikit-learn كل أداة — المعايير، والانحدار اللوجستي (Logistic Regression)، والغابة العشوائية (Random Forest)، والشبكة العصبية — عبر الطقس الثلاثي نفسه. أنشئ المقدّر، واستدع fit على بيانات التدريب، ثم استدع transform (للمعالجات المسبقة) أو predict (للتماذج) على بيانات جديدة. تعلّم هذا الطقس مرة واحدة اليوم مع معايير ومصنّف بديل مؤقت، ولن تحمل نماذج الغد التشخيصية الحقيقية أي مفاجآت نحوية.



شكل 7.2: بصمة فرط التخصص. مع ازدياد التعقيد، ينخفض خطأ التدريب باستمرار، لكن الخطأ على المرضى غير المرئيين ينعطف عائداً: بعد النقطة المثلى يكون النموذج يحفظ حالات الإقامة بدلاً من أن يتعلم الطب. المنحنى الأحمر وحده — المقيس على مجموعة الاختبار المحجوزة — يقول الحقيقة.

7.2 التطبيق الموجّه (40 دقيقة)

■ عرض تطبيقي مباشر — تحضير سجل السكري لتعلم الآلة

نأخذ `diabetes_dataset.csv` — 768 مريض تحرّ مع ثمانية قياسات روتينية ومآل سكري مؤكّد — عبر بروتوكول ما قبل النمذجة الكامل: بناء X و y ، وإجراء تقسيم طبقي قابل لإعادة الإنتاج، والتحقق من معدل الانتشار، والمعايرة القياسية بمعايير ملائم على التدريب، وفحص مريض واحد قبل التحويل وبعده، والتمرّن على قواعد `fit/predict` مع مصيّف بديل مؤقت. كل ضغطة مفتاح اليوم ستظهر من جديد دون تغيير أمام نماذج الغد الحقيقية.

الخطوة 1: تحميل السجل وتدقيق المآل

نبدأ كعادتنا دائماً: حمّل، وانظر، وعدّد. إن الرقم الأهم منفرداً في أي مجموعة بيانات تشخيصية هو معدل انتشار (Prevalence) الحالة المستهدفة — فعليه يرسو كل فحص سلامة لاحق.

```
1 import pandas as pd
2 pd.set_option("display.width", 100)
3
4 df = pd.read_csv("diabetes_dataset.csv")
5 print(df.shape)
6 print(df[["glucose", "bmi", "age", "outcome"]].head(3))
7 print(df["outcome"].value_counts(normalize=True).round(3))
```

شيفرة 7.2: تحميل العينة وقياس معدل الانتشار

```
(768, 9)
  glucose  bmi  age  outcome
0    134  41.3  31        0
1    149  34.7  37        0
2    106  31.5  46        1
outcome
0    0.632
1    0.368
Name: proportion, dtype: float64
```

يضم السجل 768 مريضاً و9 أعمدة؛ ويحمل 36.8% منهم تشخيص سكري مؤكّداً. احفظ رقم 36.8% هذا عن ظهر قلب — فستطالب باسترداده من نصفي التقسيم كليهما.

الخطوة 2: فصل الميزات X عن الهدف y

يجب ألا يرى النموذج ورقة الإجابة داخل ورقة أسئلته أبداً. لذلك نُسقط عمود outcome لتكوين X ونحتفظ به وحده بوصفه y . والمحاذاة بينهما موضعية: الصف i من X والعنصر i من y يصفان المريض نفسه.

```
1 X = df.drop(columns="outcome")
2 y = df["outcome"]
3 print("X:", X.shape, " y:", y.shape)
4 print(X.columns[:5].tolist())
5 print(X.columns[5:].tolist())
```

شيفرة 7.3: مصفوفة الميزات ومتجه الهدف

```
X: (768, 8) y: (768,)
['pregnancies', 'glucose', 'blood_pressure', 'skin_thickness', 'insulin']
['bmi', 'diabetes_pedigree', 'age']
```

ثمانية ميزات لكل مريض، ووسم مؤكد واحد لكل منهم. لاحظ أن `df.drop(columns="outcome")` تعيد إطار بيانات (DataFrame) جديداً — فالسجل الأصلي لم يُمس، وهي عادة صديقة للتدقيق اكتسبناها منذ اليوم 3.

الخطوة 3: التقسيم الطبقي القابل لإعادة الإنتاج

استدعاء دالة واحد ينفذ التصميم التجريبي كله. `test_size=0.25` يختم على رُبُع السجل في قاعة الامتحان؛ و `stratify=y` يحفظ مزيج الحالات؛ و `random_state=42` يجعل الخلط قابلاً لإعادة الإنتاج لأي مدقق.

```
1 from sklearn.model_selection import train_test_split
2
3 X_train, X_test, y_train, y_test = train_test_split(
4     X, y, test_size=0.25, stratify=y, random_state=42)
5
6 print("X_train:", X_train.shape, " y_train:", y_train.shape)
7 print("X_test :", X_test.shape, " y_test :", y_test.shape)
```

شيفرة 7.4: التقسيم الطبقي بنسبة 75/25

```
X_train: (576, 8) y_train: (576,)
X_test : (192, 8) y_test : (192,)
```

تعيد الدالة أربع قطع متحاذاة بترتيب ثابت: الميزات ثم الهدف، والتدريب ثم الاختبار. من هذا السطر فصاعداً، يوضع `X_test` و `y_test` في الحجر — فلن نحسب منهما إحصاء واحدة حتى وقت التقييم.

الخطوة 4: التحقق من الحفاظ على معدل الانتشار

ثق، ثم تحقق. إذا نجح التقسيم الطبقي، وجب أن يردد كلا القسمين معدل انتشار السجل البالغ 36.8%.

```
1 print(y_train.value_counts(normalize=True).round(3))
2 print(y_test.value_counts(normalize=True).round(3))
```

شيفرة 7.5: فحص معدل الانتشار في كلا القسمين

```
outcome
0    0.632
1    0.368
Name: proportion, dtype: float64
outcome
```

```
0    0.63
1    0.37
Name: proportion, dtype: float64
```

36.8% مقابل 37.0% — تقدّم الإقامة وامتحان البورد عبء المرض نفسه، تماماً كما وعد الشكل 7.1. يستغرق هذا الفحص سطرين اثنين، وينبغي أن يكون في كل بروتوكول نمذجة وزاري.

✖ خطأ شائع — إغفال stratify=y

أغفل stratify=y فيصبح مزيج حالات مجموعة الاختبار يانصيباً. على هذا السجل بعينه، أنتج تقسيمان غير طبقين ببذرتين أخريين معدلي انتشار في الاختبار قدرهما 41.1% و32.8% حول القيمة الحقيقية 36.8% — تأرجح قدره ثماني نقاط مئوية. الدقة المقيسة مقابل معدل انتشار منحرف ليست قابلة للمقارنة عبر الدراسات، ولا عبر المستشفيات، ولا حتى عبر إعادة تشغيل دفترك أنت. طبق التقسيم الطبقي على كل تقسيم تشخيصي، دون استثناء.

الخطوة 5: المعايرة القياسية — الملاءمة على مجموعة التدريب فقط

الآن يصبح انضباط منع التسرب ملموساً. يتعلم المعايير متوسط كل ميزة وانحرافها المعياري من X_train فقط؛ ثم يُحوّل كلا القسمين بمعاملات مجموعة التدريب نفسها. ونلف النتائج مرة أخرى في أطر بيانات DataFrame حتى تبقى أسماء الأعمدة حية.

```
1 from sklearn.preprocessing import StandardScaler
2
3 scaler = StandardScaler()
4 scaler.fit(X_train) # Learns mean and SD from train ONLY
5
6 X_train_s = pd.DataFrame(scaler.transform(X_train),
7                           columns=X.columns, index=X_train.index)
8 X_test_s = pd.DataFrame(scaler.transform(X_test),
9                          columns=X.columns, index=X_test.index)
10
11 cols = ["glucose", "blood_pressure", "bmi"]
12 print("training set after scaling:")
13 print(X_train_s[cols].agg(["mean", "std"]).round(3))
14 print("test set after the SAME (train-fitted) scaler:")
15 print(X_test_s[cols].agg(["mean", "std"]).round(3))
```

شيفرة 7.6: معايرة قياسية ملاءمة على التدريب لكلا القسمين

```
training set after scaling:
   glucose  blood_pressure  bmi
mean    0.000         0.000 -0.000
std     1.001         1.001  1.001
test set after the SAME (train-fitted) scaler:
   glucose  blood_pressure  bmi
mean   -0.072         0.041 -0.088
std     1.030         0.992  1.040
```

اقرأ الكتلتين بعناية، فهما تختزان العقيدة كلها. تحطّ مجموعة التدريب عند متوسط 0 وانحراف معياري 1 بحكم البناء. أما مجموعة الاختبار فتحطّ قرب الصفر والواحد لا عليهما تماماً — لأنها حُوّلت بمعاملات مجموعة التدريب لا بمعاملاتها هي. ذلك النقص الصغير هو البرهان على أن أي معلومة من مجموعة الاختبار لم تدخل المعايير. إذا أظهرت الكتلة الثانية يوماً أصفاً تاماً، فقد سرّبت. يعرض الجدول 7.2 ما فعلته المعايرة القياسية بالمدايات الخام.

جدول 7.2: مدايات ميزات مجموعة التدريب قبل المعايرة القياسية وبعدها (قيم حقيقية، $n = 576$). بعد المعايرة، تعيش كل ميزة على محور درجات معيارية قابل للمقارنة؛ ويبقى الذيل البيولوجي الطويل للإنسولين حداً أقصى قدره 8.34.

الميزة	قبل المعايرة			بعد المعايرة	
	الحد الأدنى	الحد الأقصى	المتوسط	الانحراف	الحد الأدنى
glucose	44.00	211.0	116.78	29.90	-2.44
blood_pressure	40.00	105.0	69.35	12.03	-2.44
insulin	15.00	850.0	114.57	88.26	-1.13
bmi	15.00	52.5	31.88	6.59	-2.56
diabetes_pedigree	0.08	2.5	0.48	0.35	-1.16
age	21.00	79.0	38.18	10.92	-1.57
					3.15

✖ خطأ شائع — ملاءمة المُعايير قبل التقسيم

علة التسرب الكلاسيكية سطر واحد مغري: `scaler.fit(X)` — أو اختصاره `scaler.fit_transform(X)` — منقذاً على السجل الكامل قبل التقسيم. على هذه العينة يبدو التلوث بريئاً (متوسط الغلوكوز على البيانات الكاملة 116.25 مقابل 116.78 على التدريب وحده)، لكن الجهات الرقابية تحكم على المبدأ لا على المقدار: قيم مرضى الاختبار قد شكّلت التحويل المطبق على مرضى الاختبار أنفسهم، فلم يعد الامتحان أعمى. الخطأ نفسه مختبئاً داخل التعويض عن القيم المفقودة أو اختيار الميزات يمكن أن يزيح الدقة المبلّغ عنها بنقاط مئوية كاملة. والصيغة الشائعة بالقدر نفسه — استدعاء `fit_transform` مرة ثانية على `X_test` — تعيد ملاءمة المُعايير على الامتحان ذاته وهي باطلة بالقدر نفسه. ملاءمة `fit` واحدة لكل دراسة، على بيانات التدريب فقط؛ و `transform` في كل مكان آخر.

الخطوة 6: مريض واحد، منظران

تتحول الأرقام إلى حدى حين نتتبع مريضاً واحداً عبر التحويل. نأخذ أول مريض في مجموعة التدريب ونطبع الصف نفسه على المقياسين.

```
1 pid = X_train.index[0]
2 print("first training index:", pid)
3 cols6 = ["glucose", "blood_pressure", "insulin", "bmi", "age"]
4 print("before scaling:")
5 print(X_train.loc[[pid], cols6])
6 print("after scaling:")
7 print(X_train_s.loc[[pid], cols6].round(2))
```

شيفرة 7.7: المريض نفسه قبل المعايرة وبعدها

```
first training index: 711
before scaling:
   glucose  blood_pressure  insulin  bmi  age
711    149             64      146  43.2  35
after scaling:
   glucose  blood_pressure  insulin  bmi  age
711    1.08          -0.45     0.36  1.72 -0.29
```

يُقرأ المريض 711 كرسالة إحالة مكتوبة بالدرجات المعيارية: غلوكوز أعلى من متوسط مجموعة التدريب بنحو انحراف معياري واحد ($+1.08$)؛ والقيمة الخام 149 ملغ/دل كانت ستستدعي فحص غلوكوز صيامي تأكيدى وفق معايير ADA)، وضغط دم أدنى قليلاً من المتوسط (-0.45)، ومؤشر كتلة جسم مرتفع بجلاء ($+1.72$)، والقيمة الخام 43.2 — سمّنة من الدرجة الثالثة، وعمر أدنى بقليل من متوسط المجموعة. لم يُفقد أي شيء سريري؛ إنما استُبدلت بالوحدات قابلية المقارنة، تماماً كما يستبدل مخطط النمو بالكيلوغرامات درجات معيارية.

الخطوة 7: القواعد التي لا تتغير أبداً

في الختام، نتمرّن على طقس المقدّر بمصنّف بديل ساذج عمداً: `DummyClassifier`، الذي يتجاهل الميزات ويتنبأ دائماً بالفئة الأغلبية. إنه عديم الفائدة سرّياً بحكم التصميم — ولهذا فهو خط الأساس المثالي: أي نموذج حقيقي نبنيه غداً يجب أن يتفوق على درجته ليبرر وجوده.

```
1 from sklearn.dummy import DummyClassifier
2
3 model = DummyClassifier(strategy="most_frequent")
4 model.fit(X_train_s, y_train) # 1. Learn from the residency
5 print(model.predict(X_test_s.head(5))) # 2. predict unseen patients
6 print("baseline accuracy:",
7       round(model.score(X_test_s, y_test), 3)) # 3. grade on the exam
```

شيفرة 7.8: `fit / predict / score` مع مقدّر بديل مؤقت

```
[0 0 0 0 0]
baseline accuracy: 0.63
```

يعلن المصنّف الوهمي أن كل مريض غير مصاب بالسكري ومع ذلك يسجل 63% — لأن 63% من مجموعة الاختبار غير مصابين بالسكري فعلاً. هذه هي أرضية الدقة التي يرسمها معدل الانتشار وحده، وهي أول رقم تستشهد به كلما عرض عليك أحدهم دقة تبدو مبهمة. غداً، حين يحل الانحدار اللوجستي محل `DummyClassifier` على هذا السطر بعينه، لاحظ ما الذي يتغير في الشيفرة: اسم المقدّر فقط. `fit` و `predict` و `score` — القواعد دائمة لا تتبدل.

7.3 التحدي السريري (30 دقيقة)

التحدي السريري — كتابة قسم المنهجية لدراسة الخطورة القلبية

تُعَدّ مديرية التشخيص بروتوكول دراسة بعنوان: التنبؤ بالخطورة القلبية المهمة من قياسات العيادات الخارجية الروتينية، باستخدام سجل الـ 900 مريض `heart_risk.csv`. أنت منهجّي البيانات في الدراسة. قبل تدريب أي نموذج، يجب أن يحدد قسم المنهجية في البروتوكول — وأن تنفذ شيفرتك — مسار تحضير محكماً ضد التسرب. باستخدام ترميزات اليوم 6 و بروتوكول اليوم:

1. حمّل `heart_risk.csv`، وأبلغ عن أبعاده، وقس معدل انتشار `target` باستخدام `.value_counts(normalize=True)`.
2. رمّز المتغيرات الفئوية بأدوات اليوم 6 ومن دون حلقات تكرارية أو جمل شرطية: اربط بـ `map` عمود `sex` إلى عمود ثنائي `sex_male`، واربط بـ `map` عمود `exercise_angina` إلى `angina_yes`، وطبّق الترميز الأحادي على `chest_pain_type` باستخدام `pd.get_dummies` (بأداة `cp`، ونوع بيانات صحيح).
3. ابن `y` من `target` و `X` من كل الميزات المرقّزة، مع إسقاط `target` والعمودين النصيين الأصليين. أبلغ عن البعدين واذكر أسماء الميزات النهائية.
4. قسّم 75/25 مع `stratify=y` و `random_state=42`؛ وأبلغ عن أبعاد القطع الأربع.
5. تحقّق من الحفاظ على توازن الفئات في `y_train` و `y_test`.
6. طبّق المعايرة القياسية بمُعَاير `StandardScaler` ملائم على مجموعة التدريب فقط؛ وأثبت نجاحها بطباعة المتوسط والانحراف المعياري لأربعة أعمدة معايرة في كل قسم.
7. اكتب فقرة المنهجية المكوّنة من ست جمل تماماً كما ستظهر في البروتوكول الوزاري، ثم أجب كتابةً: لماذا كانت ملائمة المُعايير قبل التقسيم سُبُطِل الدراسة؟

✓ حل المدرب — تحضير دراسة الخطورة القلبية

المسار الكامل، للمهام 6-1. لاحظ أننا نعيد اشتقاق كل ميزة من الملف الخام `heart_risk.csv` حتى تكون الجلسة مكتفية بذاتها — وقد تختلف أسماء الأعمدة قليلاً عن ملف `heart_risk_engineered.csv` الذي حفظته في اليوم 6:

```

1 import pandas as pd
2 from sklearn.model_selection import train_test_split
3 from sklearn.preprocessing import StandardScaler
4 pd.set_option("display.width", 100)
5
6 # Task 1 --- Load and audit
7 heart = pd.read_csv("heart_risk.csv")
8 print("raw:", heart.shape)
9 print(heart["target"].value_counts(normalize=True).round(3))
10
11 # Task 2 --- Day 6 encodings: map + one-hot, fully vectorized
12 heart["sex_male"] = heart["sex"].map({"M": 1, "F": 0})
13 heart["angina_yes"] = heart["exercise_angina"].map({"Y": 1, "N": 0})
14 heart = pd.get_dummies(heart, columns=["chest_pain_type"],
15                          prefix="cp", dtype=int)
16
17 # Task 3 --- features and target
18 y = heart["target"]
19 X = heart.drop(columns=["target", "sex", "exercise_angina"])
20 print("X:", X.shape, " y:", y.shape)
21 print(X.columns[:6].tolist())
22 print(X.columns[6:].tolist())
23
24 # Task 4 --- stratified, reproducible split
25 X_train, X_test, y_train, y_test = train_test_split(
26     X, y, test_size=0.25, stratify=y, random_state=42)
27 print("train:", X_train.shape, " test:", X_test.shape)
28
29 # Task 5 --- class balance in both partitions
30 print(y_train.value_counts(normalize=True).round(3))
31 print(y_test.value_counts(normalize=True).round(3))
32
33 # Task 6 --- scaler fitted on train only, applied to both
34 scaler = StandardScaler()
35 X_train_s = pd.DataFrame(scaler.fit_transform(X_train),
36                          columns=X.columns, index=X_train.index)
37 X_test_s = pd.DataFrame(scaler.transform(X_test),
38                        columns=X.columns, index=X_test.index)
39 chk = ["age", "cholesterol", "max_hr", "st_depression"]
40 print(X_train_s[chk].agg(["mean", "std"]).round(3))
41 print(X_test_s[chk].agg(["mean", "std"]).round(3))

```

شيفرة 7.9: تحضير محكم ضد التسرب لسجل القلب

```

raw: (900, 10)
target
0    0.572
1    0.428
Name: proportion, dtype: float64
X: (900, 12)  y: (900,)
['age', 'resting_bp', 'cholesterol', 'fasting_bs', 'max_hr', 'st_depression']
['sex_male', 'angina_yes', 'cp_ASY', 'cp_ATA', 'cp_NAP', 'cp_TA']
train: (675, 12)  test: (225, 12)
target
0    0.572

```

```

1    0.428
Name: proportion, dtype: float64
target
0    0.573
1    0.427
Name: proportion, dtype: float64
   age  cholesterol  max_hr  st_depression
mean -0.000         0.000    0.000         0.000
std   1.001         1.001    1.001         1.001
   age  cholesterol  max_hr  st_depression
mean -0.063        -0.133    0.193         0.166
std   0.981         0.950    1.001         1.198

```

التفسير السريري. يحمل السجل معدل انتشار قدره 42.8% للخطورة القلبية المهمة؛ ويعيد التقسيم الطبقي إنتاجه في كلا القسمين (42.8% تدريباً، 42.7% اختباراً)، وبذلك سيُمتحن نموذج الغد على مزيج الحالات نفسه الذي سيواجهه في العيادة. استدعاء `map` واستدعاء `get_dummies` الواحد تحوّل كل واصف سريري — الجنس، والذبحة الجهدية، وطابع ألم الصدر — إلى أعمدة عددية من دون حلقة واحدة. وتجلس كتلة التدريب المعايرة عند متوسط 0 وانحراف معياري 1 تماماً؛ بينما تجلس كتلة الاختبار قرب تلك القيم لا عندها (مثلاً متوسط الكوليسترول -0.133)، وهذه هي البصمة المرئية لمُعَايير ملاءم على التدريب على نحو صحيح. لاحظ فرق الاصطلاح الوحيد عن العرض التطبيقي: `fit_transform` على `X_train` يدمج الخطوتين الأولى والثانية — وهو مشروع هناك وهناك فقط، لأنه يلائم على بيانات التدريب وحدها.

المهمة 7 — فقرة المنهجية. حللنا 900 سجل عيادات خارجية من سجل القلب لدى المديرية، حمل 42.8% منها المآل المفصول فيه للخطورة القلبية المهمة. رُمز الجنس والذبحة المستحثة بالجهد مؤشرين ثنائيين، ورُمز نوع ألم الصدر ترميزاً أحاديّاً في أربعة أعمدة مؤشرات، بما أنتج 12 ميزة عددية لكل مريض. قُسمت العينة إلى مجموعة تدريب (675 مريضاً، 75%) ومجموعة اختبار محجوزة (225 مريضاً، 25%) باستخدام معاينة عشوائية طبقية بذرة عشوائية مثبتة (42) للحفاظ على معدل انتشار المآل في كلا القسمين ولضمان إعادة إنتاج التقسيم بدقة تامة. وخضعت جميع الميزات للمعايرة القياسية إلى متوسط صفري وتباين واحد باستخدام معاملات قُدِّرَت حصرياً على مجموعة التدريب ثم طُبِّقَت دون تغيير على مجموعة الاختبار. وأُبقيت مجموعة الاختبار في الحجر طوال جميع مراحل تطوير النموذج ولم يُطلَع عليها إلا مرة واحدة، لتقدير الأداء النهائي. وأكد التحقق معدلي انتشار للمآل قدرهما 42.8% و42.7% في قسمي التدريب والاختبار على التوالي.

المهمة 7 — لماذا تُبطل المعايرة قبل التقسيم الدراسة. يقدر المُعَايير متوسطاً وانحرافاً معيارياً لكل ميزة. فإذا لَوْنَم قبل التقسيم، امتصت تلك التقديرات قيم مرضى الاختبار، فيكون التحويل المطبّق لاحقاً على مجموعة الاختبار مشتقاً جزئياً من مجموعة الاختبار نفسها — أي أن تعمية الدراسة مكسورة بحكم البناء. عندئذٍ يبالغ الأداء الاختباري المبلغ عنه في تقدير ما سيحققه النموذج على مرضى جدد حقاً في البصرة أو نينوى، وتغدو الدراسة غير قابلة لإعادة الإنتاج كما وُصفت (إذ يدعي قسم منهجيتها تقييماً على بيانات غير مرئية لم يحدث قط)، وأي جهة رقابية تدقق المسار ستصنّفه تسرب بيانات وترفض التحقق. وتجنّب هذا الخطأ لا يكلف شيئاً: لانم بعد التقسيم، وعلى قسم التدريب فقط.

نقاش جماعي.

1. تتلقى المديرية نموذجاً من مورّد يدّعي دقة 99% للخطورة القلبية. ما الأسئلة الثلاثة من جلسة اليوم التي تطرحها قبل تصديق الرقم، وما الذي سيكشفه جواب كل سؤال؟
2. سجلاتنا لقطات زمنية. إذا دُرِّب النموذج على مرضى عام 2025 ونُشر في عام 2027، فما المكافئ السريري لعدم التطابق بين التدريب والاختبار، وكيف يمكن للمديرية أن تراقبه؟
3. سجل خط الأساس الوهمي 63% على السكري بالتنبؤ بأن لا أحد مريض. في أي برامج تحرّج لدى المديرية يكون مثل هذا الخط الأساس عالياً عددياً لكنه كارثي تشغيلاً، وما الذي يعنيه ذلك بشأن الإبلاغ عن الدقة وحدها؟

7.4 الواجب المنزلي والتقييم الذاتي

تمارين الواجب المنزلي.

1. أعد تحضير بيانات السكري مع `test_size=0.2` و `random_state=42`. أبلغ عن الأبعاد الأربعة ومعدلي الانتشار. هل يظل التقسيم الطبقي صامداً عند نسبة 80/20؟

2. شغل تقسيم السكري مرتين مع `random_state=42` ومرة مع `random_state=7`، طابعاً `X_test.index[:5]` في كل مرة. اكتب جملتين عما يبرهنه ذلك بشأن قابلية إعادة الإنتاج، ولماذا يجب على البروتوكول الوزاري تسجيل البذرة.
3. باستخدام مجموعة التدريب المعايّرة، عدّ مرضى التدريب الذين تتجاوز الدرجة المعيارية للغلوكوز لديهم $+2$ باستخدام قناع منطقي (Boolean Mask) و `sum()` (من دون حلقات). حول العتبة رجوعاً إلى ملغ/دل بالاستعانة بالجدول 7.2 وعلّق على معناها السريري.
4. حضّر `cbc_results.csv` لنموذج فقر دم مستقبلي: ابن `X` من `wbc` و `hgb` و `plt`، وهدفاً `y` بقاعدة معالجة موجهة — `(cbc["hgb"] < 12).astype(int)` — ثم قسّم مع التقسيم الطبقي وعابر على نحو صحيح. أبلغ عن معدل الانتشار في كلا القسمين.
5. (إثرائي) اقرأ توثيق `DummyClassifier` وأعد تشغيل الخطوة 7 مع ضبط الاستراتيجية على "stratified". اشرح لماذا تأتي دقتها أدنى من `most_frequent` هنا، ولماذا لا يُعد أي منهما نموذجاً تشخيصياً حقيقياً.

اختبار التقييم الذاتي.

1. في التعلم الخاضع للإشراف على `diabetes_dataset.csv`، ما هما `X` و `y` بالضبط، وما الخاصية التي يجب أن تربط صفوفهما؟
2. يبلغ زميل عن دقة 98% مقيسة على مجموعة التدريب. ما اسم الخطر الذي تشك فيه، وما التغيير الوحيد في التقييم الذي يكشفه؟
3. ما الذي كان سيغيّره `stratify=y` لو أن معدل انتشار مجموعة الاختبار من دونه جاء 32.8% بدلاً من 36.8%؟
4. لماذا تُظهر مجموعة الاختبار المعايّرة متوسطاً قدره -0.072 لا 0.000 بالضبط، ولماذا يُعد ذلك علامة جيدة؟
5. أي من هذه الأسطر يسرّب: (أ) `scaler.fit(X_train)`؛ (ب) `scaler.fit_transform(X)` قبل التقسيم؛ (ج) `scaler.transform(X_test)`؟

إجابات الاختبار. (1) `X` هي مصفوفة الميزات المختبرية والديموغرافية بأبعاد 8×768 ؛ و `y` هو متجه المآلات المؤكّدة المكوّن من 768 عنصراً؛ ويجب أن يصف الصف `i` من `X` والعنصر `i` من `y` المريض نفسه. (2) فرط التخصص — حفظ حالات التدريب؛ والحل أن يُقَيّم النموذج على مجموعة اختبار محجوزة لم يرها قط أثناء الملاءمة. (3) يلزم كلا القسمين بإعادة إنتاج معدل انتشار السجل البالغ 36.8%، فيصبح مزيج حالات الامتحان ممثلاً والنتائج قابلة للمقارنة عبر التشغيلات. (4) لأن مجموعة الاختبار حُوّلت بمتوسط مجموعة التدريب وانحرفها المعياري لا بمتوسطها هي؛ والإزاحة الصغيرة تثبت أن أي إحصاءات من مجموعة الاختبار لم تدخل المعايير — فالأصفار التامة كانت ستشير إلى تسرب. (5) (ب) وحده يسرّب: فهو يدع مرضى الاختبار يشكّلون معاملات المعايّرة قبل ختم الامتحان. (أ) هي الملاءمة الصحيحة؛ و(ج) هو التحويل الصحيح والإلزامي لمجموعة الاختبار.

♣ القيمة لمديرية التشخيص — بروتوكول تحقق قابل لإعادة الإنتاج لكل نموذج

اليوم تكسب المديرية دستوراً المنهجي للذكاء الاصطناعي. فكل مقترح نموذج — داخلياً كان أم من مورّد — يمكن الآن استجوابه بثلاثة أسئلة قابلة للتدقيق: هل قيس الأداء على مرضى محجوزين، وهل كان التقسيم طبقياً ومثبّت البذرة، وهل لوثمت كل خطوة معالجة مسبقة متعلّمة على بيانات التدريب فقط؟ الموظفون القادرون على تشغيل هذا البروتوكول يستطيعون إعادة إنتاج تحضير بيانات أي دراسة سطرّاً بسطر انطلاقاً من فقرة منهجيتها، ومقارنة النماذج المرشحة بخطط أساس انتشار نزيه بدلاً من ادعاءات الدقة الخام، ورفض عمليات التحقق المسرّبة قبل وصولها إلى مرحلة التعاقد. هذا هو الفرق بين شراء ذكاء اصطناعي تشخيصي بأرقام تسويقية واعتماده بالدليل.

خلاصات اليوم

- التعلم الخاضع للإشراف يستخلص الأنماط التشخيصية من السجلات التاريخية: ميزات `X` (اللوحة المختبرية) مقترنة صفّاً بصف بهدف مؤكّد `y`.
- مجموعة التدريب = الإقامة؛ مجموعة الاختبار = امتحان البورد. لا يُبلّغ عن الأداء أبداً إلا على مرضى لم يرههم النموذج قط.
- فرط التخصص حفظٌ يتنكر في هيئة مهارة — الخطر الأكبر منفرداً في الذكاء الاصطناعي الطبي — ومجموعة الاختبار المحجوزة هي الأداة التي تكشفه.
- `stratify=y` يحفظ معدل انتشار المرض في كلا القسمين؛ و `random_state` يجعل التقسيم قابلاً لإعادة الإنتاج. وكلاهما ينتهي إلى كل بروتوكول وزاري.
- `StandardScaler` هو الدرجة المعيارية السريرية: يضع البوتاسيوم والكوليسترول على المحور نفسه فلا تُسكّت ميزةٌ ميزةً أخرى.

- لائم المعايير على مجموعة التدريب فقط وحول المجموعتين؛ فالملاءمة قبل التقسيم تسرب بيانات يُبطل الدراسة.
 - قواعد scikit-learn لا تتغير أبداً: أنشئ المقدر، و `fit` على بيانات التدريب، و `transform` أو `predict` على البيانات الجديدة.
 - خط أساس الانتشار (63% هنا، بالتنبؤ بأن لا أحد مريض) هو الأرضية التي يجب على أي نموذج حقيقي أن يتجاوزها — استشهد به قبل الاستشهاد بالدقة.
- غداً، في الفصل 8، يفسح البديل المؤقت المجال للشيء الحقيقي: ندرّب مصيِّفات الانحدار اللوجستي وشجرة القرار على أقسام اليوم المحضّرة بإتقان، ونقرأ تنبؤاتها بوصفها قرارات تشخيصية.

د. حسام صلاح مهدي
رئيس قسم الذكاء الاصطناعي — جامعة المصطفى
Dr. Husam Salah Mahdi — Head of AI Department, Al-Mustafa University

باب 8

اليوم 8 — بناء نماذج التصنيف التشخيصية

في كل يوم عمل يجيب الأطباء في مستشفياتنا عن السؤال الصامت نفسه مئات المرات: «هل هذا المريض مصاب بالمرض أم لا؟» فالطفل المحموم في البصرة إما أن يكون مصاباً بعدوى بكتيرية أو لا يكون، والمريض الخمسيني اللاهث في بغداد إما أن يكون مصاباً بمرض الشرايين التاجية أو لا يكون. بالأمس، في الفصل 7، تعلمنا كيف تتعلم الآلة أصلاً: كيف نقسم فوج المرضى إلى مجموعة التدريب ومجموعة الاختبار، وكيف نعاير الميزات معاً عادلة، ولماذا يجب ألا تتسرب أي معلومة من مجموعة الاختبار إلى التدريب. اليوم نضع هذه الآلية موضع العمل ونبنى أولى مصنفاتنا التشخيصية الحقيقية. سندرب ثلاثة نماذج مختلفة جوهرياً — الانحدار اللوجستي (Logistic Regression)، والغابة العشوائية (Random Forest)، وآلة المتجهات الداعمة (SVM) — على فوج تحري السكري نفسه، وسنرى أن أكثر المخرجات فائدة سريراً ليس تصنيفاً مجرداً بل «احتمال الإصابة بالمرض» الذي يستطيع الطبيب موازنته مع الأعراض والسوابق المرضية وكلفة الخطأ. وبنهاية الجلسة سيكون فريقك قد نفذ تجربة استرشادية حقيقية لاختيار النموذج على مجموعة بيانات القلبية، وستفهم لماذا يستحق النموذج الذي «يدعم» الطبيب أن يُنشر في الخدمة، بينما لا يستحق ذلك النموذج الذي يحاول أن «يحل محل» الطبيب.

الدقائق	الجزء	المحتوى والأنشطة
00-05	المراجعة والمدخل	مراجعة اليوم 7: تقسيم التدريب/الاختبار، المعايرة، تسرب البيانات (Data Leakage)؛ سؤال اليوم: أي فئة، وبأي احتمال؟
05-12	التصنيف والانحدار اللوجستي	التنبؤ بفئة تشخيصية: الدالة السينية تضغط مجموعاً موزوناً إلى احتمال بين 0 و1؛ المعاملات بوصفها عوامل خطورة قابلة للتدقيق
12-17	الغابات العشوائية	لجنة من أشجار القرار: عينات التمهيد، مجموعات جزئية من الميزات، تصويت الأغلبية؛ أهميات الميزات مقابل الشفافية
17-20	آلة المتجهات الداعمة ودراسة حالة واقعية	الفصل بأوسع هامش؛ نموذج خط اليد لعام 2024 لتحري الزهايمر (دقة 94.49%) وما تعنيه هذه الأرقام وما لا تعنيه
20-30	التطبيق الموجّه 1	إعادة بناء مسار عمل اليوم 7 على diabetes_dataset.csv؛ ملاءمة LogisticRegression؛ أولى التنبؤات مقابل الحقيقة
30-40	التطبيق الموجّه 2	نسب الخطورة المئوية عبر predict_proba لخمس مريض؛ جدول المعاملات بوصفه تدقيقاً سريراً
40-50	التطبيق الموجّه 3	RandomForestClassifier بمئتي شجرة؛ أهميات الميزات مرتبة
50-60	التطبيق الموجّه 4	SVC مع مخرجات احتمالية؛ لوحة نتائج الدقة للنماذج الثلاثة؛ لماذا الدقة هي أقل المقاييس السريرية إفادة
60-85	التحدي السريري	التجربة الاسترشادية للمديرية على heart_risk.csv: هندسة الميزات، تدريب النموذج الثلاثة، ترتيب الأهميات، تحديد المرضى الثلاثة الأعلى خطورة، صياغة مذكرة اختيار
85-90	المناقشة الختامية والتمهيد	نقاش جماعي حول المذكرات؛ الواجب المنزلي؛ تشويق لليوم 9: الحساسية والنوعية ROC ومنحنى

✓ الأهداف التعليمية

بنهاية هذه الجلسة ستكون قادراً على:

1. أن تشرح، بلغة سريرية، ما الذي يتنبأ به نموذج التصنيف، ولماذا يكون احتمال الإصابة بالمرض أنفع عند سرير المريض من تصنيف مجرد 0/1.
2. أن تدرب نموذج LogisticRegression على ميزات معايرة وتستخرج احتمالات المرض لكل مريض عبر predict_proba.
3. أن تقرأ جدول معاملات الانحدار اللوجستي بوصفه قائمة عوامل خطورة اتجاهية، وتدقق ما إذا كان النموذج يتفق مع المعرفة الطبية.
4. أن تدرب RandomForestClassifier وتفسر ترتيب feature_importances_ في ضوء الحدس السريري.
5. أن تدرب آلة متجهات داعمة عبر SVC(probability=True) وتصوغ في جملة واحدة ما الذي تمنحنا إياه فكرة الهامش الأقصى.
6. أن تبني لوحة نتائج دقة جنباً إلى جنب لثلاثة نماذج، وتشرح لماذا يجب ألا تُحسم قرارات نشر أي نموذج تشخيصي بالدقة وحدها.
7. أن توصي، في مذكرة مكتوبة قصيرة، بعائلة النموذج التي ينبغي أن تجربها المديرية استرشادياً، مستشهداً بمعيار واحد على الأقل لا تستطيع الدقة التقاطه.

➔ حقيبة أدوات الذكاء الاصطناعي والاعتماديات — حزمة التصنيف لليوم 8

تستخدم جلسة اليوم حزمة المكتبات الأربع نفسها المعتمدة في اليوم 7، وتمتد الآن إلى ثلاث زوايا جديدة من مكتبة scikit-learn: النماذج الخطية، والنماذج التجميعية، وآلات المتجهات الداعمة.

```
1 import pandas as pd
2 import numpy as np
3 from sklearn.model_selection import train_test_split
4 from sklearn.preprocessing import StandardScaler
5 from sklearn.linear_model import LogisticRegression
6 from sklearn.ensemble import RandomForestClassifier
7 from sklearn.svm import SVC
8 from sklearn.metrics import accuracy_score
```

```
pip install pandas numpy scikit-learn
```

تبقى **pandas** العمود الفقري لبياناتنا السريرية: فهي تحمّل ملفات CSV الخاصة بالسجلات وتبني جداول النتائج المرتبة (قوائم الخطورة، وتدقيقات المعاملات) التي تذهب مباشرة إلى تقرير للمديرية. وتوفر **numpy** الحساب الموجه — واليوم على وجه الخصوص **np.where** لتسمية اتجاهات المعاملات دون حلقة تكرارية واحدة. أما **scikit-learn** فتقدم عائلات النماذج الثلاث كلها خلف واجهة موحدة واحدة **predict/fit**، ما يعني أن موظف بيانات المستشفى الذي يتقن نموذجاً واحداً يستطيع تشغيلها جميعاً؛ وهذه الوحدة في الواجهة هي، بالنسبة إلى مديرية صحية، ما يجعل التجارب المقارنة بين النماذج زهيدة الكلفة وقابلة للتكرار. وتمنحنا الدالة المساعدة **accuracy_score** من **sklearn.metrics** أول رقم أداء لنا — وهو رقم مؤقت عن عمد.

8.1 السياق السريري والإطار النظري (20 دقيقة)

من القياسات إلى فئة تشخيصية

علمتنا الأيام من 1 إلى 6 تنظيف البيانات المختبرية ووصفها وعرضها بيانياً؛ وعلمنا اليوم 7 انضباط التقييم النزيه. أما خطوة اليوم فهي صغيرة مفهوماً لكنها هائلة سريرياً: نطلب من الحاسوب أن يحوّل صف قياسات المريض إلى «فئة تشخيصية». هذا هو المصنّف بعينه.

التصنيف: التنبؤ بفئة تشخيصية

«نموذج التصنيف» دالة تأخذ متجه ميزات المريض — العمر والغلوكوز ومؤشر كتلة الجسم وضغط الدم — وتعيد فئة متنبأ بها: مصاب (1) أو غير مصاب (0). وثمة مخرجان مهمان: التصنيف المتنبأ به، وهو قرار قاطع 0/1، والاحتمال المتنبأ به، وهو عدد بين 0 و1 يعبر عن مدى قوة دلالة الشواهد على المرض. والاحتمال هو الأغنى سريراً بين الاثنين. فتقرير يقول «خطورة إصابة هذا المريض بالسكري 82%» يستدعي اختبار تأكيداً صائماً وجلسة إرشاد صحي: أما «1» مجردة فلا تستدعي شيئاً سوى الثقة العمياء. وعلى امتداد هذه الدورة يُشتق التصنيف من الاحتمال بمقارنته بعتبة القرار (Threshold) — وهي افتراضياً 0.5 — وسببين اليوم 9 أن هذا الافتراض عُرف لا قانون.

الانحدار اللوجستي: حسان العمل في التنبؤ السريري

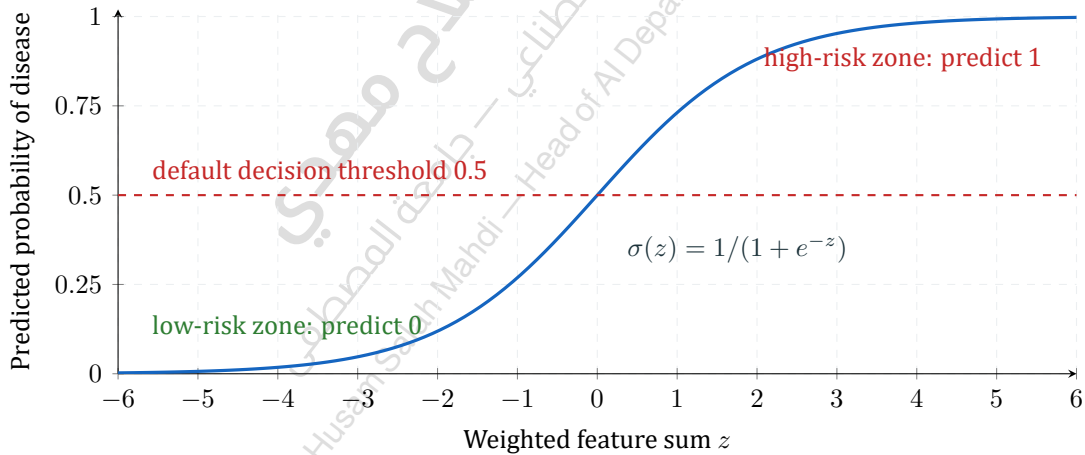
يحسب الانحدار اللوجستي أولاً مجموعاً موزوناً للميزات (المعاصرة)،

$$z = w_1x_1 + w_2x_2 + \dots + w_px_p + b,$$

تماماً مثل درجات الخطورة التي يعرفها الأطباء أصلاً (CHA₂DS₂-VASC و Wells و APACHE). ثم يمرر z عبر الدالة السينية (Sigmoid)،

$$\sigma(z) = \frac{1}{1 + e^{-z}},$$

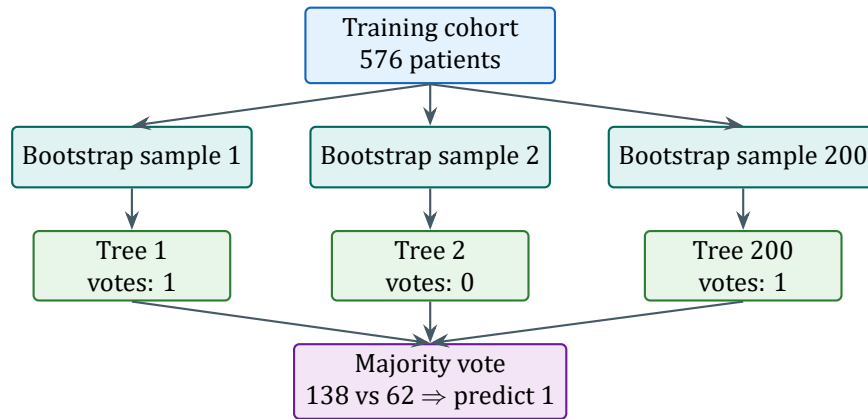
التي تضغط أي عدد حقيقي إلى المجال (0, 1) بحيث يمكن قراءته احتمالاً للإصابة بالمرض (الشكل 8.1). وتُتعلم الأوزان w من مرضى التدريب، وكل وزن منها «عامل خطورة اتجاهاً»: فالمعامل الموجب يعني أن الميزة ترفع الخطورة المتنبأ بها، والسالب يعني أنها تخفضها، والمقدار الأكبر (على ميزات معاصرة قياسياً) يعني تأثيراً أقوى. ولهذا يبقى الانحدار اللوجستي النموذج المرجعي في المجالات السريرية: يستطيع الطبيب أن يقرأ جدول المعاملات سطرًا سطرًا وأن «يدقق» هل تعلم النموذج الطب أم حفظ الضجيج.



شكل 8.1: الدالة السينية تحوّل مجموعاً موزوناً لميزات المريض إلى احتمال للإصابة بالمرض. يشير الخط المتقطع إلى العتبة الافتراضية 0.5؛ والمرضى فوقها يُصنّفون إيجابيين. سيبن اليوم 9 لماذا يجب ضبط هذه العتبة وفق الكلفة السريرية للحالة الفائتة.

الغابة العشوائية: لجنة من أشجار القرار

تطرح شجرة القرار الواحدة سلسلة من أسئلة نعم/لا («هل الغلوكوز أعلى من 144؟ هل العمر فوق 50؟») وهي سهلة القراءة لكنها غير مستقرة: فحفنة من مرضى تدريب مختلفين قد تعيد تشكيل الشجرة كلها. وتعالج «الغابة العشوائية» ذلك بالديمقراطية: إذ تُنهي مئات الأشجار، تتدرب كل منها على «عينة تمهيد» (Bootstrap Sample) مختلفة — مرضى مسحوبين مع الإرجاع من فوج التدريب — ولا تنظر كل شجرة عند كل تفرع إلا في مجموعة جزئية عشوائية من الميزات. ويكون التشخيص النهائي هو «تصويت الأغلبية» في اللجنة (الشكل 8.2). يلتقط هذا التجميع الأنماط اللاخطية وتفاعلات الميزات التي يفوتها النموذج المستقيم، ويقدم `feature_importances_`: أي مقدار إسهام كل ميزة في قرارات اللجنة. أما الثمن فهو الشفافية — فلا يستطيع أي طبيب قراءة 200 شجرة، ولذلك يجب تدقيق الغابة تدقيقاً غير مباشر عبر أهمياتها، وغداً عبر بنية أخطائها.



شكل 8.2: الغابة العشوائية بوصفها لجنة تشخيصية: تتدرب كل شجرة على عينة تمهيد مختلفة من المرضى مع مجموعات جزئية عشوائية من الميزات عند كل فرع، ويكون التصنيف النهائي هو تصويت أغلبية الأشجار المتتين جميعها.

آلة المتجهات الداعمة: أوسع هامش أمان

ترسم آلة المتجهات الداعمة الحدّ الفاصل بين الفئتين الذي يترك «أوسع هامش ممكن» — أكبر ممر أمان — بين أقرب المرضى المصابين وغير المصابين. وهؤلاء المرضى الأقرب، أي «المتجهات الداعمة»، هم الحالات الحدية التي تحدد القرار؛ ويمكن لبقية الفوج أن يختفي دون أن يتحرك الحدّ الفاصل. ومع نواة الأساس الشعاعي (RBF) يمكن للحد أن ينحني فيلتقط البنية اللاخطية. تتألق آلات المتجهات الداعمة حين تكون الفئتان قابلتين للفصل بهامش نظيف في فضاء محوّل ما، لكن النموذج الملاءم كائن هندسي بلا عوامل خطورة قابلة للقراءة المباشرة، ولا تتوفر الاحتمالات إلا عبر خطوة معايرة إضافية (probability=True)، ولهذا تقع آلة المتجهات الداعمة عند الطرف المنخفض الشفافية من الجدول 8.1.

★ لؤلؤة سريرية — عتبات التشخيص وراء تصنيفات اليوم

يرمز العمود outcome في ملف diabetes_dataset.csv تشخيصاً يستند عملياً إلى عتبات متفق عليها دولياً، يجب في النهاية أن تُقابل بها كل مخرجات النموذج:

- غلوكوز البلازما الصائم ≥ 126 ملغ/دل في مناسبتين: سكري؛ 100–125 ملغ/دل: اختلال الغلوكوز الصائم. (كما في اليوم 2، نتعامل مع عمود glucose في مجموعة بيانات اليوم على أنه قيمة صائمة تُقرأ مقابل خط 126 ملغ/دل).
 - غلوكوز البلازما بعد ساعتين في اختبار تحمل الغلوكوز الفموي (OGTT) ≥ 200 ملغ/دل: سكري؛ 140–199 ملغ/دل: اختلال تحمل الغلوكوز.
 - السكر التراكمي HbA1c $\geq 6.5\%$: سكري؛ 5.7–6.4%: ما قبل السكري.
 - مؤشر كتلة الجسم (BMI) ≥ 30 kg/m²: سمنة، وهي عامل الخطورة الأهم القابل للتعديل.
- النموذج الذي لا تكون عوامل خطورته الأولى هي الغلوكوز ومؤشر كتلة الجسم والعمر ينبغي التعامل معه بريبة مهما علت دقته — فتدقيق النموذج مقابل العتبات التشخيصية الراسخة هو خط الأمان السريري الأول.

★ لؤلؤة سريرية — نموذج SVM حقيقي في العيادة: خط اليد وداء الزهايمر

درّبت دراسة نُشرت عام 2024 (مسودة على TechRxiv، مجموعة بيانات خط اليد DARWIN) آلة متجهات داعمة على ميزات مستخلصة من مهام كتابة يدوية بسيطة بالقلم والورقة، وأفادت بدقة 94.49% وحساسية 93.53% في كشف داء الزهايمر المبكر — تحزّ غير باضع ومنخفض الكلفة لا يحتاج إلى جهاز تصوير ولا إلى بزل قطي. ناقش مع جارك قبل مواصلة القراءة: ماذا «تعني» هذه الأرقام، وماذا «لا تعني»؟ إنها تعني أن النموذج واعد على الفوج الذي اختُبر عليه. وهي «لا تعني» أن 94.49% من المرضى المؤشّر عليهم مصابون بالزهايمر (فذلك هو القيمة التنبؤية الإيجابية (PPV) التي تنهار عند انخفاض الانتشار)، ولا أن النموذج ينتقل إلى مرضى عراقيين يكتبون بالخط العربي، ولا أنه جرى التحقق منه استباقياً. واليوم 9 (الفصل 9) يمنحك العدة الكاملة لاستجواب مثل هذه الادعاءات بدقة.

الاختيار بين العائلات الثلاث

اختيار النموذج موازنة بين قابلية التفسير والمرونة، وفي بيئة وزارة الصحة تحمل قابلية التفسير وزناً تنظيمياً وأخلاقياً؛ فالنموذج الذي يضع مريضاً على قائمة مراجعة ذات أولوية يجب أن يكون قابلاً للشرح أمام طبيب ذلك المريض. ويلخص الجدول 8.1 ما سنراه مباشرة في الدقائق الأربعين القادمة.

جدول 8.1: العائلات الثلاث للمصنفات في لمحة — الموازنات التي يجب أن تزنها التجربة الاسترشادية للمديرة.

الخاصية	الانحدار اللوجستي	الغابة العشوائية	SVM (RBF)
المخرجات	احتمال	حصة الأصوات	مسافة الهامش ^a
قابلية التفسير	عالية (المعاملات)	متوسطة (الأهميات)	منخفضة
الأنماط اللاخطية	لا (افتراضياً)	نعم	نعم
كلفة التدريب	منخفضة جداً	معتدلة	مرتفعة مع البيانات الكبيرة
التدقيق السريري	قراءة الأوزان	ترتيب الأهميات	غير مباشر فقط
الدور المعتاد	النموذج المرجعي	خط أساس قوي	اختصاصي للحالات الخاصة

^a لا تُحوّل إلى احتمال إلا عبر معيار probability=True .

وأياً كانت العائلة التي تفوز بالتجربة الاسترشادية، فثمة مبدأ ثابت: النموذج «يدعم» الطبيب ولا يحل محله. فالاحتمال المتنبأ به معطى إضافي واحد على لوحة المريض — مثل قيمة اللاكتات أو التروبونين — والإنسان الذي يدمجه مع بقية المعطيات هو صاحب القرار.

8.2 التطبيق الموجّه (40 دقيقة)

■ عرض تطبيقي مباشر — ثلاثة مصنفات على فوج تحرّ واحد

الهدف: على `diabetes_dataset.csv`، نعيد بناء مسار عمل الأمس بالتقسيم والمعايرة، ثم ندرّب الانحدار اللوجستي وغابة عشوائية بمئتي شجرة وآلة متجهات داعمة بنواة RBF؛ ونستخرج نسب الخطورة المثوية لكل مريض، وندقق معاملات النموذج اللوجستي، ونرتب أهميات ميزات الغابة، ونختم بلوحة نتائج دقة للنماذج الثلاثة — مع التحذير من أن الدقة هي أقل المقاييس السريرية إفادةً. يكتب المتدربون مع المدرب سطرًا بسطر؛ وكل ناتج أدناه هو مخرج الحاسوب الفعلي.

الخطوة 1: إعادة بناء مسار عمل اليوم 7

نبدأ تماماً من حيث انتهى الفصل 7: تقسيم طبقي بنسبة 75/25 ومعايرة قياسية مُلاءمة على مرضى التدريب وحدهم. يحافظ التقسيم الطبقي على انتشار السكري في النصفين معاً؛ وملاءمة المعايير على مجموعة التدريب وحدها تمنع تسرب مجموعة الاختبار.

```
1 import pandas as pd
2 import numpy as np
3 from sklearn.model_selection import train_test_split
4 from sklearn.preprocessing import StandardScaler
5
6 pd.set_option("display.width", 100)
7
8 diab = pd.read_csv("diabetes_dataset.csv")
9 X = diab.drop(columns="outcome")
10 y = diab["outcome"]
11
12 X_train, X_test, y_train, y_test = train_test_split(
13     X, y, test_size=0.25, stratify=y, random_state=42)
14
15 scaler = StandardScaler()
16 X_train_scaled = scaler.fit_transform(X_train)
17 X_test_scaled = scaler.transform(X_test)
18
```

```

19 print("Training patients:", X_train.shape[0])
20 print("Test patients:      ", X_test.shape[0])
21 print("Diabetes prevalence in test set:", round(y_test.mean(), 3))

```

شيفرة 8.1: تقسيم فوج التحري ومعايرته

```

Training patients: 576
Test patients:      192
Diabetes prevalence in test set: 0.37

```

مرضى الاختبار البالغ عددهم 192 هم «عيادة المستقبل» المحاكاة لدينا: لا يرى أي نموذج نتائجهم أثناء التدريب، ولذلك فكل درجة أداء أدناه تقدير نزيه.

الخطوة 2: ملاءمة الانحدار اللوجستي والتنبؤ

سطر واحد يدرب حصان العمل. الخيار `max_iter=1000` يمنح المحسّن فسحة كافية للتقارب لا أكثر؛ و `random_state` يثبّت قابلية التكرار بحيث يحصل زميل في محافظة أخرى على نتائج مطابقة.

```

1 from sklearn.linear_model import LogisticRegression
2
3 log_model = LogisticRegression(max_iter=1000, random_state=42)
4 log_model.fit(X_train_scaled, y_train)
5
6 y_pred_log = log_model.predict(X_test_scaled)
7 print("First 10 predicted labels:", y_pred_log[:10])
8 print("First 10 true labels:      ", y_test.values[:10])

```

شيفرة 8.2: تدريب النموذج اللوجستي وتصنيف مرضى الاختبار

```

First 10 predicted labels: [0 0 1 0 0 0 0 1 0 0]
First 10 true labels:     [0 0 1 0 1 0 1 1 0 0]

```

اقرأ الصفين كما تقرأ تدقيقاً مختبرياً: المواقع من 1 إلى 4 متطابقة، لكن النموذج يحكم على المريض الخامس بأنه سليم بينما يقول السجل إنه سكري — هذه أول «حالة فائتة» لدينا. أمسك بهذه الملاحظة؛ فهي قلب اليوم 9.

الخطوة 3: من التصنيفات إلى نسب الخطورة المئوية

تعيد `predict_proba` عمودين لكل مريض: احتمال الفئة 0 واحتمال الفئة 1. والعمود ذو الدليل 1 هو احتمال المرض. نبني جدول خطورة قابلاً للقراءة عند سرير المريض لأول خمسة من مرضى الاختبار، ومعه النتيجة الحقيقية جنباً إلى جنب توخياً للنزاهة.

```

1 risk = log_model.predict_proba(X_test_scaled)[: , 1]
2
3 review = pd.DataFrame({
4     "glucose": X_test["glucose"].values,
5     "bmi": X_test["bmi"].values,
6     "age": X_test["age"].values,
7     "risk_pct": (risk * 100).round(1),
8     "predicted": y_pred_log,
9     "true_outcome": y_test.values,
10 })
11 print(review.head(5).to_string(index=False))

```

شيفرة 8.3: نسب الخطورة المئوية لكل مريض إلى جانب النتائج الحقيقية

glucose	bmi	age	risk_pct	predicted	true_outcome
108	33.5	38	29.5	0	0
88	30.1	30	9.0	0	0
151	48.2	65	96.8	1	1
129	26.4	31	21.9	0	0
89	31.5	45	17.5	0	1

هذا الجدول هو الثمرة السريرية لليوم كله. فالمريض الثالث يحمل خطورة متنبأ بها قدرها 96.8% — غلوكوز 151 ملغ/دل، ومؤشر كتلة جسم 48.2، وعمر 65 — وهو بالفعل سكري: حالة واضحة أصاب النموذج فيها. أما المريضة الخامسة فهي الحالة التعليمية: خطورتها المتنبأ بها 17.5% فقط، فتصنفها عتبة 0.5 سليمة، بينما يقول السجل إنها سكرية. الاحتمال يتيح لنا أن «نرى» أنها لم تكن خطأً واثقاً بل حكماً حديثاً — و يتيح لليوم 9 أن يسأل: هل ينبغي لبرنامج تحزُّ أن يخفض العتبة لالتقاطها؟

✖ خطأ شائع — عمود الاحتمال الخاطئ

تعيد predict_proba احتمالات «الفئتين معاً» مرتبةً وفق تسمية الفئات: العمود 0 هو $P(\text{disease no})$ والعمود 1 هو $P(\text{disease})$. وكتابة `predict_proba(X_test_scaled)[:, 0]` تجعل التقرير يذكر بصمت احتمال «السلامة» وكأنه الخطورة — فيتحول كل مريض منخفض الخطورة إلى مرتفع الخطورة والعكس بالعكس، وهو انقلاب كارثي في قائمة فرز لكنه غير مرئي في الشيفرة. تحقق دائماً من ترتيب العمودين عبر `log_model.classes_` (وهنا `[0 1]`) قبل أن يغادر أي جدول خطورة مكتبك. وثمة خطأ كلاسيكي ثانٍ: تمرير `X_test` الخام بدلاً من `X_test_scaled` إلى نموذج دُرب على ميزات معايرة — ستظهر تنبؤات على أي حال، لكنها بلا معنى عددي.

الخطوة 4: تدقيق المعاملات بوصفها عوامل خطورة

والآن عائد قابلية التفسير. نقرن كل ميزة بوزنها المتعلم ونسعي اتجاه الخطورة عبر `np.where` الموجبة — لا حلقات ولا جمل شرطية، تماماً كما تشترط هذه الدورة.

```
1 coef_table = pd.DataFrame({
2     "feature": X.columns,
3     "coefficient": log_model.coef_[0].round(3),
4 })
5 coef_table["direction"] = np.where(
6     coef_table["coefficient"] > 0, "raises risk", "lowers risk")
7 coef_table = coef_table.sort_values("coefficient", ascending=False)
8 print(coef_table.to_string(index=False))
```

شيفرة 8.4: جدول المعاملات: عوامل خطورة اتجاهية

feature	coefficient	direction
glucose	1.099	raises risk
age	0.614	raises risk
bmi	0.596	raises risk
diabetes_pedigree	0.107	raises risk
blood_pressure	0.089	raises risk
skin_thickness	0.022	raises risk
pregnancies	-0.088	lowers risk
insulin	-0.095	lowers risk

التدقيق ناجح: الغلوكوز هو الوزن الموجب الأقوى بفارق كبير، يليه العمر ومؤشر كتلة الجسم — وهذا بالضبط هو الثلاثي الذي كانت المعايير التشخيصية في اللؤلؤة السريرية أعلاه ستنبأ به. أما الوزنان السالبان الصغيران لعدد الحمل والإنسولين فهما قريبان من الصفر وينبغي أن يُقرأ على أنهما «لا أثر مستقلاً بعد التعديل على بقية الميزات»، لا على أنهما عاملان واقيان؛ وهذا التمييز — بين المعامل والادعاء السببي — تمييز يجب أن تصرح به تقاريرك المستقبلية دائماً.

الخطوة 5: عقد اجتماع الغابة

مثلاً شجرة، لجنة واحدة. الواجهة متطابقة عمداً مع الانحدار اللوجستي — `fit` ثم `predict` — وبعد ذلك نرتب ما وجدته اللجنة مهماً.

```
1 from sklearn.ensemble import RandomForestClassifier
2
3 rf_model = RandomForestClassifier(n_estimators=200, random_state=42)
4 rf_model.fit(X_train_scaled, y_train)
5 y_pred_rf = rf_model.predict(X_test_scaled)
6
7 importance = pd.DataFrame({
8     "feature": X.columns,
9     "importance": rf_model.feature_importances_.round(3),
10 }).sort_values("importance", ascending=False)
11 print(importance.to_string(index=False))
```

شيفرة 8.5: الغابة العشوائية وأهميات ميزاتها مرتبة

feature	importance
glucose	0.224
bmi	0.210
age	0.129
insulin	0.121
diabetes_pedigree	0.097
blood_pressure	0.087
skin_thickness	0.079
pregnancies	0.053

تتفق الغابة مع التدقيق اللوجستي على منصة التنوع — الغلوكوز ومؤشر كتلة الجسم والعمر — لكن لاحظ أن الإنسولين، الذي كان شبه خفي عن النموذج الخطي، يحتل هنا المرتبة الرابعة: فاللجنة تستثمر بنية لاختية (الإنسولين المنخفض جداً والمرتفع جداً كلاهما ينذر باضطراب التنظيم) لا يستطيع مجموع موزون مستقيم التعبير عنها. الأهميات مجموعها 1 وهي تجيب عن سؤال «كم استخدمت هذه الميزة»، لا عن سؤال «في أي اتجاه تدفع» — فالاتجاه هبة النموذج الخطي. ويضع الجدول 8.2 التدقيقين جنباً إلى جنب.

جدول 8.2: تدقيقان لنموذجين على الفوج نفسه: معاملات الانحدار اللوجستي (اتجاهية) وأهميات الغابة العشوائية (مقدار فقط). القيم حقيقية من القائمتين 8.4 و 8.5.

الميزة	المعامل اللوجستي	الاتجاه	أهمية الغابة
glucose	1.099	يرفع الخطورة	0.224
bmi	0.596	يرفع الخطورة	0.210
age	0.614	يرفع الخطورة	0.129
insulin	-0.095	يخفض الخطورة	0.121
diabetes_pedigree	0.107	يرفع الخطورة	0.097
blood_pressure	0.089	يرفع الخطورة	0.087
skin_thickness	0.022	يرفع الخطورة	0.079
pregnancies	-0.088	يخفض الخطورة	0.053

الخطوة 6: اختصاي الهامش الأقصى

تكمل آلة المتجهات الداعمة ثلاثيناً. يضيف الخيار `probability=True` معياراً داخلية تجعل `predict_proba` تعمل لاحقاً؛ وهو بسيط التدريب، وذلك أمر لا يُلاحظ مع 576 مريضاً لكنه يصبح مهماً مع نصف مليون.

```
1 from sklearn.svm import SVC
2
3 svm_model = SVC(kernel="rbf", probability=True, random_state=42)
```

```

4 svm_model.fit(X_train_scaled, y_train)
5 y_pred_svm = svm_model.predict(X_test_scaled)
6 print("SVM predictions ready:", y_pred_svm.shape[0], "patients")

```

شيفرة 8.6: آلة المتجهات الداعمة مع معايير الاحتمالات

```
SVM predictions ready: 192 patients
```

آلة المتجهات الداعمة حساسة فعلاً لمعايير الميزات: فهو أمشيها مسافات إقليدية، ولذلك فإن محور غلوكونز غير معايير (المدى 211-44 في هذا الفوج) سيطغى على محور الاستعداد الوراثي غير المعايير (المدى 0.08-2.50) بمئة ضعف. وعادة اليوم 7 — المعايير دائماً — هي ما يجعل مقارنة اليوم عادلة.

الخطوة 7: لوحة النتائج — وفحصها

ثلاثة نماذج، مجموعة اختبار نزيهة واحدة، رقم واحد لكل نموذج.

```

1 from sklearn.metrics import accuracy_score
2
3 scoreboard = pd.DataFrame({
4     "model": ["Logistic regression", "Random forest", "SVM (RBF)"],
5     "test_accuracy": [
6         accuracy_score(y_test, y_pred_log),
7         accuracy_score(y_test, y_pred_rf),
8         accuracy_score(y_test, y_pred_svm),
9     ],
10 })
11 print(scoreboard.to_string(index=False))

```

شيفرة 8.7: لوحة نتائج الدقة للنماذج الثلاثة على المرضى المحتجزين

model	test_accuracy
Logistic regression	0.802
Random forest	0.797
SVM (RBF)	0.771

يفوز النموذج الأبسط والأكثر شفافية — وهي نتيجة شائعة في البيانات السريرية الجدولية حتى إن لها اسماً في الأدبيات: «الفعالية غير المعقولة للانحدار اللوجستي». لكن ها هو الفخ. افترض أن نموذجاً كسولاً تنبأ بعدم وجود سكري لكل مريض بلا استثناء. مع انتشار قدره 37% سيحرز دقة $1 - 0.37 = 0.63$ وهو يفوت «كل مريض مصاب في العيادة». إن الدقة تخلط الخطأين اللذين يصير الطب على فصلهما — تفويت مريض مصاب وإفزاز مريض سليم — في رقم واحد يكافئ تجاهل الفئة الأقلية. ولذلك فهي «أقل» المقاييس السريرية إفادة، وكل ما في جلسة الغد وجد ليحل محلها. حتى اليوم 9، تعامل مع لوحة نتائج اليوم بوصفها بيئة مؤقتة لا أكثر.

8.3 التحدي السريري (30 دقيقة)

التحدي السريري — التجربة الاسترشادية لأمراض القلب في المديرية

تدرس مديرية التشخيص أداة دعم قرار تؤثر على حالات مرض الشرايين التاجية (CAD) المحتملة بين المرضى المحالين لتقييم ألم الصدر، باستخدام القياسات الروتينية في heart_risk.csv. أنتم فريق التجربة. بالعمل في ثنائيات، وبإعادة استخدام هندسة الميزات (Feature Engineering) التي بنيتها في الفصل 7 (الترميز Encoding) الرقمي لعمودي sex و exercise_angina، والترميز الأحادي لعمود chest_pain_type:

1. حملوا heart_risk.csv، وهندسوا الميزات الفئوية عبر map و pd.get_dummies، وأبلغوا عن العدد النهائي لأعمدة الميزات.
2. أنشئوا تقسيماً طبقياً بنسبة 75/25 (random_state=42) وعايروا عبر StandardScaler مُلاءم على مرضى التدريب

وحدهم.

3. درّبوا نماذج اليوم الثلاثة كلها وأنتجوا لوحة نتائج دقة واحدة على مرضى الاختبار.
4. استخرجوا أعلى خمس أهميات ميزات في الغابة العشوائية، واذكروا في جملة واحدة لكل منها هل تطابق الحدس القلبي.
5. اذكروا مرضى الاختبار الثلاثة أصحاب الاحتمال اللوجستي المتنبأ به الأعلى للإصابة بمرض الشرايين التاجية، مع معرفات صفوفهم وقيم الميزات التي يود طبيب القلبية رؤيتها، بوصفها قائمة مراجعة ذات أولوية.
6. صوغوا مذكرة من خمس جمل إلى المدير: أي نموذج سترشحون للتجربة ولماذا؟ سمّوا معياراً واحداً على الأقل للاختيار لا تستطيع دقة الاختبار التقاطه.

✓ حل المدرب — التجربة الاسترشادية لأمراض القلب في المديرية

المهمتان الأولى والثانية: الهندسة والتقسيم النزيه. كل شيء موجّه: `map`. تترجم الرموز الثنائية، و `pd.get_dummies` توسع فئات ألم الصدر الأربع إلى أعمدة مؤشرات. وكما في اليوم 7، نعيد اشتقاق الميزات من الملف الخام `heart_risk.csv` حتى تكون الجلسة مكثفية بذاتها — وقد تختلف أسماء الأعمدة قليلاً عن ملف `heart_risk_engineered.csv` الذي حفظتموه في اليوم 6.

```
1 heart = pd.read_csv("heart_risk.csv")
2 heart["sex_male"] = heart["sex"].map({"M": 1, "F": 0})
3 heart["angina"] = heart["exercise_angina"].map({"Y": 1, "N": 0})
4 heart = pd.get_dummies(heart, columns=["chest_pain_type"], dtype=int)
5
6 X = heart.drop(columns=["target", "sex", "exercise_angina"])
7 y = heart["target"]
8 print("Engineered features:", X.shape[1])
9 print(list(X.columns[:6]))
10 print(list(X.columns[6:9]))
11 print(list(X.columns[9:]))
12
13 X_train, X_test, y_train, y_test = train_test_split(
14     X, y, test_size=0.25, stratify=y, random_state=42)
15
16 scaler = StandardScaler()
17 X_train_scaled = scaler.fit_transform(X_train)
18 X_test_scaled = scaler.transform(X_test)
19 print("Train:", X_train.shape[0], " Test:", X_test.shape[0])
20 print("CAD prevalence (test):", round(y_test.mean(), 3))
```

شيفرة 8.8: المهمتان الأولى والثانية: الهندسة والتقسيم والمعايرة

```
Engineered features: 12
['age', 'resting_bp', 'cholesterol', 'fasting_bs', 'max_hr', 'st_depression']
['sex_male', 'angina', 'chest_pain_type_ASY']
['chest_pain_type_ATA', 'chest_pain_type_NAP', 'chest_pain_type_TA']
Train: 675 Test: 225
CAD prevalence (test): 0.427
```

المهمة الثالثة: تدريب الثلاثي وبناء لوحة النتائج.

```
1 log_model = LogisticRegression(max_iter=1000, random_state=42)
2 rf_model = RandomForestClassifier(n_estimators=200, random_state=42)
3 svm_model = SVC(kernel="rbf", probability=True, random_state=42)
4
5 log_model.fit(X_train_scaled, y_train)
6 rf_model.fit(X_train_scaled, y_train)
```

```

7 svm_model.fit(X_train_scaled, y_train)
8
9 scoreboard = pd.DataFrame({
10     "model": ["Logistic regression", "Random forest", "SVM (RBF)"],
11     "test_accuracy": [
12         accuracy_score(y_test, log_model.predict(X_test_scaled)),
13         accuracy_score(y_test, rf_model.predict(X_test_scaled)),
14         accuracy_score(y_test, svm_model.predict(X_test_scaled)),
15     ],
16 }).round(3)
17 print(scoreboard.to_string(index=False))

```

شيفرة 8.9: المهمة الثالثة: ثلاثة نماذج ولوحة نتائج واحدة

model	test_accuracy
Logistic regression	0.738
Random forest	0.653
SVM (RBF)	0.676

المهمة الرابعة: أعلى خمسة مؤشرات خطورة لدى اللجنة.

```

1 importance = pd.DataFrame({
2     "feature": X.columns,
3     "importance": rf_model.feature_importances_.round(3),
4 }).sort_values("importance", ascending=False)
5 print(importance.head(5).to_string(index=False))

```

شيفرة 8.10: المهمة الرابعة: أعلى خمس أهميات في الغابة العشوائية

feature	importance
max_hr	0.191
age	0.184
cholesterol	0.166
resting_bp	0.148
st_depression	0.146

المهمة الخامسة: قائمة المراجعة ذات الأولوية.

```

1 risk = log_model.predict_proba(X_test_scaled)[: , 1]
2 triage = X_test[["age", "sex_male", "st_depression",
3     "max_hr", "angina"]].copy()
4 triage["risk_pct"] = (risk * 100).round(1)
5 triage["true_target"] = y_test.values
6 top3 = triage.sort_values("risk_pct", ascending=False).head(3)
7 print(top3.to_string())

```

شيفرة 8.11: المهمة الخامسة: المرضى الثلاثة الأعلى خطورة لمراجعة القلبية

	age	sex_male	st_depression	max_hr	angina	risk_pct	true_target
791	56	1	4.2	137	1	96.7	1
833	77	0	3.9	147	0	96.7	1
677	48	0	4.6	154	1	96.0	1

التفسير السريري. المهمة الرابعة: المؤشرات الخمسة التي اختارتها اللجنة كلها من صميم كتب القلبية — السعة الجهدية ($\max_{hr}$)، وعلاقتها عكسية بمرض الشرايين التاجية)، والعمر، والكوليسترول، وضغط الدم أثناء الراحة، وانخفاض قطعة ST المستحث بالجهد، وهو التوقيع الكلاسيكي للإقفار على تخطيط قلب الجهد. التدقيق ناجح. المهمة الخامسة: المرضى الثلاثة المؤشر عليهم (الصفوف 791 و833 و677) يحملون انخفاضات في قطعة ST بمقدار 3.9-4.6 ملم — وهي شديدة بكل أعراف اختبار الجهد، حيث يُعد ≥ 2 ملم خطورة عالية أصلاً — واثنان منهم يبلغان عن ذبحة جهدية؛ والثلاثة جميعاً مؤكدة الإصابة بمرض الشرايين التاجية في السجل. قائمة فرز كهذه، مرفقة بطابور الحالات الصباحي، هي بالضبط الكيفية التي «يدعم» بها النموذج طبيب القلبية بدلاً من أن يحل محله. المهمة السادسة، مذكرة نموذجية (نموذج يُحتذى). نوصي بتجربة الانحدار اللوجستي استرشادياً. فقد حقق أعلى دقة اختبار (0.738) بين المرشحين الثلاثة، والأهم على نحو حاسم أن جدول معاملاته قابل للتدقيق سطرًا سطرًا من لجنة القلبية — فانخفاض قطعة ST والذبحة الجهدية والعمر ترفع الخطورة المتنبأ بها، بينما يخفضها معدل ضربات القلب الأقصى المُحرز، بما يطابق المعرفة السريرية. لا تستطيع الدقة التقاط هذه القابلية للتدقيق، ولا أن تخبرنا كم حالة شرايين تاجية حقيقية يفوتها النموذج مقابل كم إنذاراً كاذباً يطلقه — ونوعا الخطأ هذان يحملان كلفتين مختلفتين جداً في عيادة ألم الصدر. لذلك نطلب تقييماً على طريقة اليوم 9 للحساسية والنوعية قبل أي قرار نشر. وسيظهر ناتج النموذج على استمارة الإحالة نسبة مئوية، ويبقى قرار الفرز النهائي بيد الطبيب المراجع.

أسئلة النقاش الجماعي.

1. تفوقت الغابة العشوائية على الانحدار اللوجستي في التقاط إشارة الإنسولين في فوج السكري لكنها خسرت بوضوح في فوج القلب (0.653 مقابل 0.738). ما خصائص مجموعة البيانات التي قد ترجح عائلة على أخرى، وكيف تتحقق من ذلك قبل الالتزام بتجربة استرشادية؟
2. المريضة 833 في قائمة الأولوية امرأة في السابعة والسبعين بلا ذبحة جهدية، ومع ذلك يمنحها النموذج خطورة 96.7% . أي الميزات قادت هذا الحكم، وكيف تشرح هذا التأشير لطبيبها المعالج في جملتين؟
3. أفاد نموذج خط اليد لكشف الزهايمر بدقة 94.49% . بعد جلسة اليوم، اذكر ثلاثة أسئلة ينبغي للمديرية طرحها قبل تصديق أن هذا الرقم ينطبق على برنامج تحريّ عراقي.

8.4 الواجب المنزلي والتقييم الذاتي

التمارين.

1. فضول العتبة (تمهيد لليوم 9). على مجموعة اختبار السكري، ابن جدول الخطورة الكامل للمرضى الـ 192 من الخطوة 3، واستخدم قناعاً منطقياً (Boolean Mask) لعدّ كم من المرضى «السكريين حقاً» تقع $risk_pct$ لديهم بين 20 و50 — حالات تفوّتها العتبة الافتراضية لكن عتبة تحرّ قد تلتقطها.
 2. حجم الغابة. أعد تدريب غابة السكري العشوائية بقيم $n_estimators$ مساوية لـ 10 و50 و500 (ثلاث ملاءمات منفصلة، بلا حلقات — ثلاث خلايا معنونة) وجدول دقات الاختبار الثلاث. هل اللجنة الأكبر تفيد دائماً؟
 3. تدقيق معاملات نموذج القلب. أنتج جدول المعاملات المرتب لنموذج التحدي اللوجستي، واكتب جملة واحدة لكل ميزة من الميزات الثلاث الأولى تقارن فيها ترتيبها مع ترتيب أهميات الغابة العشوائية.
 4. فحص سلامة الاحتمالات. على مجموعة اختبار القلب، تحقق عبر $\log_model.classes_$ ومطبوعة $predict_proba$ بعمودها لثلاثة مرضى من أن العمودين يجمعان إلى 1.0 — ثم اكتب جملة واحدة عن سبب وجوب أن يصرح أي تقرير فرز بالعمود الذي استخدم.
- اختبار التقييم الذاتي.**

1. أخرج نموذج لوجستي القيمة 0.82 لمرضى. حدد بدقة ما الذي يدعيه هذا الرقم، وما التصنيف الذي ستعيده $predict$ افتراضياً.
2. في جدول معاملات على ميزات معيّنة قياسياً، وزن الميزة A هو $+1.1$ ووزن الميزة B هو -0.5 . فسّر كلاً منهما في جملة واحدة.
3. ما مصدرا العشوائية اللذان يجعلان كل شجرة في الغابة العشوائية مختلفة عن جارائها؟
4. لماذا تحتاج آلة المتجهات الداعمة إلى ميزات معيّنة بإلحاح أشد من حاجة الغابة العشوائية؟
5. يبلغ زميل عن دقة 91% لنموذج إنتان في ردهة نسبة الإنتان فيها 8% من المرضى. لماذا يكون هذا الرقم وحده غير مفيد؟

إجابات الاختبار. (1) يقدر النموذج احتمالاً قدره 82% لانتماء المريض إلى فئة المرض في ضوء الميزات؛ وبما أن $0.82 \geq 0.5$ فإن $predict$ تعيد 1. (2) زيادة قدرها انحراف معياري واحد في A ترفع لوغاريتم أرجحية المرض المتنبأ به بمقدار 1.1 (عامل خطورة اتجاهي)؛ والزيادة نفسها في B تخفض لوغاريتم الأرجحية بمقدار 0.5. بعد التعديل على بقية الميزات. (3) كل شجرة تتدرب على عينة تمهيد مختلفة من المرضى وتنظر في مجموعة جزئية عشوائية من الميزات عند كل فرع. (4) هامش آلة المتجهات الداعمة مسافة إقليدية، فتطغى الميزات ذات المقاييس الخام الكبيرة على الهندسة؛ أما تفرعات الغابة فمبنية

على عتبات لكل ميزة على حدة وهي غير متأثرة بالمقياس. (5) التنبؤ بعدم وجود إنتان للجميع يحرز أصلاً دقة 92% عند انتشار 8%. ولذلك فإن 91% متوافقة مع نموذج لا يكشف أي إنتان على الإطلاق؛ المطلوب هو الحساسية (Sensitivity) والنوعية (Specificity) — وهما موضوع اليوم 9.

❖ القيمة لمديرية التشخيص — من السجلات الاسترجاعية إلى فرز استباقي

تكسب المديرية اليوم أول نموذج أولي متكامل لدعم القرار من البداية إلى النهاية: مسار عمل يحوّل ملف CSV لإحالات روتينية إلى قائمة مراجعة ذات أولوية مرتبة مع جدول عوامل خطورة قابل للتدقيق — بُني بالكامل بأدوات مجانية مفتوحة المصدر وبأيدي ملاكات داخلية. وهذا يغيّر حوار المشتريات: فبدلاً من شراء صناديق سوداء تجارية مبهمة، تستطيع الوزارة أن تجرب استرشادياً نماذج شفافة يمكن للجائها السريرية أن تفحص معاملاتها وتطعن فيها وتقرّها، ويمكن لمحلليها أن يتحققوا من أدائها على مرضى عراقيين قبل أن يُبنى عليها أي قرار سريري.

خلاصات اليوم

- يحوّل المصنّف ميزات المريض إلى فئة تشخيصية؛ والاحتمال الكامن خلف التصنيف هو الناتج القيم سريرياً.
 - يضغط الانحدار اللوجستي مجموعاً موزوناً للميزات عبر الدالة السينية إلى خطورة بين 0 و1؛ ومعاملاته عوامل خطورة اتجاهية قابلة للتدقيق — وهذا سبب بقائه النموذج المرجعي في الطب.
 - عتبة القرار 0.5 افتراض لا قانون سريري؛ والمریضة السكرية الفائتة ذات الخطورة 17.5% بينت لماذا يجب ضبطها على كلفة الخطأ.
 - الغابة العشوائية لجنة من أشجار مدربة على عينات تمهيد تصوّت بالأغلبية: أقوى على الإشارات اللاحظية (فقد أظهرت الإنسولين) وأضعف في الشفافية؛ وأهمياتها ترتب الاستخدام لا الاتجاه.
 - تبحث آلة المتجهات الداعمة عن أوسع هامش أمان بين الفئتين؛ وهي تتطلب ميزات معايرة ولا تعطي احتمالات إلا عبر المعايرة.
 - على فوجينا كان النموذج الأكثر شفافية هو الأدق أيضاً — فالمرونة ليست أداءً تلقائياً.
 - تخلط الدقة الحالات الفائتة والإنذارات الكاذبة في رقم واحد وتكافئ تجاهل الفئة الأقلية: إنها أقل المقاييس السريرية إفادةً.
 - النموذج يدعم والطبيب يقرر: كل ناتج من مسار عمل اليوم معطى إضافي واحد على لوحة المريض، يملكه طبيب.
- غداً، في الفصل 9، تُخضع هذه النماذج للفحص الذي تستحقه: الحساسية، والنوعية، والقيم التنبؤية، ومنحنيات ROC، وعتبات مضبوطة على الكلفة الحقيقية للتشخيص الفائت.

باب 9

اليوم 9 — التقييم النقدي للأداء التشخيصي

بالأمس، في الفصل 8، درّينا أولى مصتفاتنا التشخيصية واحتفلنا بدقة (Accuracy) بلغت نحو 80 بالمئة. أما اليوم فنطرح السؤال الذي يفصل هاوي البيانات عن الممارس التشخيصي: 80 بالمئة من ماذا بالضبط؟ فالمختبر لا يُطلق فحصاً جديداً لمجرد أنه بدا جيداً؛ بل يطالب بالحساسية (Sensitivity) والنوعية (Specificity) والقيم التنبؤية، مقيسةً على مجموعة سكانية محددة، قبل أن تُصدّر نتيجة مريض واحد. والنموذج التنبؤي لا يستحق أقل من ذلك، لأن النموذج هو اختبار تشخيصي — كاشفه البيانات. هذه قمة الدورة. فكل مهارة من الأيام الثمانية الماضية — تحميل السجلات، وتنظيف القيم، وفلترة الأفواج، وهندسة الميزات (Feature Engineering)، والتقسيم والمعايرة، وتدريب المصنّفات — تلقتي عند سؤال اليوم الواحد: هل يمكن ائتمان هذا النموذج على المرضى، وعند أي عتبة؟ بنهاية الجلسة ستقرأ مصفوفة الارتباك (Confusion Matrix) كما تقرأ صبغة غرام، وتحسب كل مقياس رئيس بحساب يدوي بسيط بدل الوثوق الأعمى بمكتبة، وتدافع عن عتبة تشغيل أمام مدير برنامج بلغة الحالات المفوّنة والإنذارات الكاذبة لكل ألف شخص مفحوص.

الدقائق	الجزء	المحتوى والأنشطة
00-05	مراجعة اليوم الثامن	إعادة تحميل نموذج السكري اللوجستي وبديل الغابة العشوائية من الفصل 8؛ إعادة التأكيد على الانضباط التزيه بين مجموعة التدريب ومجموعة الاختبار: كل رقم اليوم يأتي من مجموعة الاختبار غير الممسوسة.
05-12	النظرية الأولى: فخ الدقة	لماذا قد تعني دقة 98 بالمئة اختباراً عديم الفائدة: مصفوفة الارتباك بلغة سريرية؛ السالب الكاذب بوصفه الخطأ الأخطر في الطب.
12-20	النظرية الثانية: عدة المقاييس	الحساسية والنوعية وPPV وNPV؛ معدل الانتشار وانحياز PPV؛ نسب الأرجحية؛ منحني ROC والمساحة AUC واختيار نقطة التشغيل؛ مأزق AUC للمنتج الثنائي.
20-35	التطبيق الموجّه الأول	إعادة بناء نموذج اليوم الثامن: confusion_matrix و ravel()؛ حساب المقاييس السريرية الأربعة بحساب يدوي بسيط؛ التحقق المتقابل مع classification_report.
35-50	التطبيق الموجّه الثاني	auc و roc_curve؛ رسم منحني ROC مع قطر رمية العملة؛ تحريك العتبة إلى 0.30 ومراقبة ارتفاع الحساسية مع هبوط النوعية.
50-60	التطبيق الموجّه الثالث	مسح موجّه للعتبات 0.90-0.10؛ تراكب منحنى ROC للوجستي والغابة العشوائية؛ إيجاد نقطة تشغيل يودن باستخدام np.argmax.
60-85	التحدي السريري	ملف التقييم التشخيصي الكامل لنموذجي heart_risk.csv التجريبيين؛ المصفوفات، وجدول المقاييس، وتراكب ROC، وعتبة المسح، والحالات المفوّنة لكل 1000، والمذكّرة إلى المدير العام.
85-90	الاستخلاص والواجب المنزلي	تقرأ الفرق مذكراتها بصوت عالٍ؛ الخلاصات؛ شرح الواجب المنزلي والجسر إلى اليوم العاشر.

✓ الأهداف التعليمية

بنهاية هذه الجلسة ستكون قادراً على:

1. أن تشرح، بمثال رقي، لماذا تكون الدقة الكلية مضللة في الأمراض منخفضة الانتشار، وأن تسمي المقاييس التي تحل محلها.

2. أن تبني مصفوفة الارتباك لمصنّف مدرب وتحسب الحساسية والنوعية والقيمة التنبؤية الإيجابية والقيمة التنبؤية السلبية بحساب يدوي بسيط — بلا صندوق أسود.
3. أن تبين حسابياً كيف تتغير PPV وNPV مع معدل الانتشار بينما لا تتغير الحساسية والنوعية.
4. أن تولّد منحى ROC وترسمه وتفسّره مع مساحته، AUC وتقرّن نموذجين مرشحين على المحاور نفسها.
5. أن تحرك عتبة القرار عمداً، وتقاس كمياً المقايضة الناتجة بين الحساسية والنوعية، وتبرر نقطة تشغيل (يودن افتراضياً أو بحسب المهمة) لغرض سريري معلن.
6. أن تكشف مأزق AUC المستوفاة خطياً للمنبتات الثنائية الصّرفة، وتدقق أي مقياس مولّد آلياً قبل وصوله إلى صانع القرار.
7. أن تجمع ملف تقييم كاملاً — مصفوفات، وجدول مقاييس، وتراكب، ROC، وتوصية عتبة، والحالات المفوتة المتوقعة لكل 1000 — وتلخصه في مذكرة من خمس جمل.

➤ حقبة أدوات الذكاء الاصطناعي والاعتماديات — حزمة التقييم

استيرادات اليوم — وكلها مثبتة مسبقاً على حواسيب الدورة:

```
1 import pandas as pd
2 import numpy as np
3 import matplotlib.pyplot as plt
4 from sklearn.model_selection import train_test_split
5 from sklearn.preprocessing import StandardScaler
6 from sklearn.linear_model import LogisticRegression
7 from sklearn.ensemble import RandomForestClassifier
8 from sklearn.metrics import (confusion_matrix, classification_report,
9                               roc_curve, auc)
```

```
pip install pandas numpy scikit-learn matplotlib
```

pandas تحمل جداول المرضى وتتيح تجميع جداول مقارنة مرتبة للمقاييس يمكن لصقها مباشرة في تقرير للمديرية. **numpy** توفر الحساب الموجه — الشبكات المنطقية و **argmax** — الذي يستبدل كل حلقة تكرارية عندما نمسح العتبات. **scikit-learn** تقدم **confusion_matrix** و **roc_curve** و **auc**؛ نستخدمها لتوليد الأعداد الخام والمنحنيات، ثم نتحقق بأنفسنا من كل رقم رئيس، وهذه بالضبط وضعية التدقيق التي يطبقها مختبر وطني على أي جهاز مؤرّد. **matplotlib** تحوّل مقارنة ROC إلى الشكل الوحيد الذي سنتنظر إليه لجنة المشتريات فعلاً.

9.1 السياق السريري والإطار النظري (20 دقيقة)

فخ الدقة

تخيّل أن المديرية تجرب نموذج مسح لمرض معدل انتشاره (Prevalence) 2 بالمئة — وهو تقريباً نطاق السل الرئوي النشط في قضاء عالي العبء، أو نطاق ورم خبيث غير شائع في مجموعة مسح عامة. نموذج كسول يعلن أن كل شخص بلا استثناء سليم يكون مصيباً في 9,800 شخص سليم من كل 10,000 ومخطئاً فقط في المرضى المئتين. دقته 98 بالمئة. وقيمتة السريرية صفر تماماً: لا يكتشف أحداً، ولا يعالج أحداً، ولا ينقذ أحداً. وأي لوحة معلومات تنصدها الدقة على مجموعة غير متوازنة هي دعاية لا تقييم.

فخ الدقة

الدقة $(TP + TN) / N$ = تكافئ التوافق مع الفئة الأغلبية. عندما يكون انتشار المرض منخفضاً تكون الفئة الأغلبية هي الأصحاء، فيستطيع نموذج أن يسجل دقة عالية وهو يفوت كل مريض حقيقي. عند انتشار 2 بالمئة يسجل نموذج «الجميع سلمي» دقة 0.98 بحساسية 0.00. قاعدة المديرية: لا تُذكر الدقة وحدها أبداً؛ كل تقرير عن نموذج يجب أن يذكر الحساسية والنوعية وPPV وNPV ومعدل انتشار مجموعة التقييم.

مصفوفة الارتباك بلغة سريرية

كل قرار تشخيصي ثنائي — فحص مختبري أو قراءة تصويرية أو نموذج — يقع في خلية واحدة بالضبط من أربع خلايا عند مقارنته بالحقبة. يعرض الشكل 9.1 المصفوفة كما تطبعها scikit-learn: الصفوف هي الحقيقة، والأعمدة هي قرار النموذج، والفئة السلبية (الأصحاء) تأتي أولاً.

	Model says: HEALTHY	Model says: DISEASED	
Truth: HEALTHY	True Negative (TN) health correctly confirmed <i>reassurance, resources saved</i>	False Positive (FP) false alarm <i>anxiety, cost, extra testing</i>	Specificity $= \frac{TN}{TN + FP}$ (this row)
Truth: DISEASED	False Negative (FN) missed diagnosis <i>the most dangerous error</i>	True Positive (TP) disease correctly detected <i>timely treatment</i>	Sensitivity $= \frac{TP}{TP + FN}$ (this row)
	$NPV = \frac{TN}{TN + FN}$ (this column)	$PPV = \frac{TP}{TP + FP}$ (this column)	

شكل 9.1: مصفوفة الارتباك بلغة سريرية. تتبع الصفوف اصطلاح scikit-learn (الحقيقة: الأصحاء أولاً). الخلايا الخضراء قرارات صحيحة؛ والخلية الكهرمانية هي الإنذار الكاذب؛ أما الخلية الحمراء — التشخيص المفوت — فهي الخطأ الأقل مغفرة في الطب.

خلينا الخطأ ليستا متساويتين أخلاقياً. فالإيجابي الكاذب يكلف فحصاً تأكيدياً ومريضاً قلقاً ووقتاً في العيادة — تكاليف حقيقية لكنها قابلة للاسترداد. أما السالب الكاذب فهو مريض أعيد إلى بيته حاملاً المرض الذي كنا نبحت عنه: حالة السل التي تواصل نقل العدوى، ونقص التروية القلبية الذي خرج من قسم الطوارئ، والورم الخبيث الذي قيل لصاحبه عُذ في العام المقبل. في كل سياق مسح تقريباً، السالب الكاذب هو الخطأ الذي لا نستطيع وزارة تحمله، وكل قرار عتبة اليوم سيُحاجج فيه بلغة هاتين الخليتين.

المقاييس السريرية الأربعة

من الخلايا الأربع تأتي الأسئلة الأربعة التي يطرحها الطبيب فعلاً:

• **الحساسية** $TP/(TP + FN)$ — من بين المرضى الحقيقيين جميعاً، ما النسبة التي التقطناها؟ (وتسمى أيضاً الاستدعاء Recall أو معدل الإيجابية الحقيقية).

• **النوعية** $TN/(TN + FP)$ — من بين الأصحاء جميعاً، ما النسبة التي برأناها بحق؟

• **القيمة التنبؤية الإيجابية (PPV)** $TP/(TP + FP)$ — المريض الذي أمامي نتيجته إيجابية: ما احتمال أن يكون المرض موجوداً فعلاً؟

• **القيمة التنبؤية السلبية (NPV)** $TN/(TN + FN)$ — نتيجة المريض سلبية: ما مدى الثقة بإعلان الخلو من المرض؟

الحساسية والنوعية خاصيتان للاختبار وترتعلان معه: أما PPV و NPV فخاصيتان للاختبار في مجموعة سكانية وتتغيران تغيراً هائلاً مع معدل الانتشار.

جدول 9.1: عدة المقاييس في لمحة واحدة.

المقياس	الصيغة	السؤال السريري
الحساسية	$TP/(TP + FN)$	من بين المرضى، كم التقطنا؟
النوعية	$TN/(TN + FP)$	من بين الأصحاء، كم برأنا؟
القيمة التنبؤية الإيجابية	$TP/(TP + FP)$	نتيجة إيجابية — هل نصدقها؟
القيمة التنبؤية السلبية	$TN/(TN + FN)$	نتيجة سلبية — هل نصدقها؟
الدقة	$(TP + TN)/N$	التوافق الكلي (احذر معدل الانتشار)
LR+	$Sens/(1 - Spec)$	كم تُزج النتيجة الإيجابية الأرجحية
AUC	المساحة تحت منحنى ROC	مهارة الترتيب عبر كل العتبات

معدل الانتشار: المتغير الصامت خلف كل قيمة تنبؤية

إليك الحساب الذي ينبغي لكل ضابط في المديرية أن يستطيع إجراءه على ظهر ورقة. خذ اختباراً واحداً ثابتاً: حساسية 90 بالمئة ونوعية 90 بالمئة. انشره في بيئتين وتنبّع 10,000 شخص في كل منهما. الجدول 9.2 ينجز المحاسبة الكاملة.

جدول 9.2: الاختبار نفسه (حساسية 0.90 ونوعية 0.90) في مجموعتين سكانييتين من 10,000 شخص. لا تتحرك الحساسية ولا النوعية: بينما تهاجر PPV من 69 بالمئة إلى 16 بالمئة عند انخفاض معدل الانتشار.

المسح السكاني (انتشار 2%)	عيادة الإحالة (انتشار 20%)	
200 المصابون فعلاً	2,000	
9,800 الأصحاء فعلاً	8,000	
$TP = 0.90 \times$ المصابين	1,800	
FN (الحالات المفقودة)	200	
$TN = 0.90 \times$ الأصحاء	7,200	
FP (الإنذارات الكاذبة)	800	
$PPV = TP/(TP + FP)$	$1800/2600 = 69.2\%$	$180/1160 = 15.5\%$
$NPV = TN/(TN + FN)$	$7200/7400 = 97.3\%$	$8820/8840 = 99.8\%$

الانتشار والقيم التنبؤية

في مجموعة المسح، النتيجة الإيجابية خاطئة في 84 مرة من كل 100 — لا لأن الاختبار تدهور، بل لأن المرضى الحقيقيين نادرون والإنذارات الكاذبة تتراكم من الأغلبية السليمة الضخمة. نسبة الأرجحية الإيجابية تجعل الآلية صريحة: $LR+ = Sens/(1 - Spec) = 0.90/0.10 = 9$. أرجحية ما قبل الاختبار عند انتشار 2% هي 0.0204؛ اضربها في 9 تحصل على أرجحية ما بعد الاختبار 0.184، أي احتمال ما بعد الاختبار 15.5% — وهي بالضبط PPV في الجدول 9.2. النتائج العملية: (1) نموذج جرى التحقق منه في مستشفى إحالة سيُظهر PPV أدنى بكثير عند نقله إلى المسح المجتمعي؛ (2) كل PPV أو NPV يُذكر بلا معدل الانتشار الذي حُسب عنده كلام بلا معنى؛ (3) برامج المسح يجب أن تخطط دائماً لمسار التأكيد للإنذارات الكاذبة المحتملة.

العتبة ومنحنى ROC والمساحة تحت المنحنى

النموذج اللوجستي لا يُخرج مريضاً أو سليماً؛ بل يُخرج احتمالاً، ونحن من يختار القيمة الفاصلة التي تحوِّله إلى قرار — عتبة القرار (Threshold). اخفض الفاصل تتسع الشبكة: مرضى حقيقيون أكثر يلتقطون (ترتفع الحساسية) لكن أصحاء أكثر يُعلمون (تنخفض النوعية). ارفعه يحدث العكس. فالحساسية والنوعية إذن ليستا صفتين مستقلتين للنموذج — إنهما قرص واحد بوجهين. منحنى خاصية تشغيل المستقبل (ROC) يجعل القرص كله مرئياً دفعة واحدة: امسح العتبة من 1 نزولاً إلى 0 وارسم معدل الإيجابية الحقيقية (الحساسية) مقابل معدل الإيجابية الكاذبة (واحد ناقص النوعية) عند كل إعداد. الاختبار المثالي يمر بالركن الأعلى الأيسر (0, 1) — كل المرضى مُلتقطون ولا أحد من الأصحاء مُعَلَّم. أما القطر فهو رمية عملة: أي منحنى يلتصق به لا يحمل أي معلومة تشخيصية. والمساحة تحت المنحنى (AUC) تضغط المنحنى كله في رقم واحد له تفسير جميل: احتمال أن يرتب النموذج مريضاً مختاراً عشوائياً أعلى من شخص سليم مختار عشوائياً.

جدول 9.3: نطاقات تفسير عملية للمساحة تحت المنحنى AUC السياق هو الحكم: قد تكون $AUC = 0.75$ ممتازة لسؤال مسح يكلف دولاراً واحداً وغير مقبولة لقرار إجراء عملية جراحية.

AUC	التفسير العملي
0.50	لا تمييز — رمية عملة
0.50–0.70	ضعيف: نادراً ما يفيد سريراً وحده
0.70–0.80	مقبول: مفيد كأداة فرز مساعدة
0.80–0.90	ممتاز: مرشح لدعم القرار السريري
> 0.90	استثنائي: دقق بحثاً عن تسرب البيانات (Data Leakage) قبل الاحتفال

اختيار نقطة التشغيل

يعرض منحنى ROC كل مقايضة ممكنة؛ والمهمة هي التي تختار النقطة. المسح الوطني (العثور على كل حالة تقريباً؛ ويوجد اختبار تأكيد رخيص) → قدّم الحساسية: اقبل مزيداً من الإنذارات الكاذبة لإبقاء FN قرب الصفر. التأكيد قبل علاج باضع أو سام (الإيجابي الكاذب يستدعي جراحة أو علاجاً كيميائياً) → قدّم النوعية. وعندما لا تهيمن حجة كلفة، يكون مؤشر يودن (Youden) $J = \text{Sens} + \text{Spec} - 1$ افتراضاً يمكن الدفاع عنه: اختر العتبة التي تعظم J ، وهي نقطة المنحنى الأبعد فوق القطر. يتراوح J من 0 (عديم الفائدة) إلى 1 (مثالي). إنه عرض افتتاحي يُقدّم للجنة السريرية، لا بديل عنها.

ملاحظة نقدية متقدمة: مأزق AUC للمنبئ الثنائي

يحتاج منحنى ROC إلى ترتيب — درجة متصلة تُمسح. المنبئ الثنائي الصّرف (عدة المستضد السريع: إيجابي أو سلبي) لا قرص له: كل معلوماته على ROC نقطة واحدة (Sens, $1 - \text{Spec}$). برمجيات مثل scikit-learn سترسم مع ذلك، وبكل سرور، منحنى بوصل (0, 0) وتلك النقطة الوحيدة و(1, 1) بخطوط مستقيمة، فتتجاهل المساحة شبه المنحرفة إلى $AUC = (\text{Sens} + \text{Spec})/2$ — وهو مأزق موثق يمكن أن يبخر القيمة التمييزية التي يحملها الاختبار الثنائي فعلاً (انظر مقالة PubMed Central 7695228). عدة بحساسية 0.85 ونوعية 0.97 تطبع AUC مستوفاة قدرها 0.91؛ ومن يرتّب هذا الرقم مقابل $AUC = 0.93$ لمنحنى أملس لنموذج متصل ثم يعلن تفوق النموذج يرتكب خطأ تصنيفياً. والدرس المؤسسي أكبر من الصيغة: الانحياز للأتمتة (Automation Bias) — قبول رقم محسوب لمجرد أن حاسوباً أنتجه — خطأ تشخيصي، تماماً كالتسّخ عند قراءة أشعة أولية. لا تقراً AUC من تقرير على عماها أبداً؛ اعرف كيف حُسبت، وأعد حساب المقاييس الأولية بنفسك. تطبيق اليوم الموجّه كله هو هذه العادة، مدربةً بالتكرار.

★ لؤلؤة سريرية — كل قيمة فاصلة مختبرية هي نقطة تشغيل

منطق العتبة الذي تتعلمه اليوم يحكم طاولة المختبر أصلاً. سكر البلازما الصائم ≥ 126 ملغ/دل يعرف السكري؛ والسكر التراكمي HbA1c عند 6.5% فأكثر سكريٌّ وعند 5.7–6.4% ما قبل السكري؛ وكلا الفاصلين اختير حيث يتصاعد خطر اعتلال الشبكية — تسوية بين الحساسية والنوعية تَبَّتْها لجنة، لا الطبيعة. والتروبونين عالي الحساسية يستخدم المئين التاسع والتسعين لمجموعة مرجعية سليمة تحديداً لرفع الحساسية في قسم الطوارئ، مع قبول مزيد من الإنذارات الكاذبة تصفّحها عينة ثانية موقوتة. وفي RT-PCR، قيمة عتبة الدورة (Ct) هي نفسها عتبة على عبور التآلق: قيمة Ct المنخفضة (أقل من 25 تقريباً) تدل على حمل فيروسي مرتفع، بينما العبور المتأخر جداً (Ct 35–40) يقع قرب حد كشف الفحص، وكثير من المختبرات تعيد الاختبار قبل الإبلاغ عن إيجابي ضعيف. إن تحريك عتبة نموذج من 0.50 إلى 0.30 هو الفعل نفسه الذي يقوم به مختبر يعيد اعتماد قيمة فاصلة مرجعية — ويستحق التوثيق نفسه.

9.2 التطبيق الموجّه (40 دقيقة)

■ عرض تطبيقي مباشر — تدقيق نموذج سكري اليوم الثامن كما يُدقق فحص مختبري

الهدف: أخذ انحدار الأمس اللوجستي (Logistic Regression) على diabetes_dataset.csv وإخضاعه لتقييم طريقة كامل، تماماً كما يقيم المختبر جهاز تحليل جديداً: مصفوفة الارتباك، والمقاييس السريرية الأربعة محسوبة يدوياً، ROC وAUC، وتحريك متعمد للعتبة، ومسح عتبات موجّه، ومواجهة مباشرة مع الغابة العشوائية (Random Forest)، وأخيراً نقطة تشغيل

يودن. يكتب المتدربون كل سطر؛ ويروي المدرب المعنى السريري لكل رقم.

الخطوة 1: إعادة بناء نموذج اليوم الثامن واستخراج الاحتمالات

نعيد إنتاج مسار عمل الأمس بالقيمة نفسها لـ `random_state` كي ترى القاعة كلها أرقاماً متطابقة. الجديد الجوهرى: نحتفظ بالاحتمالات من `predict_proba`، لا بالتسميات الصلبة فحسب — فعمود الاحتمال هو الإشارة الخام، والتسمية ليست سوى تلك الإشارة مدفوعة عبر عتبة. لاحظ أن `predict()` مثبّنة على فاصل 0.50؛ وكتابة العتبة صراحةً تجعل القرار السريري مرئياً في الشيفرة.

```
1 import pandas as pd
2 import numpy as np
3 from sklearn.model_selection import train_test_split
4 from sklearn.preprocessing import StandardScaler
5 from sklearn.linear_model import LogisticRegression
6
7 pd.set_option("display.width", 100)
8
9 df = pd.read_csv("diabetes_dataset.csv")
10 X = df.drop(columns="outcome")
11 y = df["outcome"]
12
13 X_train, X_test, y_train, y_test = train_test_split(
14     X, y, test_size=0.25, random_state=42, stratify=y)
15
16 scaler = StandardScaler()
17 X_train_s = scaler.fit_transform(X_train)
18 X_test_s = scaler.transform(X_test)
19
20 logit = LogisticRegression(max_iter=1000)
21 logit.fit(X_train_s, y_train)
22
23 proba = logit.predict_proba(X_test_s)[: , 1]
24 y_pred = (proba >= 0.50).astype(int)
25
26 print("Test patients          :", len(y_test))
27 print("True diabetics in test :", int(y_test.sum()))
28 print("First 5 probabilities  :", np.round(proba[:5], 3))
```

شيفرة 9.1: إعادة بناء نموذج السكري والاحتفاظ بإشارة الاحتمال.

```
Test patients          : 192
True diabetics in test : 71
First 5 probabilities  : [0.295 0.09  0.968 0.219 0.175]
```

تضم مجموعة الاختبار 192 مريضاً لم يَرهم التدريب قط، منهم 71 مصاباً فعلاً بالسكري — أي معدل انتشار 37 بالمئة، وهو نموذجي لمجموعة إحالة وأعلى بكثير من أي بيئة مسح مجتمعي. أبقى هذا الرقم في ذهنك عندما نناقش PPV.

الخطوة 2: مصفوفة الارتباك مفككةً خلية خلية

```
1 from sklearn.metrics import confusion_matrix
2
3 cm = confusion_matrix(y_test, y_pred)
4 print("Confusion matrix (rows = truth, cols = prediction):")
5 print(cm)
6
```

```

7 tn, fp, fn, tp = cm.ravel()
8 print("TN =", tn, "| FP =", fp, "| FN =", fn, "| TP =", tp)

```

شيفرة 9.2: الخلايا الأربع، مع `ravel()` التي تبسطها بالترتيب الثابت، TP، FN، FP، TN.

```

Confusion matrix (rows = truth, cols = prediction):
[[106  15]
 [ 23  48]]
TN = 106 | FP = 15 | FN = 23 | TP = 48

```

اقرأها في ضوء الشكل 9.1: 106 أشخاص أصحاء بُرِّئوا بحق، و15 إنذاراً كاذباً، و48 مريض سكري نُقِلوا — و23 مريض سكري أُعيدوا إلى بيوتهم بشهادة سلامة. هؤلاء الثلاثة والعشرون هم الرقم الذي ينبغي للقاعة أن تحدِّق فيه. فقرابة مريض سكري واحد من كل ثلاثة يخرج دون كشف عند العتبة الافتراضية.

✖ خطأ شائع — تبديل الوسيطين يبادل أخطاءك بصمت

التوقع هو `confusion_matrix(y_true, y_pred)` — الحقيقة أولاً. تبديل الوسيطين لا يثير أي خطأ؛ بل ينقل المصفوفة إلى منقولها، فيتبادل FN وFP مكانهما بصمت، ويمكن لتقرير أن يبخل التشخيصات المفوتة بينما يضخم الإنذارات الكاذبة. وبالمثل، ترتيب التفكيك `tn, fp, fn, tp = cm.ravel()` صحيح فقط للتسميات الثنائية المرتبة [0, 1] حيث يعني الصفر السليم. تحقق مرة واحدة في كل مشروع بفحص أن `tn + fp + fn + tp == len(y_test)` وأن `tp + fn` يساوي العدد المعروف للمرضى الحقيقيين (هنا $48 + 23 = 71$ ، صحيح).

الخطوة 3: المقاييس السريرية الأربعة بحساب يدوي بسيط

لا صندوق أسود. أربع عمليات قسمة يستطيع عضو اللجنة التحقق منها يدوياً — وينبغي له.

```

1 sensitivity = tp / (tp + fn)
2 specificity = tn / (tn + fp)
3 ppv = tp / (tp + fp)
4 npv = tn / (tn + fn)
5 accuracy = (tp + tn) / (tp + tn + fp + fn)
6
7 print(f"Sensitivity : {sensitivity:.3f} ({tp} of {tp+fn} diabetics caught)")
8 print(f"Specificity : {specificity:.3f} ({tn} of {tn+fp} healthy cleared)")
9 print(f"PPV : {ppv:.3f} (a positive call is right {ppv:.1%})")
10 print(f"NPV : {npv:.3f} (a negative call is right {npv:.1%})")
11 print(f"Accuracy : {accuracy:.3f}")

```

شيفرة 9.3: الحساسية والنوعية و PPV و NPV والدقة من المبادئ الأولى.

```

Sensitivity : 0.676 (48 of 71 diabetics caught)
Specificity : 0.876 (106 of 121 healthy cleared)
PPV : 0.762 (a positive call is right 76.2%)
NPV : 0.822 (a negative call is right 82.2%)
Accuracy : 0.802

```

النموذج نفسه الذي يبلغ عن دقة محترمة قدرها 80 بالمئة لا يلتقط سوى 68 بالمئة من مرضى السكري. هذا فخ الدقة مصغراً: المقياس الذي يجمال ليس المقياس المهم. ولاحظ أيضاً أن PPV البالغة 76 بالمئة إنما هي عند انتشار 37 بالمئة — انشر هذا النموذج دون تغيير في مسح مجتمعي انتشاره 5 بالمئة، وسيتنبأ حساب بايز في الجدول 9.2 بهبوط حاد في PPV.

الخطوة 4: التحقق المتقابل مع `classification_report`

نحسب يدوياً ونتحقق بالمكتبة — في الاتجاهين. في مفردات `scikit-learn`، استدعاء فئة السكري هو `classification_report`، واستدعاء فئة الأصحاء هو نوعيتنا، ودقة التنبؤ (*Precision*) لفئة السكري هي قيمتنا التنبؤية الإيجابية.

```

1 from sklearn.metrics import classification_report
2
3 print(classification_report(y_test, y_pred,
4                             target_names=["healthy", "diabetic"]))

```

شيفرة 9.4: منظور المكتبة — تأكيداً للحساب اليدوي لا بديل عنه.

	precision	recall	f1-score	support
healthy	0.82	0.88	0.85	121
diabetic	0.76	0.68	0.72	71
accuracy			0.80	192
macro avg	0.79	0.78	0.78	192
weighted avg	0.80	0.80	0.80	192

كل رقم يطابق الخطوة 3: الاستدعاء 0.68 في صف السكري هو حساسيتنا، والاستدعاء 0.88 في صف الأصحاء هو نوعيتنا، ودقة التنبؤ 0.76 في صف السكري هي PPV عندنا. عادة التدقيق تكلف ثلاثين ثانية وتشتري ثقة مؤسسية.

الخطوة 5: منحنى ROC والمساحة تحت المنحنى

تنجز roc_curve المسح نيابة عنا: ترتب الاحتمالات وتحسب معدلي الإيجابية الحقيقية والكاذبة عند كل فاصل متمايز.

```

1 from sklearn.metrics import roc_curve, auc
2 import matplotlib.pyplot as plt
3
4 fpr, tpr, thresholds = roc_curve(y_test, proba)
5 auc_logit = auc(fpr, tpr)
6 print("Thresholds swept by roc_curve:", len(thresholds))
7 print(f"AUC, logistic regression : {auc_logit:.3f}")
8
9 plt.figure(figsize=(5, 5))
10 plt.plot(fpr, tpr, color="teal", label=f"Logistic (AUC = {auc_logit:.3f})")
11 plt.plot([0, 1], [0, 1], "k--", label="Coin flip (AUC = 0.500)")
12 plt.xlabel("1 - Specificity (FPR)")
13 plt.ylabel("Sensitivity (TPR)")
14 plt.legend(loc="lower right")
15 plt.tight_layout()
16 plt.savefig("roc_diabetes.png", dpi=150)
17 print("Saved: roc_diabetes.png")

```

شيفرة 9.5: مسح كل العتبات دفعة واحدة وقياس المساحة.

```

Thresholds swept by roc_curve: 58
AUC, logistic regression : 0.854
Saved: roc_diabetes.png

```

AUC = 0.854: إذا أعطي النموذج مريض سكري عشوائياً وشخصاً سليماً عشوائياً من مجموعة الاختبار هذه، فإنه يمنح مريض السكري الاحتمال الأعلى في نحو 85 مرة من كل 100. وبحسب الجدول 9.3 هذا ممتاز — مرشح حقيقي لدعم القرار، رهناً بعمل العتبة أدناه.

الخطوة 6: تحريك العتبة — رافعة المسح

تسأل المديرية: هل يمكن لهذا النموذج أن يخدم برنامج مسح يكون فيه تفويت مريض سكري أسوأ من إنذار كاذب؟ نخفض الفاصل من 0.50 إلى 0.30 ونعيد حساب كل شيء. لا شيء في النموذج يتغير — قاعدة قرارنا وحدها.

```

1 y_pred_030 = (proba >= 0.30).astype(int)
2 tn3, fp3, fn3, tp3 = confusion_matrix(y_test, y_pred_030).ravel()
3 print("TN =", tn3, "| FP =", fp3, "| FN =", fn3, "| TP =", tp3)
4
5 compare = pd.DataFrame(
6     {"thr_0.50": [sensitivity, specificity, ppv, npv],
7      "thr_0.30": [tp3 / (tp3 + fn3), tn3 / (tn3 + fp3),
8                  tp3 / (tp3 + fp3), tn3 / (tn3 + fn3)]},
9     index=["Sensitivity", "Specificity", "PPV", "NPV"]).round(3)
10 print(compare)

```

شيفرة 9.6: النموذج نفسه، قاعدة قرار جديدة: العتبة 0.30.

	thr_0.50	thr_0.30
TN = 86 FP = 35 FN = 14 TP = 57		
Sensitivity	0.676	0.803
Specificity	0.876	0.711
PPV	0.762	0.620
NPV	0.822	0.860

ينخفض عدد مرضى السكري المفوتين من 23 إلى 14 — تسعة مرضى إضافيون يدخلون الرعاية. والثمن مطبوع في الجدول نفسه: ترتفع الإنذارات الكاذبة من 15 إلى 35، وتهبط PPV من 0.76 إلى 0.62. لا حساسية مجانية؛ لا يوجد سوى مقايضة نتحكم الآن بشروطها ونستطيع الدفاع عنها.

✖ خطأ شائع — لا تضبط العتبة على مجموعة الاختبار

حركنا العتبة هنا على مجموعة الاختبار لغرض الوضوح التعليمي. أما في تقييم إنتاجي فهذا تسرب خفي: اختر العتبة التي تبدو الأفضل على بيانات الاختبار تكن قد لاءمت نموذجك عليها بصمت، وسيخيب أداء الغد الظن. الانضباط الصحيح: اختر عتبة التشغيل على بيانات التدريب أو التحقق (أو على أسس الكلفة السريرية قبل رؤية أي بيانات)، ثم بلغ عن مجموعة الاختبار غير الممسوسة عند تلك العتبة المسجلة سلفاً — المنطق نفسه لتثبيت القيمة الفاصلة لفحص قبل جولة الاعتماد، لا بعدها.

الخطوة 7: مسح العتبات 0.10-0.90 بمعالجة موجهة بالكامل

مقارنة بثّ واحدة تقيّم العتبات التسع كلها دفعة واحدة: شبكة منطقية 9×192 ، ثم جموع على مستوى الأعمدة. لا حلقات تكرارية — إنها طريقة numpy في المعالجة الموجهة (Vectorization)، وهي سريعة بما يكفي لمليون مريض.

```

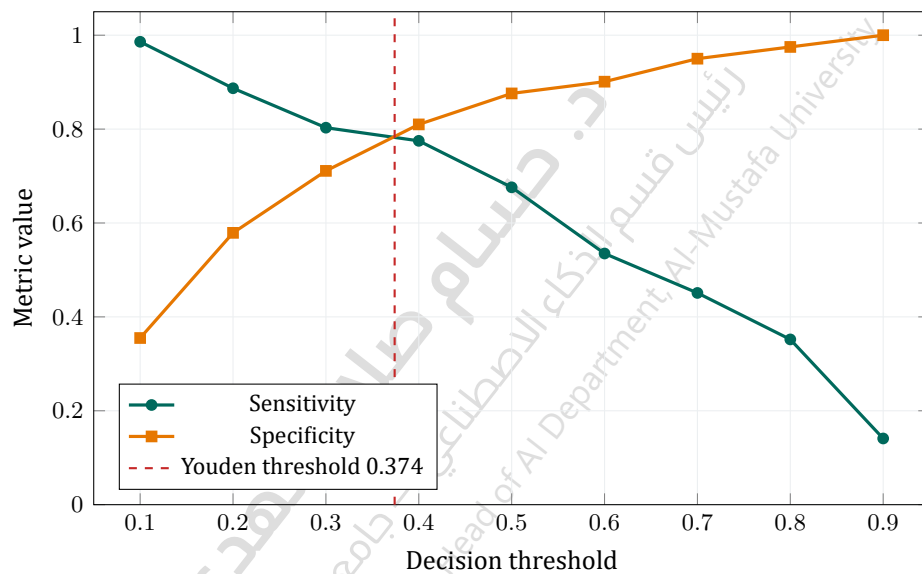
1 thr_grid = np.round(np.arange(0.10, 0.91, 0.10), 2)
2 pred_grid = (proba[:, None] >= thr_grid).astype(int)
3 truth = y_test.to_numpy()[:, None]
4
5 tp_g = ((pred_grid == 1) & (truth == 1)).sum(axis=0)
6 fn_g = ((pred_grid == 0) & (truth == 1)).sum(axis=0)
7 tn_g = ((pred_grid == 0) & (truth == 0)).sum(axis=0)
8 fp_g = ((pred_grid == 1) & (truth == 0)).sum(axis=0)
9
10 sweep = pd.DataFrame(
11     {"threshold": thr_grid,
12      "sensitivity": tp_g / (tp_g + fn_g),
13      "specificity": tn_g / (tn_g + fp_g),
14      "missed_FN": fn_g,
15      "false_alarms_FP": fp_g}).round(3)
16 print(sweep.to_string(index=False))

```

شيفرة 9.7: شبكة تنبؤات بأبعاد 9×192 عبر البث؛ كل مقياس لكل عتبة بأربع عمليات جمع.

threshold	sensitivity	specificity	missed_FN	false_alarms_FP
0.1	0.986	0.355	1	78
0.2	0.887	0.579	8	51
0.3	0.803	0.711	14	35
0.4	0.775	0.810	16	23
0.5	0.676	0.876	23	15
0.6	0.535	0.901	33	12
0.7	0.451	0.950	39	6
0.8	0.352	0.975	46	3
0.9	0.141	1.000	61	0

المقايضة كلها في جدول واحد: عند العتبة 0.10 نفوت مريض سكري واحداً لكننا نعلم 78 شخصاً سليماً؛ وعند 0.90 لا نعلم أحداً تقريباً ونفوت 61 من 71 مريضاً. الشكل 9.2 يرسم هذه القيم المقيسة بالضبط — نقطة تقاطع المنحنيين هي النظام المتوازن، وكل مهمة سريرية تختار جانبها من التقاطع.



شكل 9.2: الحساسية والنوعية المقيستان لنموذج السكري اللوجستي عند العتبات 0.10-0.90 (قيم حقيقية من الشيفرة 9.7). خفض العتبة يشتري الحساسية على حساب النوعية؛ ويحدد الخط المتقطع نقطة يودن المثلى التي نجدها في الخطوة 9.

الخطوة 8: اللوجستي مقابل الغابة العشوائية — رسم واحد ورقم واحد

نموذج مرشح ثانٍ يخضع للتدقيق نفسه، وعلى المحاور نفسها. مجموعات الأشجار لا تحتاج معايرة للميزات، لذا تتدرب الغابة على مصفوفة الميزات الخام.

```

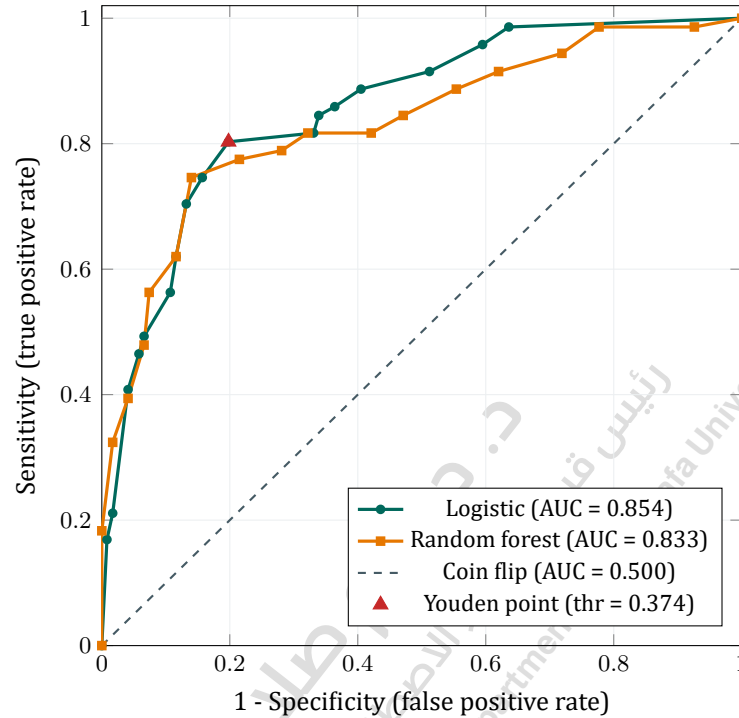
1 from sklearn.ensemble import RandomForestClassifier
2
3 forest = RandomForestClassifier(n_estimators=200, random_state=42)
4 forest.fit(X_train, y_train)
5 proba_rf = forest.predict_proba(X_test)[: , 1]
6
7 fpr_rf, tpr_rf, thr_rf = roc_curve(y_test, proba_rf)
8 auc_rf = auc(fpr_rf, tpr_rf)
9 print(f"AUC, logistic regression : {auc_logit:.3f}")
10 print(f"AUC, random forest : {auc_rf:.3f}")

```

شيفرة 9.8: غابة اليوم الثامن العشوائية تدخل حلبة ROC نفسها.

```
AUC, logistic regression : 0.854
AUC, random forest      : 0.833
```

النموذج الأبسط، القابل للتفسير بالكامل، يفوز على هذه البيانات — نتيجة تستحق التذوق. الشكل 9.3 يراكب المنحنيين المقيسين؛ والمنحنى اللوجستي يهيمن على معظم المنطقة اليسرى ذات الأهمية السريرية.



شكل 9.3: منحني ROC المقيسان لنموذجي السكري (خُففت النقاط من مخرجات roc_curve الكاملة لتسهيل القراءة). يهيمن النموذج اللوجستي في منطقة النوعية العالية حيث تعيش نقاط التشغيل السريرية؛ ويعلم المثلث الأحمر عتبهته المثلى بحسب يودن.

الخطوة 9: نقطة تشغيل يودن باستخدام np.argmax

مؤشر يودن $J = \text{Sens} + \text{Spec} - 1 = \text{TPR} - \text{FPR}$: عملية طرح واحدة على المصفوفات التي أعطينا إياها roc_curve أصلاً، و `argmax` واحدة لإيجاد العتبة الأبعد فوق القطر.

```
1 j = tpr - fpr
2 best = np.argmax(j)
3 print(f"Youden J max : {j[best]:.3f}")
4 print(f"Best threshold : {thresholds[best]:.3f}")
5 print(f"Sensitivity : {tpr[best]:.3f}")
6 print(f"Specificity : {1 - fpr[best]:.3f}")
```

شيفرة 9.9: نقطة التشغيل الافتراضية المتوازنة في ثلاثة أسطر موجهة.

```
Youden J max : 0.604
Best threshold : 0.374
Sensitivity : 0.803
Specificity : 0.802
```

عند العتبة 0.374 يلتقط النموذج 80 بالمئة من مرضى السكري بينما يبرئ 80 بالمئة من الأصحاء — التسوية المتوازنة، وهي المثلث الأحمر في الشكل 9.3. أما هل ينبغي للمديرية أن تفضل التوازن فسؤال سريري لا رياضي: فلمهمة مسح، يحتاج جدول المسح بالدفع أبعد نحو اليسار. وتلك الحاجة هي بالضبط ما يطلبه منك التحدي الآن.

9.3 التحدي السريري (30 دقيقة)

📌 التحدي السريري — ملف المسح القلبي المرفوع إلى المدير العام

تقيم المديرية نموذجين تجريبيين مبنيين على مجموعة بيانات القلبية heart_risk.csv (900 مريض محالين، وعمود هدف ثنائي target لمرض قلبي معتبر) بوصفهما أداتي فرز مرشحتين لبرنامج مسح قلبي وعائي وطني مقترح. سيقدر المدير العام في الأسبوع المقبل ما إذا كان المشروع التجريبي سيمضي. على فريقك إنتاج ملف التقييم التشخيصي الكامل. اعملوا في أزواج؛ وكل رقم يجب أن يأتي من قسم الاختبار غير المسوس.

1. حمّل heart_risk.csv؛ ورّمز sex و chest_pain_type و exercise_angina بالترميز الأحادي عبر pd.get_dummies(..., drop_first=True)؛ وأنشئ تقسيماً 75/25 محافظاً على نسب الفئات مع random_state=42؛ وعابر الميزات للنموذج اللوجستي فقط.

2. درّب النموذجين التجريبيين: الانحدار اللوجستي (max_iter=1000) على ميزات معايرة، والغابة العشوائية (random_state=42, n_estimators=200) على ميزات خام؛ واستخرج احتمالات مجموعة الاختبار من كليهما.

3. أنتج مصفوفة الارتباك لكل نموذج عند العتبة الافتراضية 0.50 وفكك TN و FP و FN و TP.

4. جمّع جدول المقاييس الكامل للنموذجين: الحساسية والنوعية و PPV و NPV والدقة و AUC — محسوبةً بحساب بسيط من الخلايا المفككة، في DataFrame واحد.

5. راكب منحني ROC مع قطر رمية العملة في رسم واحد موسوم، واذكر النموذج الذي ينبغي للملف أن يمضي به قدماً.

6. للنموذج المختار، امسح العتبات 0.10-0.90 بالبحث، وبافتراض انتشار مسح قدره 10 بالمئة أضف عمودين للمنتوقع من الحالات المفوتة لكل 1000 مفحوص والإنذارات الكاذبة لكل 1000 مفحوص؛ وأوص بعتبة تشغيل للبرنامج الوطني وبررها بلغة كلف FN مقابل FP.

7. اذكر العدد المتوقع للحالات المفوتة لكل 1000 مفحوص عند عتبتك، واختم الملف بمذكرة توصية من خمس جمل إلى المدير العام.

✓ حل المدرب — ملف المسح القلبي كاملاً

المهام 1-3: البيانات والنماذج والمصفوفات.

```
1 import pandas as pd
2 import numpy as np
3 from sklearn.model_selection import train_test_split
4 from sklearn.preprocessing import StandardScaler
5 from sklearn.linear_model import LogisticRegression
6 from sklearn.ensemble import RandomForestClassifier
7 from sklearn.metrics import confusion_matrix, roc_curve, auc
8
9 pd.set_option("display.width", 100)
10
11 heart = pd.read_csv("heart_risk.csv")
12 X = pd.get_dummies(heart.drop(columns="target"),
13                   columns=["sex", "chest_pain_type", "exercise_angina"],
14                   drop_first=True)
15 y = heart["target"]
16 X_train, X_test, y_train, y_test = train_test_split(
17     X, y, test_size=0.25, random_state=42, stratify=y)
18 scaler = StandardScaler()
19 X_train_s = scaler.fit_transform(X_train)
20 X_test_s = scaler.transform(X_test)
21 print("Encoded features:", X.shape[1],
22       "| test patients:", len(y_test),
```

```

23     "| diseased:", int(y_test.sum()))
24
25 logit = LogisticRegression(max_iter=1000).fit(X_train_s, y_train)
26 forest = RandomForestClassifier(n_estimators=200,
27                                random_state=42).fit(X_train, y_train)
28 proba_lr = logit.predict_proba(X_test_s)[: , 1]
29 proba_rf = forest.predict_proba(X_test)[: , 1]
30
31 tn_l, fp_l, fn_l, tp_l = confusion_matrix(
32     y_test, (proba_lr >= 0.50).astype(int)).ravel()
33 tn_r, fp_r, fn_r, tp_r = confusion_matrix(
34     y_test, (proba_rf >= 0.50).astype(int)).ravel()
35 print("Logistic: TN", tn_l, "FP", fp_l, "FN", fn_l, "TP", tp_l)
36 print("Forest : TN", tn_r, "FP", fp_r, "FN", fn_r, "TP", tp_r)

```

شيفرة 9.10: الجزء الأول من الملف — الترميز والتقسيم والنموذجان ومصفوفا الارتباك.

```

Encoded features: 11 | test patients: 225 | diseased: 96
Logistic: TN 107 FP 22 FN 37 TP 59
Forest : TN 99 FP 30 FN 57 TP 39

```

المهمة 4: جدول المقاييس. حساب بسيط من الخلايا المفككة، مجمّع في إطار واحد جاهز للتقرير.

```

1 fpr_l, tpr_l, thr_l = roc_curve(y_test, proba_lr)
2 fpr_r, tpr_r, thr_r = roc_curve(y_test, proba_rf)
3 n = len(y_test)
4 dossier = pd.DataFrame(
5     {"Logistic": [tp_l/(tp_l+fn_l), tn_l/(tn_l+fp_l), tp_l/(tp_l+fp_l),
6                  tn_l/(tn_l+fn_l), (tp_l+tn_l)/n, auc(fpr_l, tpr_l)],
7      "Forest": [tp_r/(tp_r+fn_r), tn_r/(tn_r+fp_r), tp_r/(tp_r+fp_r),
8                tn_r/(tn_r+fn_r), (tp_r+tn_r)/n, auc(fpr_r, tpr_r)]},
9     index=["Sensitivity", "Specificity", "PPV", "NPV",
10           "Accuracy", "AUC"]).round(3)
11 print(dossier)

```

شيفرة 9.11: الجزء الثاني من الملف — جدول المقاييس الكامل.

	Logistic	Forest
Sensitivity	0.615	0.406
Specificity	0.829	0.767
PPV	0.728	0.565
NPV	0.743	0.635
Accuracy	0.738	0.613
AUC	0.791	0.683

المهام 5-7: تراكم ROC وعتبة المسح والحالات المفقودة. يهيمن النموذج اللوجستي في كل صف — $AUC = 0.791$ (مقبول، درجة فرز) مقابل 0.683 (ضعيف)؛ فيمضي الملف قدماً بالنموذج اللوجستي. الشكل 9.4 يعرض التراكم المقيس.

```

1 thr_grid = np.round(np.arange(0.10, 0.91, 0.10), 2)
2 grid = (proba_lr[:, None] >= thr_grid).astype(int)
3 truth = y_test.to_numpy()[:, None]
4 tp_g = ((grid == 1) & (truth == 1)).sum(axis=0)
5 fn_g = ((grid == 0) & (truth == 1)).sum(axis=0)
6 tn_g = ((grid == 0) & (truth == 0)).sum(axis=0)
7 fp_g = ((grid == 1) & (truth == 0)).sum(axis=0)

```

```

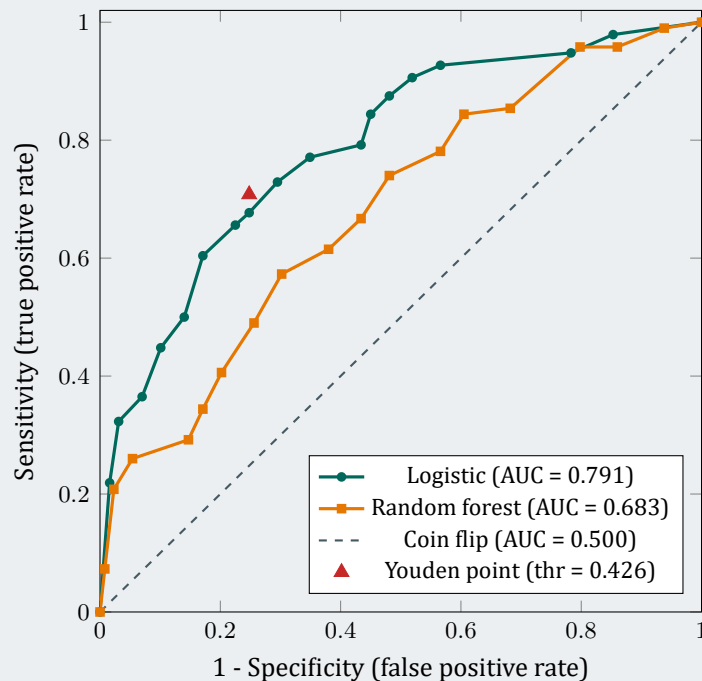
8
9 prev = 0.10
10 sweep = pd.DataFrame(
11     {"threshold": thr_grid,
12      "sensitivity": tp_g / (tp_g + fn_g),
13      "specificity": tn_g / (tn_g + fp_g)})
14 sweep["missed_per_1000"] = 1000 * prev * (1 - sweep["sensitivity"])
15 sweep["alarms_per_1000"] = 1000 * (1 - prev) * (1 - sweep["specificity"])
16 print(sweep[sweep["threshold"].isin([0.20, 0.30, 0.40, 0.50])
17       .round(3).to_string(index=False))
18
19 j = tpr_1 - fpr_1
20 best = np.argmax(j)
21 print(f"Youden: thr={thr_1[best]:.3f} "
22       f"sens={tpr_1[best]:.3f} spec={1-fpr_1[best]:.3f}")

```

شيفرة 9.12: الجزء الثالث من الملف — مسح العتبات مع حساب المسح لكل 1000 عند انتشار 10 بالمئة.

threshold	sensitivity	specificity	missed_per_1000	alarms_per_1000
0.2	0.938	0.380	6.250	558.140
0.3	0.865	0.535	13.542	418.605
0.4	0.719	0.713	28.125	258.140
0.5	0.615	0.829	38.542	153.488

Youden: thr=0.426 sens=0.708 spec=0.752



شكل 9.4: تراكم ROC المقيس لنموذجي heart_risk.csv التجريبيين (نقاط مخففة). يهيمن المنحنى اللوجستي على الغابة في كل مكان؛ ولا يقترب أي منهما من الركن الأعلى الأيسر، ولهذا توصي المذكرة أدناه بنشر مرحلي لا بالاحتفال.

التفسير السريري. عند العتبة الافتراضية 0.50 سيفوت النموذج اللوجستي التجريبي 385 من كل 1000 مريض مصاب فعلاً (حساسية 0.615) — أمر لا يمكن الدفاع عنه في المسح. خفض العتبة إلى 0.30 يرفع الحساسية إلى 0.865 (نحو 14 حالة مفوتة لكل 1000 مفحوص عند انتشار 10 بالمئة، مقابل 39 عند العتبة 0.50)؛ والخفض أبعد إلى 0.20 يبلغ حساسية 0.938 ونحو 6

حالات مفوّنة لكل 1000، لكنه يدفع الإنذارات الكاذبة إلى 558 لكل 1000 — أي أن أكثر من نصف المفحوصين سيحتاجون فحوصات تأكيدية، وهو ما سيُشبع الطاقة الاستيعابية لقسم القلب. نقطة يودن (0.426) هي الافتراض المتوازن لكنها تقصّر عن خدمة مهمة مسح يمكن أن تكون فيها الحالة القلبية المفوّنة قاتلة بينما الاختبار التأكيدي (تخطيط قلب مع استشارة) رخيص وآمن. لذا فالعتبة 0.30 هي التوصية القابلة للدفاع: تحصر الحالات المفوّنة في نحو 14 لكل 1000 مفحوص مع إبقاء حمل التأكيد قرب 419 لكل 1000، على حافة الجدوى لنشر مرحلي. أما PPV المسح الضمنية فمخفضة (نحو 17 بالمئة عند انتشار 10 بالمئة، بحساب الجدول 9.2) — وهذا طبيعي لأدوات المسح، لكن يجب إبلاغه كي تتوقع العيادات تأكيدات معظمها سلبية.

مذكّرة المهمة 7 إلى المدير العام (خمس جمل). (1) قيّمنا نموذجين تجريبيين على 225 مريضاً محجوزين عن التدريب؛ والنموذج اللوجستي متفوق في كل مقياس ($AUC = 0.791$ مقابل 0.683) وهو المرشح الوحيد الذي نوصي بالمضي به قدماً. (2) عند العتبة الافتراضية للبرمجات سيفوت 38 بالمئة من الحالات القلبية الحقيقية، لذا نوصي بالتشغيل عند عتبة مخفضة قدرها 0.30، مع تقديم الحساسية كما تقتضي مهمة المسح. (3) عند نقطة التشغيل تلك نتوقع نحو 14 حالة مفوّنة و419 إنذاراً كاذباً مولّداً للإحالة لكل 1000 مواطن مفحوص، بافتراض انتشار حقيقي قدره 10 بالمئة، ويجب رصد موازنة لطاقة تخطيط القلب التأكيدي وفقاً لذلك. (4) ولأن $AUC = 0.791$ درجة فرز لا درجة تشخيص، يجب أن يعمل النموذج حصراً كمساعد إحالة — لا حكماً مستقلاً أبداً — وننصح بمشروع تجريبي خاضع للإشراف على مستوى قضاء واحد مع إعادة حساب شهرية لهذا الملف قبل أي نشر وطني. (5) كل الأرقام في هذا الملف أُعيد حسابها يدوياً من مصفوفات الارتباك بواسطة محلّلين مستقلّين، وفق سياسة التدقيق في المديرية.

نقاش جماعي. اختتم التحدي بعشر دقائق حول ثلاثة أسئلة:

1. في سياق المسح العراقي، ما الكلف اللاحقة الفعلية لسالب كاذب واحد مقابل إيجابي كاذب واحد في مرض القلب — سريراً ومالياً وفي الثقة العامة؟ هل تصمد العتبة 0.30 أمام هذه المحاسبة، أم ستختار لجنتك 0.20؟
2. قيمة AUC في الملف البالغة 0.791 مقبولة بحسب الجدول 9.3. هل توقّع عليها للنشر الوطني، أم لمشروع تجريبي في قضاء، أم تعيدها طلباً لميزات أفضل؟ وما الدليل الإضافي الذي سيغير قرارك؟
3. عرضٌ من مورّد يُظهر لوحة معلومات تبلغ عن ذكاء اصطناعي واحد بدقة 97 بالمئة. باستخدام مفردات اليوم، عدّد الأسئلة الخمسة التي ستطرحها قبل نهاية الاجتماع — وأي جواب سيُنبي الاجتماع فوراً.

9.4 الواجب المنزلي والتقييم الذاتي

الواجب المنزلي.

1. أعد تدقيق نموذج السكري عند عتبة يودن (0.374): أعد بناء مصفوفة الارتباك، وأعد حساب المقاييس الأربعة، واكتب ثلاث جمل تقارنها بنقطتي التشغيل 0.50 و0.30 من الشيفرة 9.6.
2. باستخدام الحساسية المقيسة (0.865) والنوعية المقيسة (0.535) فقط للنموذج اللوجستي القلبي عند العتبة 0.30، احسب PPV وNPV يدوياً عند انتشار 5 بالمئة، متبعاً طريقة الجدول 9.2. لا حاجة إلى شيفرة — أظهر حساب الخلايا الأربع لعدد 10,000 مفحوص.
3. وسّع مسح البث في الشيفرة 9.7 إلى العتبات 0.05-0.95 بخطوات قدرها 0.05، ثم استخدم قناعاً منطقياً على DataFrame الناتج للإبلاغ عن أدنى عتبة تبلغ نوعيتها 0.80 على الأقل.
4. اقرأ ملخص مقالة PubMed Central رقم PMC7695228 عن AUC المستوفاة خطياً للاختبارات الثنائية، ثم اكتب فقرة واحدة تشرح فيها لزميل لماذا قد تبخس قيمة AUC المبلّغة 0.91 لعدة مستضد سريع قيمةً العدة نفسها — ولماذا مقارنتها مباشرة بقيمة AUC لنموذج لوجستي خطأ تصنيفي.
5. (تحديّ إضافي) مع `prev = np.arange(0.01, 0.51, 0.01)`، احسب منحني PPV لاختبار حساسيته 0.90 ونوعيته 0.90 كتعبير موجّه واحد، وارسم PPV مقابل معدل الانتشار. عند أي معدل انتشار تتجاوز PPV قيمة 0.50 لأول مرة؟

اختبار التقييم الذاتي.

1. اختبار حساسيته 80% ونوعيته 95%. من بين 100 مريض مصاب فعلاً، كم مريضاً يفوت؟
2. أي من هذه المقاييس يتغير عندما ينتقل النموذج نفسه من عيادة إحالة إلى مسح مجتمعي: الحساسية، النوعية، PPV، AUC؟
3. ماذا تعني $AUC = 0.5$ في جملة واحدة؟

4. لبرنامج مسح وطني مع اختبار تأكيد رخيص، هل ينبغي أن تتحرك عتبة القرار فوق القيمة الافتراضية 0.50 أم تحتها، وأي مقياس تحمي بذلك؟

5. في scikit-learn، ما ترتيب التفكيك الدقيق لناتج $\text{cm.ravel}()$ لمصفوفة ارتباط ثنائية بالتسميات $[0, 1]$ ؟
أجوبة الاختبار. (1) 20 مريضاً — $FN = (1 - 0.80) \times 100$. (2) فقط PPV وNPV: القيم التنبؤية تعتمد على معدل الانتشار؛ أما الحساسية والنوعية وAUC فخصائص للاختبار نفسه. (3) يترتب النموذج مريضاً عشوائياً أعلى من سليم عشوائياً نصف الوقت فقط — رمية عملة بلا أي معلومة تشخيصية. (4) تحت 0.50: خفض العتبة يرفع الحساسية ويحمي من السوابب الكاذبة، بثمن مقبول هو مزيد من الإنذارات الكاذبة. (5) tn , fp , fn , tp — السوابب الحقيقية أولاً والإيجابيات الحقيقية أخيراً.

♣ القيمة لمديرية التشخيص — ثقافة تقييم، لا مجرد مقياس

مهارة اليوم تحوّل المديرية من مستهلك لادعاءات الموردين إلى مدقق لها. عملياً: (1) كل عملية شراء لعدة تشخيصية أو منتج ذكاء اصطناعي يمكنها الآن أن تطالب — وأن تعيد حساب — الحساسية والنوعية والقيم التنبؤية عند الانتشار المحلي، والعتبة التي أنتجت عندها الأرقام؛ (2) النماذج التجريبية المبنية داخلياً تُسلّم مع ملف قياسي (مصفوفة، وجدول مقاييس، وتراكب ROC ونقطة تشغيل، وحالات مفوّنة لكل 1000) يستطيع مدير عام غير تقني التصرف بناءً عليه في خمس جمل؛ (3) قرارات العتبات للبرامج الوطنية تصبح خيارات سياسة موثقة قابلة للدفاع — الحساسية أولاً للمسح، والنوعية أولاً قبل التأكيد الباضع — بدل افتراضات برمجية صامتة؛ و(4) عادة تدقيق AUC الثنائية تحمي الوزارة من الانحياز للأتمتة في كل لوحة معلومات تُباع لها مستقبلاً. هذا هو الفرق بين شراء التقنية وحوكمتها.

خلاصات اليوم

- الدقة وحدها فخ: عند انتشار 2 بالمئة، نموذج «الجميع سلمي» دقته 98 بالمئة وقيّمته السريرية معدومة. بلّغ دائماً عن مجموعة المقاييس الكاملة مع معدل الانتشار الذي قيست عنده.
 - مصفوفة الارتباك هي الوثيقة الأولية. بلغة سريرية: TP التّقط، TN بُرئ، FP إنذار كاذب، FN تشخيص مفوّت — وFN هو الخطأ الأقل مغفراً في الطب.
 - الحساسية $(TP / (TP + FN))$ والنوعية $(TN / (TN + FP))$ ترتحلان مع الاختبار؛ أما PPV وNPV فتخصان الاختبار في مجموعة سكانية وتنهان مع هبوط الانتشار — من 69 بالمئة إلى 16 بالمئة في مثالنا المحلول.
 - عتبة القرار رافعة سياسة سريرية، لا افتراض برمجيات: تحريك نموذج السكري من 0.50 إلى 0.30 خفض مرضى السكري المفوّتين من 23 إلى 14 بثمن مرئي مقيس قدره 20 إنذاراً كاذباً إضافياً.
 - منحى ROC يعرض كل المقايضات المتاحة دفعة واحدة؛ وAUC تلخص مهارة الترتيب (0.854 لنموذجنا اللوجستي — ممتاز؛ و0.791 للنموذج القلي التجريبي — درجة فرز فقط).
 - اختر نقطة التشغيل بحسب المهمة: الحساسية أولاً للمسح، والنوعية أولاً قبل علاج باضع، ومؤشر يودن J هو الافتراض المتوازن القابل للدفاع (يوجد بعملية np.argmax واحدة).
 - دقق كل مقياس محسوب: قيمة AUC المستوفاة لاختبار ثنائي هي مجرد $(\text{Sens} + \text{Spec}) / 2$ وقد تبخس الاختبار — والانحياز للأتمتة خطأ تشخيصي، فأعد حساب الأرقام الأولية بنفسك.
 - التقييم لا يكتمل حتى يسعه مذكرة: النموذج، والعتبة، والحالات المفوّنة لكل 1000، وكلفة الطاقة الاستيعابية، وشرط النشر — خمس جمل يستطيع المدير العام التصرف بناءً عليها.
- غداً، في الفصل 10، نجمع كل شيء في صيغة إنتاجية: مسارات عمل scikit-learn التي تسلسل التنظيف والمعايرة والنمذجة في كائن واحد قابل للتدقيق، وأتمتة وأخلاقيات نشر هذه الأنظمة داخل وزارة، ومشروع التخرج الذي يسلم فيه فريقك أداة تشخيصية كاملة — مقيّمة ومضبوطة العتبة وموثقة وفق المعيار الذي وضعت أنت اليوم.

باب 10

اليوم 10 — مسارات العمل والأتمتة والأخلاقيات ومشروع التخرج

قبل تسعة أيام لم يكن معظم الحاضرين في هذه القاعة قد كتبوا سطرًا واحدًا من بايثون. ومنذ ذلك الحين حملتم ملفات المختبر، وأصلحتهم جداول بيانات بدت ميؤوساً منها، ورسمتم أشكالاً بمستوى النشر العلمي، وهندستم الميزات، ودربتم المصنفات، ثم — بالأمس — شرحتهم أداءها بشكوكية اختصاصي علم أمراض خبير يراجع مقاييس جديدة. اليوم نعبّر الجسر الأخير: الجسر بين التحليل والتشغيل. فالنموذج الذي يعيش في دفتر ملاحظات يفيد مجللاً واحداً؛ أما النموذج الذي يعمل كل صباح في الساعة 07:30 فيقيم عينات الليل ويسلم طبيب الخفارة قائمة عمل مرتبة، فيفيد مديرية بأكملها. تتطلب هذه الخطوة عتاداً جديداً — مسار العمل (Pipeline) في scikit-learn، والتحقق المتقاطع (Cross-Validation)، والحفظ الدائم بمكتبة joblib — وتتطلب شيئاً لا يوفره العتاد: إطاراً أخلاقياً. من يحق له رؤية البيانات؟ ومن الذي يضره النموذج بصمت؟ ومن يوقع التقرير حين تختلف الخوارزمية مع الطبيب؟ سنجيب عن كل ذلك، ثم ستثبتون مهاراتهم العشر في مشروع التخرج: نظام فرز متكامل لخطورة أمراض القلب مبني من heart_risk.csv، يُسلم مع ملخص تنفيذي يليق بالمدير العام.

الدقائق	الجزء	المحتوى والأنشطة
00-05	المراجعة والجسر	حكم اليوم التاسع: المقاييس وحدها لا تكفي؛ اليوم ننقل النموذج إلى التشغيل ونحوكمه
05-12	النظرية الأولى: نحو التشغيل	مسار العمل = المعايير + النموذج في مقدّر واحد؛ cross_val_score بخمس طيات؛ الحفظ الدائم بـ joblib؛ التقييم الدفعي
12-20	النظرية الثانية: الأخلاقيات	إخفاء الهوية: HIPAA وGDPR وسياسة وزارة الصحة العراقية؛ التحيز الخوارزمي والانحياز للأتمتة؛ المراقبة والانجراف؛ آفاق الحوسبة الكمومية والنماذج اللغوية
20-35	العرض الموجّه أ	بناء مسار عمل السكري، والتحقق المتقاطع بنزاهة، وقراءة AUC متوسطاً ± انحرافاً معيارياً
35-45	العرض الموجّه ب	الملاءمة، ثم joblib.dump، ثم إعادة التحميل، ثم التحقق على المجموعة المحجوزة
45-60	العرض الموجّه ج	محاكاة صباح الغد: تقييم new_patients.csv، والترتيب، وتصدير قائمة عمل Excel
60-65	إحاطة مشروع التخرج	مشروع التخرج على heart_risk.csv: المخرجات وسلم الدرجات
65-85	العمل على المشروع	فرادى أو أزواجاً؛ المدرب يتجول؛ ويُكشف الحل النموذجي في النهاية
85-90	التخرج	قائمة المهارات العشر، وكلمة مدير الدورة، وختام الدورة

✓ الأهداف التعليمية

بنهاية هذه الجلسة ستكون قادراً على:

1. تجميع Pipeline محكم ضد التسرب يسلسل المعايير القياسية (StandardScaler) ومصنفاً في مقدّر واحد بملاءمة واحدة fit وتنبؤ واحد predict.

2. تقدير الأداء الواقعي بـ `cross_val_score` بخمس طيات وتسجيل ROC-AUC، والإبلاغ عنه متوسطاً $\pm$ انحرافاً معيارياً بدل تقسيم واحد محظوظ.
3. حفظ مسار عمل مدرّب بـ `joblib.dump`، وإعادة تحميله بـ `joblib.load`، والتحقق من أن الكائن المعاد تحميله يسجل النتائج نفسها تماماً.
4. التقييم الدفعي لملف CSV صباحي من مرضى جدد بـ `predict_proba`، وتصدير قائمة عمل مرتبة ومعلّمة إلى Excel بـ `.to_excel`.
5. الشرح، بأمثلة ذات صلة بالمديرية، لأركان الذكاء الاصطناعي الطبي المسؤول: إخفاء الهوية (Anonymization)، والأطر القانونية (HIPAA) و (GDPR) وسياسة وزارة الصحة العراقية، والتحيز الخوارزمي (Algorithmic Bias)، والانحياز للأتمتة (Automation Bias)، والمراجعة بإبقاء الإنسان في حلقة القرار، ومراقبة الانجراف.
6. تسليم مشروع نموذج تشخيصية متكامل — تدقيق، وتحليل استكشافي، وميزات، ومسار عمل، وملف تقييبي، وعتبة مبررة، وقائمة عمل، وملخص تنفيذي — من دون مساعدة.

➔ حقبة أدوات الذكاء الاصطناعي والاعتماديات — مكتبات اليوم

كل استيراد يُستخدم اليوم، في مكان واحد:

```
1 import numpy as np
2 import pandas as pd
3 import matplotlib.pyplot as plt
4 import joblib
5
6 from sklearn.pipeline import Pipeline
7 from sklearn.preprocessing import StandardScaler
8 from sklearn.linear_model import LogisticRegression
9 from sklearn.model_selection import train_test_split, cross_val_score
10 from sklearn.metrics import roc_auc_score, roc_curve, confusion_matrix
```

شيفرة 10.1: استيرادات اليوم العاشر

```
pip install pandas numpy scikit-learn joblib matplotlib openpyxl
```

تبقى **pandas** و **numpy** العمود الفقري لبياناتنا: قراءة ملف CSV الصباحي، وبناء قائمة العمل في إطار بيانات (DataFrame)، والتعليم الموجه بـ `np.where` — لا حلقات تكرارية أبداً. وتوفر **scikit-learn** مسار العمل والتحقق المتقاطع والمقاييس؛ وهي بالنسبة إلى المديرية تحوّل سلسلة هشّة من الخطوات اليدوية إلى كائن واحد قابل للتدقيق. وتسلسل **joblib** ذلك الكائن الملاءم إلى القرص، فيُدرّب النموذج مرة واحدة على يد المحلل ويعاد استخدامه كل صباح على يد أي شخص — إنها الذاكرة المؤسسية لمسار العمل. وترسم **matplotlib** أشكال الملف التقييبي التي تدخل ملف المشروع. وتعمل **openpyxl** بصمت خلف `DataFrame.to_excel`: فهي ما يتيح لسكرتير العيادة فتح قائمة العمل في Excel عادي من دون أن يلمس بايثون أبداً.

10.1 السياق السريري والإطار النظري (20 دقيقة)

تأمل كيف تصل مقايسة مختبرية مصادق عليها إلى الخدمة الروتينية. لا أحد يعيد اشتقاق منحني المعايرة عند كل طاولة عمل، كل صباح، من الذاكرة. تُصادق الطريقة مرة واحدة، وتوثّق، وتُقلّد، ثم تُنفَّذ تنفيذاً متطابقاً آلاف المرات، مع ضبط جودة يراقب الانجراف. ويجب أن يخضع تعلم الآلة للمعيار نفسه. عدو ذلك المعيار هو دفتر الملاحظات المفكك: معايرة مُلاءمة في الخلية 12، ونموذج مُلاءم في الخلية 30، وعتبة مكتوبة يدوياً في الخلية 41. أعد ترتيب الخلايا — أو سلّم الدفتر إلى زميل — فتتغير الطريقة بصمت. مسار العمل Pipeline هو ملف طريقتنا المقفل.

مسار العمل في scikit-learn: مقدّر واحد بلا تسرب

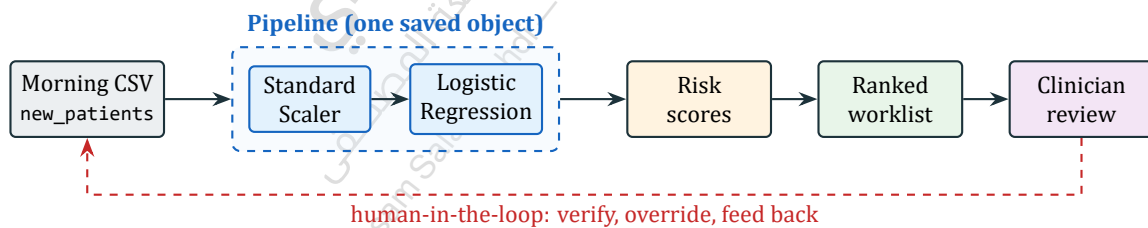
مسار العمل Pipeline قائمة مرتبة من خطوات مسماة، كل خطوة فيها ما عدا الأخيرة محوّل، والأخيرة نموذج — هنا خطوة باسم scaler تحمل معايرة قياسية StandardScaler، ثم خطوة باسم model تحمل انحداراً لوجستياً (Logistic Regression). يسلك الكائن المركب سلوك مقدّر واحد تماماً: `pipe.fit(X_train, y_train)` يلائم المعايرة على بيانات التدريب فقط، ثم يحوّل، ثم يلائم النموذج؛ و `pipe.predict_proba(X_new)` يطبق متوسطات التدريب وانحرافاته المعايرة المخزونة على البيانات الجديدة قبل التنبؤ. وهكذا يصبح فشلان مزمان مستحيلين بنويًا: (1) تسرب البيانات (Data Leakage)، حيث تسترق المعايرة النظر إلى بيانات الاختبار أو المستقبل فتجامل كل المقاييس؛ و (2) نسيان الخطوات (Step Amnesia)، حيث يتلقى نموذج منشور قيمياً خاماً غير معايرة لأن أحدهم نسي المعالجة المسبقة — وهو المكافئ السريري لتشغيل مقايسة ضوئية من دون محلولها الشاهد.

التحقق المتقاطع الخماسي: حكم أعدل من تقسيم واحد

تقسيم واحد بين التدريب والاختبار سحبة واحدة من يانصيب: فمع بضع مئات من المرضى، قد يؤدي انتقال حفنة من الحالات الصعبة بين الجانبين إلى إزاحة AUC عدة نقاط. أما `cross_val_score` مع `cv=5` وتسجيل ROC-AUC فيقسم البيانات إلى خمس طيات، يدرب على أربع ويتحقق على الخامسة، ويدور حتى يخدم كل مريض في التحقق مرة واحدة بالضبط. نورد المتوسط تقديراً لنا والانحراف المعياري مقياساً لاستقراره. ولأن مسار العمل يعاد ملاءمته داخل كل طية، تُعاد أيضاً تقدير المعايرة على كل حصة تدريب — فإجراء التحقق المتقاطع على نموذج مجرد فوق بيانات معايرة مسبقاً يعني التسرب. والتباعد الواسع بين الطيات نتيجة بحد ذاته: أداء النموذج يعتمد على المرضى الذين يراهم، وذلك إنذار بعينة صغيرة أو غير متجانسة.

الحفظ الدائم والتقييم الدفعي: درّب مرة، وقمّ كل صباح

يكتب `joblib.dump` مسار العمل الملاءم بأكمله — متوسطات المعايرة، ومعاملات النموذج، وترتيب الخطوات — في ملف ثنائي واحد مثل `diagnostic_model.joblib`؛ ويبحثه `joblib.load` من جديد في أي جلسة لاحقة. عندئذ يصبح النمط التشغيلي لخدمة في المديرية ثلاثة أسطر: اقرأ ملف الصباح بـ `pd.read_csv`، واحسب عمود الخطوة بـ `predict_proba`، والترتيب بحسب الخطوة والتصدير بـ `to_excel` يحوّلان الاحتمالات إلى قائمة عمل: فالمرور الشحيح في العيادة هو انتباه الكبار، والقائمة المرتبة تنفقه على المرضى الأعلى خطورة أولاً. ويجب أن يعامل ملف النموذج معاملة دفعة الكواشف: مرقّم الإصدار، ومؤرّخ، ولا يحزّر في مكانه أبداً.



شكل 10.1: حلقة التشغيل لليوم العاشر. يطبق مسار العمل المحفوظ المعايرة والتنبؤ فعلاً واحداً لا ينقسم؛ ويغلق الطبيب السريري الحلقة — كل راية تُراجع، وتغذي التجاوزات التدقيق التالي.

الأخلاقيات والحوكمة: الثوابت غير القابلة للتفاوض

نشر النموذج يغيّر منزلته الأخلاقية. ففي دفتر الملاحظات يكلف الخطأ المحلل أمسية من وقته؛ أما في قائمة العمل فهو يقرر أي مريض يُرى أولاً. لذلك تلحق المديرية ستة أركان حوكمة بكل نموذج منشور، يلخصها الشكل 10.2.

إخفاء الهوية وتقليل البيانات

تنتقل المعارف على أساس الحاجة إلى المعرفة. فقبل أي تحليل أو تدريب نموذج تُزال المعارف المباشرة (الاسم، والرقم الوطني، والهاتف، والعنوان الدقيق)، ولا يُحتفظ بالمفاتيح المستعارة مثل `patient_id` إلا حيث يكون ربط السجلات مطلوباً حقاً — ثم تُحذف من جدول النمذجة نفسه. ولاحظ تصميم هذه الدورة: ملفات CSV السبعة كلها اصطناعية، ولدت لتحاكي التوزيعات الحقيقية

من دون مرضى حقيقيين خلفها، وذلك تحديداً ليتسنى التدريب بلا أي خطر على الخصوصية. أما بيانات المديرية الحقيقية فلا يجوز التمرن عليها باستخفاف أبداً. واحذر إعادة تحديد الهوية: رجل في الرابعة والتسعين من العمر أُدخل قسم الأورام من محافظة صغيرة قد يكون قابلاً للتحديد حتى من دون اسم — فالتركيبات النادرة من أشباه المعرفات معرفاتٌ بحد ذاتها.

ينبغي أن تكون ثلاثة أطر قانونية حاضرة على مكتب كل محل. نادراً ما ستُدقق بموجها الثلاثة معاً، لكن كل واحد منها يصوغ واجباً يُعد ممارسة جيدة في كل مكان.

جدول 10.1: الأطر القانونية والسياساتية النازمة للبيانات الصحية.

الإطار	نطاق الاختصاص	الواجب الجوهري عند طاوله العمل
HIPAA (1996)	الولايات المتحدة	لا تُستخدم المعلومات الصحية المحمية ولا يُفصح عنها إلا للأغراض المسموح بها؛ وإخفاء الهوية يجري عبر أسلوب الملاذ الآمن (Safe Harbor) بإزالة 18 نوعاً من المعرفات أو عبر تقدير خبير.
GDPR (2018)	الاتحاد الأوروبي	البيانات الصحية فئة خاصة: أساس قانوني صريح، وتقليل للبيانات، وحق في المحو، ومسألة — يجب أن تكون قادراً على إثبات الامتثال لا مجرد ادعائه.
سياسة وزارة الصحة العراقية	العراق	التزامات سرية المريض بموجب أنظمة وزارة الصحة ومدونة أخلاقيات المهنة الطبية: تبقى السجلات داخل المؤسسة، والوصول قائم على الأدوار، والإفصاح يتطلب تخويلاً؛ ومشاركة البيانات على مستوى المديرية تمر عبر القنوات الرسمية.

التحيز الخوارزمي: التمييز الصامت

لا يستطيع النموذج أن يتعلم إلا الجمهرة التي عُرضت عليه. فإذا أفرط سجل التدريب في تمثيل منطقة واحدة، أو جنس واحد، أو علامة تجارية واحدة من أجهزة التحليل، تدهور الأداء بهدوء لكل الآخرين — ولا تكشفه قيمة AUC على مجموعة اختبارنا نحن أبداً، لأن مجموعة الاختبار تحمل الاختلال نفسه. دقق التمثيل قبل التدريب؛ والفحص سطر واحد:

```
1 import pandas as pd
2
3 demo = pd.read_csv("patients_demographics.csv")
4 print(demo["governorate"].value_counts())
```

شيفرة 10.2: تدقيق التمثيل الجغرافي قبل أي تدريب

```
governorate
Baghdad    336
Basra      154
Nineveh    153
Erbil      149
Anbar      111
Najaf      109
Diyala     97
Karbala    91
Name: count, dtype: int64
```

تساهم بغداد بما يعادل 3.7 أضعاف مرضى كربلاء. والنموذج الإجمالي المدرب على هذا السجل هو، في واقع الأمر، نموذج بغدادي مع ضيوف. والعلاجات صريحة: أورد الأداء مصنفًا بحسب الفئات الفرعية (تماماً كما صُنّفت المقاييس في اليوم التاسع)، واجمع بيانات الفئات ناقصة التمثيل أو أعد وزنها، وصّح بالجمهرة المصادق عليها في بطاقة هوية النموذج — تماماً كما يصحّ المجال المرجعي بالجمهرة التي اشتق منها.

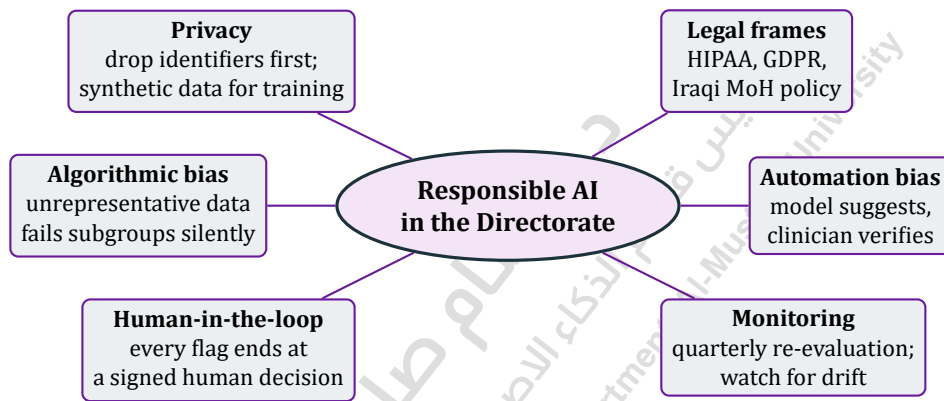
✖ خطأ شائع — الانحياز للأتمتة — منعكس الختم المطاطي

رأيتم بالأمس أن اختباراً نوعيته 95% ما يزال يغمر عيادة منخفضة الانتشار بالإيجابيات الكاذبة. وتعميم اليوم: يفرط البشر في الثقة بالآلات الوائقة. فدراسات دعم القرار السريري تُظهر باستمرار أطباء يقبلون اقتراحات آلية خاطئة كانوا سيلتقطونها من دون مساعدة. قائمة العمل التي ستبنونها تطبع احتمالاً لا حكماً — وسياسة المديرية مكتوبة تبعاً لذلك: النموذج يقترح، والطبيب السريري يتحقق،

وتوقيع الطبيب هو التوقيع الوحيد على التقرير. وإذا ناقضت راية النموذج الصورة السريرية، فالصورة السريرية تنتصر ويُسجل الخلاف لتدقيق النموذج.

الإنسان في حلقة القرار ومراقبة الانجراف سياسةً دائمة

إبقاء الإنسان في حلقة القرار ليس مجاملة؛ إنه بنية التحكم في الشكل 10.1: كل راية آلية تنتهي عند قرار بشري، وتعود التجاوزات البشرية بوصفها بيانات تدقيق. وتكمل المراقبة الحلقة عبر الزمن. فالجملهرات تشيخ، وأنماط الإحالة تتبدل، وجهاز تحليل جديد يغيّر توزيع الكرياتينين، ومتحورة جديدة تغيّر قيم Ct — هذا هو الانجراف (Drift)، وهو يبلي نموذجاً مجمداً تماماً كما ينجرّف جهاز غير معايير خارج ضبط الجودة. جدول المديرية: أعد حساب الملف التقييبي على بيانات موسومة حديثة كل ربع سنة؛ وحقق إذا هبطت AUC أكثر من 0.05 تحت قيمة المصادقة أو إذا تغيّرت نسبة الحالات المعلّمة تغيّراً جوهرياً؛ وأعد التدريب وأصدر نسخة جديدة بدل التحرير في المكان.



شكل 10.2: خارطة الحوكمة التي يجب أن يستوفها كل نموذج منشور في المديرية. لا شيء من الأركان الستة اختياري.

★ لؤلؤة سريرية — رايات النموذج فرز لا تشخيص

نموذج عرض اليوم يتحرى خطورة السكري من النوع الثاني. وعتبة التعليم فيه (خطورة 0.5 $\geq$) خيار تشغيلي بشأن من يُراجع أولاً — وليست معياراً تشخيصياً. فالتشخيص يظل تابعاً للحدود المعتمدة: غلوكوز بلازما صائم ≥ 126 ملغ/دل في مناسبتين، أو سكر تراكمي $\geq 6.5\%$ HbA1c، مع نطاق ما قبل السكري المحدد بين 100–125 ملغ/دل أو بقيم HbA1c بين 5.7–6.4%. وغلوكوز عشوائي ≥ 200 ملغ/دل مع أعراض يُعد تشخيصاً أيضاً. كل مريض يعلمه النموذج يمضي إلى هذا المسار التأكيدي بالضبط؛ وكل مريض لا يعلمه يبقى مستحقاً له عند الاشتباه السريري. أبقى المفردتين منفصلتين في كل تقرير تكتبته: خطورة النموذج مقابل المعيار التشخيصي.

الطريق أمامنا: تقنيتان تستحقان المتابعة

الحوسبة الكمومية. تتعثر الحواسيب التقليدية في محاكاة الجزيئات لأن كلفة تتبع الحالات الكمومية تنفجر مع حجم الجزيء؛ بينما تمثل المعالجات الكمومية تلك الحالات تمثيلاً أصيلاً. وفي الجينوميّات واكتشاف الأدوية يعد ذلك بتحرّ حاسوبي أسرع لمرشحات الأدوية، ومحاكاة أفضل لطّي البروتين والالتحام الجزيئي، وأمثلة توافقية لتخطيط العلاج. أما التقرير الصادق عن الحالة الراهنة فهو: الأجهزة الحالية صغيرة وكثيرة الضجيج، ولا وجود اليوم لأفضلية كمومية في التشخيص الروتيني، ولا شيء مما بنيتومه هذا الأسبوع ينتظرها — لكن المديرية التي تدير بياناتها جيداً أصلاً ستكون أول القادرين على استخدام تلك الخدمات حين تنضج.

النماذج اللغوية الكبيرة. تصوغ النماذج اللغوية الكبيرة (LLM) بالفعل مسودات ملخصات الخروج، وترجم تعليمات المرضى، وتلخص الأدبيات، وتهيك تقارير علم الأمراض الحرة النص — وذلك عون حقيقي لخدمة مثقلة بالتوثيق، ومساعد معقول لبحوث الفرز. أما المحاذير فثابتة وراهنة: يمكن للنماذج اللغوية أن تهلوس عبارات سلسلة وثيقة خاطئة؛ وهي تحمل تحيزات نصوص تدريبها؛ ولصق بيانات المرضى في روبوت محاكاة خارجي خرق للخصوصية بموجب كل إطار في الجدول 10.1. وموقف المديرية يعكس سياسة اليوم نفسها: مخرجات النموذج اللغوي مسودة تخضع لتحقق الطبيب، لا تصريح سريري غير موقع أبداً، وبيانات المرضى المعرّفة للهوية لا تغادر البنية التحتية المؤسسية أبداً.

10.2 التطبيق الموجّه (40 دقيقة)

■ عرض تطبيقي مباشر — من النموذج المدرب إلى قائمة عمل الغد

الهدف: أخذ مسألة تحري السكري من مجموعة بيانات اليوم الثامن إلى التشغيل الكامل. سوف (1) نسلسل المعايير والانحدار اللوجستي في Pipeline واحد، و(2) نحصل على AUC نزيه بتحقيق متقاطع خماسي، و(3) نلائم مسار العمل ونحفظه دائماً بـ joblib، و(4) نعيد تحميله كما سيفعل محلل الخفارة غداً، و(5) نقيم ملف صباح جديداً إلى قائمة عمل Excel مرتبة للعيادة. اعمل في مجلد datasets: فكل الملفات تُقرأ باسمها المجرد.

الخطوة 1: تحميل البيانات وتأطير المهمة

نفصل مصفوفة الميزات عن الناتج. لاحظ معدل الانتشار: نحو 37% — وتذكر من اليوم التاسع أن الانتشار يقود القيمة التنبؤية الإيجابية (PPV)، لذا سنبقيه في أذهاننا حين نقرأ قائمة العمل.

```
1 import pandas as pd
2 import numpy as np
3
4 pd.set_option("display.width", 100)
5
6 df = pd.read_csv("diabetes_dataset.csv")
7 X = df.drop(columns=["outcome"])
8 y = df["outcome"]
9 print("Features:", X.shape, "| prevalence:", y.mean().round(3))
```

شيفرة 10.3: تحميل مجموعة بيانات التحري

```
Features: (768, 8) | prevalence: 0.368
```

الخطوة 2: تجميع مسار العمل

خطوتان مسماتان، وكائن واحد. طباعة مسار العمل تستحق شرحاً بمفردها: هذا النص هو ملف الطريقة — يستطيع أي شخص أن يقرأ بالضبط ما يحدث للعينة، وبأي ترتيب.

```
1 from sklearn.pipeline import Pipeline
2 from sklearn.preprocessing import StandardScaler
3 from sklearn.linear_model import LogisticRegression
4
5 pipe = Pipeline([
6     ("scaler", StandardScaler()),
7     ("model", LogisticRegression(max_iter=1000)),
8 ])
9 print(pipe)
```

شيفرة 10.4: المعايير والنموذج في مقدر واحد

```
Pipeline(steps=[('scaler', StandardScaler()),
                  ('model', LogisticRegression(max_iter=1000))])
```

الخطوة 3: التحقق المتقاطع بنزاهة

خمس طيات، والمساحة تحت المنحنى (AUC) مقياساً. داخل كل طية يعيد scikit-learn ملاءمة مسار العمل بأكمله — بما فيه المعايير — على طيات التدريب الأربع فقط. المتوسط هو تقديرنا؛ والانحراف المعياري يخبرنا كم يعتمد الحكم على المرضى الذين صادف حجزهم للتحقق.

```
1 from sklearn.model_selection import cross_val_score
2
3 auc_folds = cross_val_score(pipe, X, y, cv=5, scoring="roc_auc")
4 print("Fold AUCs:", auc_folds.round(3))
5 print("Mean AUC :", auc_folds.mean().round(3),
6       "+/-", auc_folds.std().round(3))
```

شيفرة 10.5: المساحة تحت المنحنى بالتحقق المتقاطع الخماسي

```
Fold AUCs: [0.819 0.88 0.843 0.821 0.815]
Mean AUC : 0.836 +/- 0.025
```

قيمة $AUC = 0.836 \pm 0.025$ تُقرأ كما يُقرأ تقرير ضبط الجودة: تمييز جيد، واستقرار عبر الطيات. ولو أعادت إحدى الطيات 0.65 لتوقفنا وسألنا أي المرضى يقيمون في تلك الطية قبل الوثوق بأي شيء لاحق.

✖ خطأ شائع — المعايير قبل التقسيم — التسرب الذي لا تراه

الخطأ الكلاسيكي الذي وُجد هذا المسار لمنعه: استدعاء `StandardScaler().fit_transform(X)` على مجموعة البيانات الكاملة ثم إخضاع النموذج المجرد للتحقق المتقاطع بعد ذلك. عندئذٍ تكون كل طية تحقق قد همست بمتوسطاتها وانحرافاتها المعيارية في عملية التدريب — فتتضخم المقاييس، ولا شيء في المخرجات يندرك. القاعدة مطلقة: المعالجة المسبقة تعيش داخل مسار العمل، ومسار العمل هو ما تخضعه للتحقق المتقاطع. وإذا انخفضت قيمة AUC قليلاً حين نقلت المعايير إلى الداخل — فذلك خير؛ لقد غادرت المجاملة الغرفة للتو.

الخطوة 4: الملاءمة على مجموعة التدريب والحفظ الدائم

نحجز الآن مجموعة اختبار نهائية — بتقسيم طبقي، كما هو المعيار منذ اليوم السابع — ونلائم مسار العمل على حصة التدريب، ونجمّد النتيجة على القرص. في عرض النشر هذا نحجز عمداً 20%، وهو عرف إنتاجي شائع؛ بينما يعود مشروع التخرج إلى بروتوكول 75/25 المؤلف. يحتوي ملف `joblib` على كل شيء: إحصاءات المعايير، ومعاملات النموذج، وترتيب الخطوات.

```
1 from sklearn.model_selection import train_test_split
2 import joblib
3
4 X_train, X_test, y_train, y_test = train_test_split(
5     X, y, test_size=0.2, stratify=y, random_state=42)
6 pipe.fit(X_train, y_train)
7 joblib.dump(pipe, "diagnostic_model.joblib")
8 print("Saved: diagnostic_model.joblib")
```

شيفرة 10.6: ملاءمة الطريقة وتجميدها

```
Saved: diagnostic_model.joblib
```

الخطوة 5: إعادة التحميل والتحقق — أول أعمال الغد

محلل مختلف، ويوم مختلف، وجلسة بايثون جديدة: استدعاء واحد لـ `joblib.load` يبعث الطريقة من جديد. نتحقق على مجموعة الاختبار المحجوزة قبل أن نأتمنه على عمل حقيقي — وهو ما يعادل عند المحلل تشغيل عينة ضبط قبل وجبة المرضى.

```

1 from sklearn.metrics import roc_auc_score
2
3 model = joblib.load("diagnostic_model.joblib")
4 proba_test = model.predict_proba(X_test)[: , 1]
5 print("Hold-out AUC of reloaded pipeline:",
6       round(roc_auc_score(y_test, proba_test), 3))

```

شيفرة 10.7: إعادة تحميل مسار العمل المجمع والتحقق منه

```
Hold-out AUC of reloaded pipeline: 0.848
```

قيمة AUC على المجموعة المحجوزة (0.848) تجلس بارتياح داخل نطاق التحقق المتقاطع من الخطوة 3 — فالملف المجمع يسلك تماماً سلوك الكائن الذي صادقنا عليه. يدخل هذا الرقم سجل النموذج بوصفه قيمة إصداره.

الخطوة 6: محاكاة عينات الغد

لتمرين روتين الصباح نقتطع عشرين مريضاً غير مرئيين من مجموعة الاختبار في ملف يؤدي دور وافيدي الليل. في الإنتاج يأتي هذا الملف من تصدير نظام معلومات المختبر (LIS) ولن تتغير الشيفرة التالية بحرف واحد.

```

1 X_test.head(20).to_csv("new_patients.csv", index=False)
2 print("new_patients.csv written:", X_test.head(20).shape)

```

شيفرة 10.8: إنشاء ملف الصباح المحاكى

```
new_patients.csv written: (20, 8)
```

الخطوة 7: التقييم الدفعي والترتيب وتصدير قائمة العمل

هذا قلب سكرت الصباح. يعيد predict_proba عمودين (احتمال الفئة 0 واحتمال الفئة 1)؛ والعمود [1, :] هو خطورة السكري. ويحول np.where الخطورة إلى راية من دون أي جملة if، والترتيب التنازلي يضع المرضى الأكثر إيجاءً بالمرض في الأعلى. ثم يسلم to_excel العيادة ملفاً تستطيع فتحه قبل أول كوب شاي.

```

1 new_patients = pd.read_csv("new_patients.csv")
2 risk = model.predict_proba(new_patients)[: , 1]
3 worklist = pd.DataFrame({
4     "case_id": "NEW-" + (new_patients.index + 1).astype(str).str.zfill(2),
5     "glucose": new_patients["glucose"],
6     "bmi": new_patients["bmi"],
7     "risk_pct": (risk * 100).round(1),
8     "flag": np.where(risk >= 0.5, "REVIEW", "routine"),
9 })
10 worklist = worklist.sort_values("risk_pct", ascending=False)
11 print(worklist.head(8).to_string(index=False))
12 worklist.to_excel("morning_worklist.xlsx", index=False)
13 print("Exported morning_worklist.xlsx |", (risk >= 0.5).sum(),
14       "of 20 flagged for review")

```

شيفرة 10.9: قائمة عمل الصباح في كتلة موجهة واحدة

```

case_id  glucose  bmi  risk_pct  flag
NEW-16    177  41.7    91.2  REVIEW
NEW-10    170  42.1    85.3  REVIEW
NEW-14    144  30.2    72.4  REVIEW

```

```
NEW-09      123 32.6      70.8 REVIEW
NEW-11      147 38.7      67.9 REVIEW
NEW-05      149 34.7      67.6 REVIEW
NEW-04      134 38.7      64.8 REVIEW
NEW-06      172 33.2      59.2 REVIEW
Exported morning_worklist.xlsx | 9 of 20 flagged for review
```

اقرأ الصفوف العليا قراءة سريرية: يجمع المريض NEW-16 غلوكوزاً قدره 177 ملغ/دل — وهو أصلاً فوق معيار الصيام 126 ملغ/دل إن كانت العينة صائمة — مع مؤشر كتلة جسم 41.7؛ فخطورة 91% ليست مفاجأة، وقد دفع النموذج هذا المريض بحق إلى رأس الطابور. وتسع رايات من عشرين حمل صباحي قابل للمراجعة. يعمل الطبيب الآن على القائمة من أعلاها، ويؤكد عبر المسار التشخيصي الوارد في اللؤلؤة السريرية، ويوقع — فالنموذج لم يكتب حرفاً واحداً من التقرير.

✖ خطأ شائع — لا تستدعِ fit عند طاولة العمل أبداً

يحتوي سكرت الصباح على `joblib.load` و `predict_proba` — ولا يحتوي عمداً على أي `fit`. لإعادة الملاءمة على ملف الصباح ستحاول التعلم من عشرين مريضاً بلا وسوم (وهي تفضل — إذ لا يوجد `y`)، وإعادة الملاءمة على ملفات الصباح مع وسوم متعجلة ستحوّر الطريقة المصادق عليها بصمت: فالملف على القرص لم يعد يطابق الملف التقييمي الذي اعتمده. وبالقدر نفسه من الأهمية: يجب أن يحمل ملف CSV الجديد أعمدة التدريب بالضبط — إذ يفحص `scikit-learn` أسماء الميزات وسيرفرض ملفاً أعمدته ناقصة أو معاد تسميتها. وذلك الخطأ صديق: إنه مسار عملك يرفض عينة ملصقها تالف.

10.3 التحدي السريري (30 دقيقة)

تحدي اليوم هو مشروع التخرج. وهو يحل محل التحدي الموجه المعتاد: تتسلمون السيناريو والمخرجات المطلوبة وتبنون — فرادى أو أزواجاً — والمدرّب يتجول بينكم. كل ما تحتاجونه دُرّس بين اليوم الأول واليوم العاشر.

📌 التحدي السريري — مشروع التخرج: خدمة فرز الخطورة القلبية للمديرية

عيادة إحالات أمراض القلب في المديرية مثقلة: كل مريض محال ينتظر حالياً في طابور واحد غير مفاضل. يُسلم إليك `heart_risk.csv` — 900 مريض مقيّم بناتج مؤكد (`target: 1` = مرض قلبي معتد به) — ويُطلب منك بناء نموذج الفرز الاسترشادي، من البداية إلى النهاية، تماماً كما لو كان موجهاً للإنتاج. المخرجات:

1. الاستيراد والتدقيق: الشكل، والتكرارات، والقيم المفقودة (Missing Values)، وتوازن الفئات، وفحص معقولية على `resting_bp` و `cholesterol` و `max_hr` باستخدام أقنعة `between` الموجهة (انضباط اليوم الرابع).
2. تحليل استكشافي سريع: متوسطات مجمعة بحسب `target` مع شكلين محفوظين (مخطط صندوق `st_depression` بحسب الناتج، ومخطط انتشار للعمر مقابل `max_hr` ملوّن بحسب الناتج).
3. هندسة الميزات (Feature Engineering) بوصفة اليوم السادس: ترميز ثنائي لـ `sex` و `exercise_angina` بالأقنعة المنطقية (Boolean Masks)؛ وترميز أحادي التنشيط لـ `chest_pain_type` بـ `pd.get_dummies`.
4. تقسيم طبقي (75/25)، مع `random_state=42` و `Pipeline` من `StandardScaler` ومصنف من اختيارك؛ وأورد AUC بالتحقق المتقاطع الخماسي على مجموعة التدريب.
5. الملف التقييمي على مجموعة الاختبار: AUC، ومصفوفة الارتباك (Confusion Matrix)، والحساسية، والنوعية، والقيمة التنبؤية الإيجابية والسلبية عند العتبة الافتراضية 0.50.
6. اختر عتبة تشغيل وبررها لحالة استخدام فرز يكون فيها تفويت المرض الحقيقي هو الخطأ المكلف؛ ثم أعد حساب الملف التقييمي عند عتبتك.
7. التصدير: قائمة عمل مرتبة (`to_excel`) وحفظ مسار العمل دائماً بـ `joblib.dump`.
8. الملخص التنفيذي: عشرة أسطر إلى المدير العام — ما الذي بُني، وما مدى جودة عمله، وما حدوده، وما توصيتك بشأن النشر.

✓ حل المدرب — الحل النموذجي لمشروع التخرج

(أ) الاستيراد والتدقيق وتنظيف المعقولة. التدقيق يأتي أولاً، دائماً: الشكل، والتكرارات، والنقص، والتوازن، والمدايات.

```
1 import pandas as pd
2 import numpy as np
3
4 pd.set_option("display.width", 100)
5
6 hd = pd.read_csv("heart_risk.csv")
7 print("Shape:", hd.shape, "| duplicates:", hd.duplicated().sum())
8 print("Missing values per column (total):", hd.isna().sum().sum())
9 print(hd["target"].value_counts(normalize=True).round(3))
10 print(hd[["resting_bp", "cholesterol", "max_hr"]].describe().round(1))
11
12 plausible = (hd["resting_bp"].between(70, 250)
13             & hd["cholesterol"].between(80, 700)
14             & hd["max_hr"].between(50, 230))
15 print("Rows failing plausibility checks:", (~plausible).sum())
16 hd = hd[plausible].drop_duplicates().copy()
17 print("Clean shape:", hd.shape)
```

شيفرة 10.10: الجزء أ من مشروع التخرج: التدقيق والتنظيف

```
Shape: (900, 10) | duplicates: 0
Missing values per column (total): 0
target
0    0.572
1    0.428
Name: proportion, dtype: float64
      resting_bp  cholesterol  max_hr
count      900.0         900.0    900.0
mean       132.0         245.2    163.7
std        17.3          52.9     19.2
min         92.0         110.0    101.0
25%        120.0         210.0    152.0
50%        131.0         246.0    164.0
75%        144.0         282.0    177.0
max        180.0         422.0    202.0
Rows failing plausibility checks: 0
Clean shape: (900, 10)
```

ملف نظيف — ومع ذلك فقد استحق التدقيق أجره، لأن بوسعنا الآن أن نصرح بأنه نظيف بدل أن نأمل. معدل الانتشار 42.8% أعلى بكثير من أي انتشار مجتمعي: هذه جمهرة محالة، والقيمة التنبؤية الإيجابية التي سنحسبها لن تنتقل إلى سياق تحرّ مجتمعي (درس اليوم التاسع).

(ب) تحليل استكشافي سريع. تُظهر المتوسطات المجمعة التدرجات السريية المتوقعة: الحالات أكبر سناً، وتبلغ ذروة نبض أدنى، وتحمل ضعف انخفاض القطعة ST.

```
1 import matplotlib.pyplot as plt
2
3 print(hd.groupby("target")[["age", "max_hr", "st_depression"]].
4       .mean().round(1))
5 fig, ax = plt.subplots(1, 2, figsize=(10, 4))
6 hd.boxplot(column="st_depression", by="target", ax=ax[0])
7 hd.plot.scatter(x="age", y="max_hr", c="target", cmap="coolwarm",
```

```

8         colorbar=False, ax=ax[1])
9 plt.savefig("capstone_eda.png", dpi=150)
10 print("Saved capstone_eda.png")

```

شيفرة 10.11: الجزء ب من مشروع التخرج: التحليل الاستكشافي

	age	max_hr	st_depression
target			
0	52.0	167.8	0.6
1	56.8	158.2	1.0

Saved capstone_eda.png

(ج) هندسة الميزات — وصفة اليوم السادس. أقنعة منطقية للمتغيرات الفئوية، و `get_dummies` لنوع ألم الصدر رباعي المستويات.

```

1 hd["sex_male"] = (hd["sex"] == "M").astype(int)
2 hd["angina_flag"] = (hd["exercise_angina"] == "Y").astype(int)
3 hd = pd.get_dummies(hd, columns=["chest_pain_type"], prefix="cp",
4                     dtype=int)
5 features = ["age", "sex_male", "resting_bp", "cholesterol",
6             "fasting_bs", "max_hr", "angina_flag", "st_depression",
7             "cp_ASY", "cp_ATA", "cp_NAP", "cp_TA"]
8 X = hd[features]
9 y = hd["target"]
10 print("Feature matrix:", X.shape)

```

شيفرة 10.12: الجزء ج من مشروع التخرج: الميزات

```
Feature matrix: (900, 12)
```

(د) التقسيم الطبقي ومسار العمل والتحقق المتقاطع والملاءمة.

```

1 from sklearn.pipeline import Pipeline
2 from sklearn.preprocessing import StandardScaler
3 from sklearn.linear_model import LogisticRegression
4 from sklearn.model_selection import train_test_split, cross_val_score
5 from sklearn.metrics import roc_auc_score
6
7 X_train, X_test, y_train, y_test = train_test_split(
8     X, y, test_size=0.25, stratify=y, random_state=42)
9 pipe = Pipeline([
10     ("scaler", StandardScaler()),
11     ("model", LogisticRegression(max_iter=1000)),
12 ])
13 cv_auc = cross_val_score(pipe, X_train, y_train, cv=5,
14                          scoring="roc_auc")
15 print("5-fold CV AUC:", cv_auc.mean().round(3),
16       "+/-", cv_auc.std().round(3))
17 pipe.fit(X_train, y_train)
18 proba = pipe.predict_proba(X_test)[: , 1]
19 print("Test-set AUC :", round(roc_auc_score(y_test, proba), 3))

```

شيفرة 10.13: الجزء د من مشروع التخرج: مسار العمل والمصادقة

```
5-fold CV AUC: 0.751 +/- 0.032
Test-set AUC : 0.791
```

(هـ) الملف التقييمي واختيار العتبة. عند العتبة الافتراضية 0.50 وعند 0.35، وهي عتبة الفرز التي سندافع عنها.

```
1 from sklearn.metrics import confusion_matrix
2
3 pred = (proba >= 0.50).astype(int)
4 tn, fp, fn, tp = confusion_matrix(y_test, pred).ravel()
5 print("At 0.50 -> TP", tp, "FP", fp, "FN", fn, "TN", tn)
6 print("Sens", round(tp / (tp + fn), 3),
7       "Spec", round(tn / (tn + fp), 3),
8       "PPV", round(tp / (tp + fp), 3),
9       "NPV", round(tn / (tn + fn), 3))
10
11 pred = (proba >= 0.35).astype(int)
12 tn, fp, fn, tp = confusion_matrix(y_test, pred).ravel()
13 print("At 0.35 -> TP", tp, "FP", fp, "FN", fn, "TN", tn)
14 print("Sens", round(tp / (tp + fn), 3),
15       "Spec", round(tn / (tn + fp), 3),
16       "PPV", round(tp / (tp + fp), 3),
17       "NPV", round(tn / (tn + fn), 3))
```

شيفرة 10.14: الجزء هـ من مشروع التخرج: الملف التقييمي عند عتبتين

```
At 0.50 -> TP 59 FP 22 FN 37 TN 107
Sens 0.615 Spec 0.829 PPV 0.728 NPV 0.743
At 0.35 -> TP 75 FP 47 FN 21 TN 82
Sens 0.781 Spec 0.636 PPV 0.615 NPV 0.796
```

منحنى خاصية تشغيل المستقبل (ROC) الخاص بالملف التقييمي:

```
1 from sklearn.metrics import roc_curve
2
3 fpr, tpr, thr = roc_curve(y_test, proba)
4 plt.figure(figsize=(5, 4))
5 plt.plot(fpr, tpr, color="teal")
6 plt.plot([0, 1], [0, 1], linestyle="--", color="gray")
7 plt.xlabel("1 - Specificity")
8 plt.ylabel("Sensitivity")
9 plt.savefig("capstone_roc.png", dpi=150)
10 print("Saved capstone_roc.png")
```

شيفرة 10.15: الجزء هـ من مشروع التخرج: شكل المنحنى

```
Saved capstone_roc.png
```

تبرير العتبة. عند 0.50 يفوّت النموذج 37 من أصل 96 حالة قلبية حقيقية (حساسية 61.5%) — وهذا غير مقبول في الفرز، حيث تعني الحالة المفوتة أن مريضاً قلبياً ينتظر في مؤخرة الطابور. وخفض العتبة إلى 0.35 يستعيد 16 من تلك التفويتات (حساسية 78.1%، وعدد السلبيات الكاذبة 21) لقاء 25 إيجابية كاذبة إضافية؛ والإجراء التالي — تخطيط قلب معجل واستشارة — رخيص وغير ضار، فاللاتناظر يرجّح كفة الحساسية. وترتفع القيمة التنبؤية السلبية إلى 79.6%، فالمرضى غير المعلم آمن بدرجة معقولة، وإن لم تكن كاملة، في الطابور الروتيني — ولهذا يبقى ممر المراجعة البشرية مفتوحاً في الاتجاهين.

(و) قائمة العمل والحفظ الدائم والملخص التنفيذي.

```

1 import joblib
2
3 capstone_list = pd.DataFrame({
4     "case_id": "CARD-" + (X_test.reset_index(drop=True).index + 1)
5         .astype(str).str.zfill(3),
6     "age": X_test["age"].to_numpy(),
7     "st_dep": X_test["st_depression"].to_numpy(),
8     "risk_pct": (proba * 100).round(1),
9     "flag": np.where(proba >= 0.35, "REVIEW", "routine"),
10 }).sort_values("risk_pct", ascending=False)
11 print(capstone_list.head(6).to_string(index=False))
12 capstone_list.to_excel("cardiac_worklist.xlsx", index=False)
13 joblib.dump(pipe, "heart_risk_model.joblib")
14 print("Exported cardiac_worklist.xlsx |",
15       (proba >= 0.35).sum(), "of", len(proba), "flagged")

```

شيفرة 10.16: الجزء و من مشروع التخرج: قائمة العمل والحفظ الدائم

case_id	age	st_dep	risk_pct	flag
CARD-064	56	4.2	96.7	REVIEW
CARD-185	77	3.9	96.7	REVIEW
CARD-067	48	4.6	96.0	REVIEW
CARD-093	48	5.6	95.1	REVIEW
CARD-076	71	1.3	93.9	REVIEW
CARD-209	71	0.0	93.9	REVIEW

Exported cardiac_worklist.xlsx | 122 of 225 flagged

الملخص التنفيذي إلى المدير العام (إجابة نموذجية، عشرة أسطر).

1. بنينا نموذج فرز استرشادياً يرتب إحالات أمراض القلب بحسب احتمال وجود مرض قلبي معتد به.
 2. المدخلات: بيانات الإحالة الروتينية — العمر، والجنس، ونوع ألم الصدر، وضغط الدم، والكوليسترول، ونتائج اختبار الجهد.
 3. الطريقة: مسار عمل مقفل من معايرة قياسية وانحدار لوجستي، دُرِّب على 675 مريضاً واختُبر على 225 لم يرهّم التدريب قط.
 4. قدرة التمييز: بلغت قيمة AUC بالتحقق المتقاطع 0.751 ± 0.032 ؛ وقيمة AUC على اختبار مستقل 0.791.
 5. نقطة التشغيل 0.35، المختارة للفرز: حساسية 78.1%، ونوعية 63.6%، وقيمة تنبؤية سلبية 79.6%.
 6. الأثر: يُعلّم 54% من الإحالات للمراجعة المعجلة؛ ويتركز المرضى الأعلى خطورة في رأس القائمة.
 7. الحدود: دُرِّب النموذج على فوج إحالة واحد؛ وأداؤه في التحري المجتمعي أو في المحافظات الأخرى غير مثبت.
 8. الحوكمة: النموذج يقترح، وطبيب القلب يتحقق ويوقع؛ ولا تغادر أي بيانات معرفّة للهوية المؤسسة.
 9. المراقبة: إعادة تقييم فصلية مقابل نتائج مؤكدة؛ وإعادة تدريب إذا هبطت قيمة AUC دون 0.74.
 10. التوصية: تجربة استرشادية خاضعة للإشراف لثلاثة أشهر في عيادة واحدة قبل أي نشر أوسع.
- التفسير السريري. يقول الملف التقييمي: مفيد، لا حاسم. فقيمة AUC القريبة من 0.79 ترتب المرضى ترتيباً أفضل بكثير من الصدفة وأفضل من أي متغير مفرد، لكن 21 حالة حقيقية ما تزال تحت خط 0.35 — لذا يجب أن يبقى الطابور الروتيني متحركاً بوتيرة آمنة، ويجب أن يُسمح دائماً للاشتباه السريري بأن يتجاوز موضع الطابور. ومعدل التعليم البالغ 54% هو الكلفة الصادقة لترجيح الحساسية في تيار إحالات عالي الانتشار؛ أما قدرة العيادة على استيعابه فقرار تشغيلي على التجربة الاسترشادية أن تجيب عنه — وهو بالضبط نوع الأسئلة التي ينبغي أن يحسمها مدير عام، لا دالة خسارة.

أسئلة النقاش الجماعي.

1. تعلّم عتبة 0.35 نسبة 54% من الإحالات. فإذا كانت طاقة العيادة تكفي لتعجيل 30% فقط من المرضى، فما سياسة العتبة التي ستدافع عنها أمام المدير العام — وماذا ستقول عن الحالات المفوتة الإضافية التي تستلزمها؟

2. دُرِبَ هذا النموذج على فوج إحالة واحد. أي الفئات الفرعية — محافظات، وأجناس، وفئات عمرية — ستدققها قبل السماح بالتجربة الاسترشادية في البصرة، وما إشارة المراقبة التي ستوجب إيقافها؟
3. يخالف طبيب قلب راية REVIEW ويخفّض تصنيف المريض؛ ثم يتبين لاحقاً أن المريض مصاب بالمرض. من المسؤول؟ وماذا ينبغي أن يسجل محضر الحادثة؟ وهل تتغير إجابتك لو كان النموذج قد قال routine؟

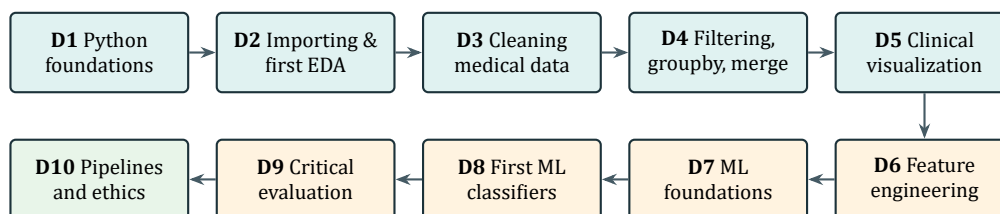
10.4 الواجب المنزلي والتقييم الذاتي

1. دراسة الاستقرار. أعد تشغيل التحقق المتقاطع للسكري بـ $cv=10$. أورد المتوسط $\pm$ الانحراف المعياري، واكتب جملتين عن سبب ثبات المتوسط تقريباً بينما يتغير الانحراف المعياري.
2. نطاقات الخطورة. وسّع قائمة عمل الصباح بتطبيق `pd.cut` على عمود الخطورة، بحدود تقسيم 0 و 0.2 و 0.5 و 1.0 وبالسوم `low` و `borderline` و `high`، وأورد العدد في كل نطاق بـ `value_counts()`. واقترح إجراءً سريرياً مختلفاً لكل نطاق.
3. رقيب الانجراف. باستخدام شيفرة موجبة فقط، قارن `new_patients.csv` بمجموعة التدريب: اطبع `X_train.mean().round(1)` و `new_patients.mean().round(1)` جنباً إلى جنب بـ `pd.concat` على `axis=1`. واكتب ثلاث جمل عن الفروق التي تثير قلقك وإعادة التقييم التي ستجديها.
4. إجراء إخفاء الهوية. في صفحة واحدة، صُغِ إجراء العمل القياسي للمديرية لتحضير مستخلص حقيقي من `patients_demographics.csv` للتحليل: أي الأعمدة معرفات مباشرة، وأياها أشباه معرفات، وما الذي يُحذف، وما الذي يُعمّم (مثلاً العمر إلى فئات من خمس سنوات)، ومن الذي يوافق على الإصدار.

اختبار التقييم الذاتي

1. سمّ نمطي الفشل اللذين يجعلهما Pipeline مستحيلين بنياً.
 2. لماذا يكون متوسط خمس قيم AUC من التحقق المتقاطع تقديراً أعدل من قيمة AUC لتقسيم 80/20 واحد؟
 3. ما الذي يُخزّن بالضبط داخل `diagnostic_model.joblib` — شيفرة أم بيانات أم شيء آخر؟
 4. نموذج دُرِبَ في معظمه على مرضى بغداد يؤدي أداءً أسوأ في نينوى. سمّ هذا الفشل واذكر علاجين له.
 5. عرّف الانحياز للأتمتة، واذكر جملة سياسة المديرية التي تخففه.
- إجابات الاختبار. (1) تسرب البيانات — معالجة مسبقة مُلاءمة على بيانات الاختبار أو المستقبل — ونسيان الخطوات، أي نشر النموذج من دون معالجته المسبقة الدقيقة. (2) لأن كل مريض يخدم في التحقق مرة واحدة بالضبط، فلا يعتمد التقدير على تقسيم واحد محظوظ أو منحوس، والانحراف المعياري يكفي الاستقرار. (3) كائن المقدر الملاءم: متوسطات المعايير وانحرافاتها المتعلمة، ومعاملات النموذج، وترتيب الخطوات — حالة متعلمة، لا بيانات مصدرة. (4) التحيز الخوارزمي الناجم عن جمهرة تدريب غير ممثلة؛ والعلاجات: تقييم مصنّف بحسب الفئات الفرعية، وجمع بيانات الفئات ناقصة التمثيل أو إعادة وزنها، والتصريح بالجمهرة المصادق عليها في بطاقة هوية النموذج. (5) الانحياز للأتمتة هو ميل الإنسان إلى قبول اقتراح آلة واثقة من دون نقد؛ والسياسة: النموذج يقترح، والطبيب السريري يتحقق، والطبيب وحده يوقع التقرير.

10.5 التخرج: عشرة أيام وعشر مهارات



شكل 10.3: رحلة الأيام العشرة: من أول جلسة بايثون إلى مسار عمل تشخيصي محوكم قابل للنشر.

جدول 10.2: قائمة تحقق الخريج: عشر مهارات في يد المديرية الآن.

اليوم	المهارة المكتسبة	مُتقنة
1	قراءة البيانات السريرية وهيكلتها والاستدلال بها في بايثون	✓
2	استيراد السجلات الطبية وإجراء أول تحليل استكشافي بـ pandas	✓
3	تدقيق ملفات المختبر الواقعية الفوضوية وإصلاحها	✓
4	الفلتر والتجميع والدمج في سجلات العمل التشغيلي	✓
5	إيصال النتائج بأشكال بمستوى النشر العلمي	✓
6	هندسة ميزات ذات معنى سريري وترميزها	✓
7	تقسيم البيانات طبقياً ومعايرتها من دون تسرب	✓
8	تدريب مصنفات تشخيصية خاضعة للإشراف وضبطها	✓
9	تقييم النماذج بعين الطبيب: الحساسية والنوعية وقيمتا التنبؤ ROC	✓
10	النشر والأتمتة والحوكمة — مسارات العمل والأخلاقيات والمراقبة	✓

كلمة مدير الدورة

أيها الزملاء: قبل عشرة أيام قيل لكم إن هذه الدورة ستنتهي ببناء كل واحد منكم مسار عمل تشخيصياً من الملف الخام إلى قائمة عمل موقعة، ولم تصدّقوا ذلك تماماً. اليوم بنيتم واحداً. تذكروا ثلاثة أشياء وأنتم تحملون هذا إلى مختبراتكم. أولاً، الانضباط أهم من الشيفرة — التدقيق قبل التحليل، والمصادقة قبل الادعاء، والتوقيع البشري بعد اقتراح الآلة. ثانياً، التزامكم تجاه المرضى لم يتغير حين تغيرت أدواتكم؛ بل ازداد مداه فحسب. ثالثاً، علّموه لمن بعدكم: كل واحد منكم يقف الآن حيث وقف في اليوم الأول، والمديرية تتحول حين تصبح هذه القاعة ثمانين قاعة. لقد كان شرفاً لي. انتهت الدورة؛ والعمل — وهو عمل نبيل — يبدأ صباح الغد، في الساعة 07:30، حين تعمل أول قائمة عمل حقيقية. — د. حسام صلاح مهدي

♣ القيمة لمديرية التشخيص — من عشرة متدربين إلى قدرة مؤسسية

اليوم يحوّل المهارة الفردية إلى بنية تحتية. فمسار عمل محفوظ دائماً ومرقّم الإصدار يعني أن المديرية تمتلك طريقة، لا دفتر ملاحظات شخص ما: أي محلل مدرب يستطيع تقييم ملف الصباح، والنتائج قابلة للاستنساخ حتى المنزل العشرية، وكل خطوة قابلة للتدقيق. وقوائم العمل المرتبة تحوّل مخرجات النموذج مباشرة إلى وقت سريري موثّر — فتصل مراجعة الكبار إلى المرضى الأعلى خطورة أولاً، كل يوم، من دون أن يعيد أحد اشتقاق أي شيء. وإطار الحوكمة الذي بُني اليوم — إجراءات إخفاء الهوية، وتدقيقات التحيز، وتوقعيات الإنسان في حلقة القرار، ومراجعة الانجراف الفصلية — هو بالضبط ما سيتيح للوزارة اعتماد هذه الأدوات والدفاع عنها وتوسيعها. والعائد الدائم هو الفوج المتخرج نفسه: عشر مهارات، مدرجة في الجدول ، صارت الآن مقيمة في المديرية وجاهزة لأن نُعلّم لمن بعدهم.

خلاصات اليوم

- يلحم Pipeline المعالجة المسبقة والنموذج في مقدّر واحد: fit واحدة و predict واحدة، ويصبح التسرب أو نسيان المعايير مستحيلاً بنوياً.
- أورد الأداء متوسطاً $\pm$ انحرافاً معيارياً بتحقيق متقاطع خماسي — فالتقسيم الواحد سحبة واحدة من يانصيب، لا مصادقة.
- يحوّل joblib.dump / joblib.load مسار عمل ملاءماً إلى أصل مؤسسي مرقّم الإصدار: درّب مرة، وقمّ كل صباح، ولا تحرر في المكان أبداً.
- التقييم الدفعي ثلاثة أفعال: حمّل النموذج، واقراء ملف CSV، ونقّد predict_proba؛ والترتيب مع to_excel يحوّلان الاحتمالات إلى قائمة عمل جاهزة للعيادة.
- الأخلاقيات جزء من الطريقة: أخفّ الهوية قبل التحليل، واعرف الأطر القانونية، ودقق التمثيل، وتوقع الانجراف.
- النموذج يقترح، والطبيب يتحقق، والطبيب يوقع — إبقاء الإنسان في حلقة القرار سياسة مديريةية، لا تفضيل شخصي.
- عتبة التشغيل قرار سريري بشأن كلف غير متناظرة: دافع عنها بالملف التقييمي، وراجعها ببيانات المراقبة.
- الحوسبة الكمومية والنماذج اللغوية الكبيرة في الأفق — قابلوهما بالانضباط نفسه الذي تعلمتموه هنا: صادق أولاً، وحوكم دائماً.

لا يوجد فصل لليوم الحادي عشر — فدرس الغد تكتبه أنت، عند طاولة عملك، في أول صباح يعمل فيه مسار عملك على بيانات المديرية الحقيقية.

الملاحق A

مرجع سريع لمكتبي NumPy وPandas

يلخص هذا الملحق كل أدوات مكتبي Pandas وNumPy التي دُرست في الأيام 6-1 في خمسة جداول منظّمة بحسب المهمة. أبقى هذا الملحق مفتوحاً إلى جانب دفترتك البرمجي: فأسرّع طريقك إلى الطلاقة أن تتوقف عن الحفظ وتبدأ بالبحث عند الحاجة. يعرض كل صف الطريقة، وما تفعله، واستخداماً سريريّاً بسطر واحد مستمداً من مجموعات بيانات الدورة.

نموذج ذهني واحد لكل شيء

كل عملية في هذه الدورة جملة من القواعد اللغوية نفسها: خذ جدولاً من المرضى، واطرح سؤالاً موجّهاً على عمود أو أكثر، ثم نقذ الإجراء على كل الصفوف المطابقة دفعة واحدة. التحميل يبني الجدول، والأقنعة المنطقية (Boolean Masks) تطرح السؤال، والتنظيف يصلح الإجابات، والتجميع يلخصها، وهندسة الميزات تشتق منها متغيرات سريرية جديدة — لا صفّاً واحداً في كل مرة أبداً، بل الفوج كله في تعبير واحد دائماً.

A.1 تحميل البيانات وفحصها

الطريقة	ما تفعله	استخدام سريري بسطر واحد
<code>pd.read_csv("f.csv")</code>	تحميل ملف CSV مُصدّر إلى إطار بيانات (DataFrame)	<code>pd.read_csv("cbc_results.csv")</code> عمل صورة الدم الكاملة (CBC)
<code>pd.read_excel("f.xlsx")</code>	تحميل مصنّف Excel (عبر openpyxl)	قراءة التصدير الشهري لنظام معلومات المختبر (LIS) المُرسَل كجدول بيانات
<code>df.head(5)</code>	عرض الصفوف الأولى	معاينة أول خمسة مرضى مباشرة بعد التحميل
<code>df.info()</code>	أسماء الأعمدة وأنواعها وعدد القيم غير الفارغة	معرفة أي التحاليل فيها نتائج مفقودة بنظرة واحدة
<code>df.describe()</code>	إحصاءات ملخّصة للأعمدة الرقمية	المتوسط والأرباع والقيم القصوى لكل تحليل في استدعاء واحد
<code>df.shape</code>	الزوج (rows, columns)	التأكد من أن السجل يحتوي فعلاً على 1200 مريض
<code>df.dtypes</code>	نوع بيانات كل عمود	اكتشاف الغلوكوز المخزّن كنص قبل أن تفشل العمليات الحسابية بصمت
<code>df["col"].value_counts()</code>	تكرار كل فئة	عدد نتائج RT-PCR الإيجابية مقابل السلبية
<code>df.sample(10)</code>	عينة عشوائية من الصفوف	فحص عشوائي لعشرة سجلات للتحقق من معقوليتها

جدول A.1: التحميل والفحص الأول (يُدْرَس في اليومين 1-2).

A.2 الاختيار والفلتر

الطريقة	ما تفعله	استخدام سريري بسطر واحد
<code>df["col"]</code>	اختيار عمود واحد (سلسلة Series)	سحب عمود الهيموغلوبين لرسم مدرج تكراري
<code>df[["a", "b"]]</code>	اختيار عدة أعمدة	الإبقاء على <code>patient_id</code> و <code>glucose_mgdl</code> فقط لإعداد قائمة إحالة
<code>df.loc[mask, cols]</code>	الصفوف بالشرط والأعمدة بالاسم	قيم الجلوكوز لمرضى العناية المركزة فقط
<code>df.iloc[0:5, 0:3]</code>	الصفوف والأعمدة بالموضع	أول خمسة صفوف وأول ثلاثة أعمدة عندما يكون مخطط الجدول مجهولاً
<code>df[df["wbc"] > 11]</code>	فلتر بالقياس المنطقي	عزل كل حالات ارتفاع كريات الدم البيضاء في الفوج
<code>~ &</code>	عمليات NOT / OR / AND على الأقنعة (ضع كل طرف بين قوسين)	مصابة بفقر الدم وأنثى؛ مدخن أو معاد إدخاله
<code>s.isin(["Basra", "Nineveh"])</code>	اختبار العضوية	المرضى من المحافظات الجنوبية أو الشمالية
<code>s.between(12, 15)</code>	اختبار مجال مغلق	هيموغلوبين ضمن المجال المرجعي للإنثى
<code>s.str.contains("Regional")</code>	مطابقة جزئية للنص	كل العينات المعالجة في المختبرات الإقليمية
<code>df.sort_values("crea")</code> <code>df.nlargest(10, "crea")</code>	ترتيب الصفوف؛ أخذ أعلى n	أعلى عشر قيم كرياتينين لمراجعة كلوية عاجلة

جدول A.2: الاختيار والفلتر المنطقية (تُدْرَس في اليومين 2 و4).

✖ خطأ شائع — الخطأ الكلاسيكي في الأقنعة

أولاً، كلمتا بايثون `or/and` لا تعملان على السلاسل (Series) — استخدم `&` و `|`، وضع كل مقارنة بين قوسين الخاصين: `df[(df["age"] > 60) & (df["smoker"] == "Yes")]`. ثانياً، علامة `=` المفردة تُسند بينما `==` تقارن: داخل الأقواس المعقوفة يكون `df["sex"] = "M"` خطأً نحوياً (SyntaxError) (بل إن بايثون تقترح عليك `==`): أما كسطر مستقل فهو يستبدل بالعمود بأكمله قيمة واحدة بصمت — فاعرف أي الخطأين ارتكبت.

A.3 التنظيف

الطريقة	ما تفعله	استخدام سريري بسطر واحد
<code>df.isna().sum()</code>	عدّ القيم المفقودة في كل عمود	تحديد حجم لوحات CBC غير المكتملة قبل أي تحليل
<code>df.dropna(subset=["id"])</code>	حذف الصفوف الفاقدة للحقول الأساسية	إزالة السجلات التي يتعذر ربطها بمرضى
<code>s.fillna(s.median())</code>	تعويض القيم المفقودة	استبدال وسيط الفوج بنتائج الجلوكوز الغائبة
<code>df.drop_duplicates()</code>	إزالة الصفوف المكررة تكراراً تاماً	حذف التصديرات المختبرية المدخلة مرتين
<code>s.replace({"male": "M"})</code>	إعادة ترميز القيم عبر قاموس	توحيد التهجئات الخمس لكلمة <code>male</code> في رمز واحد
<code>s.str.upper()</code> , <code>s.str.strip()</code>	إزالة الفراغات؛ توحيد حالة الأحرف	توحيد ترميز الجنس قبل أي <code>groupby</code>
<code>df.astype({"age": int})</code>	تحويل أنواع الأعمدة	جعل عمود العمر النصي قابلاً للحساب
<code>pd.to_numeric(s, errors="coerce")</code>	فرض التحويل الرقمي؛ الإدخالات التالفة تصبح NaN	تحييد إدخالات مثل N/A في تصدير مختبري
<code>pd.to_datetime(s)</code>	تحليل سلاسل التواريخ	جعل <code>admission_date</code> صالحاً لرسم منحنيات الوباء

جدول A.3: تنظيف تصدير تالف (يُدْرَس في اليوم 3 على `messy_lab_data.csv`).

A.4 التجميع والدمج

الطريقة	ما تفعله	استخدام سريري بسطر واحد
<code>df.groupby("dep")["los"].mean()</code>	إحصاء واحد لكل مجموعة	متوسط مدة الإقامة لكل قسم
<code>.agg(["mean", "std", "count"])</code>	عدة إحصاءات دفعة واحدة	ملف غلوكوز كامل لكل محافظة
<code>df.groupby("dep").size()</code>	عدد الصفوف في كل مجموعة	عدد الإدخالات لكل قسم لتقرير مؤشرات الأداء (KPI) الأسبوعي
<code>pd.crosstab(a, b)</code>	جدول توافقي لمتغيرين فئويين	جدول 2×2 بين التدخين وإعادة الإدخال
<code>pd.merge(a, b, on="patient_id")</code>	ربط بأسلوب SQL على مفتاح مشترك	إلحاق نتائج الكيمياء بالبيانات الديموغرافية
<code>pd.concat([a, b])</code>	تكديس الجداول عمودياً	دمج تصديرات RT-PCR الشهرية من مختبرين
<code>df.pivot_table(values, index, columns)</code>	ملخص مجمّع على شكل شبكة	متوسط HbA1c بحسب القسم والجنس في عرض واحد

جدول A.4: التجميع وربط السجلات (يُدرّس في اليومين 4-5).

A.5 هندسة الميزات

الطريقة	ما تفعله	استخدام سريري بسطر واحد
<code>df["a"] / df["b"]</code>	حساب موجّه بين الأعمدة	نسبة العدلات إلى اللمفاويات لجميع المرضى البالغ عددهم 1200 دفعة واحدة
<code>df["new"] = expr</code>	تخزين عمود مشتق	إضافة درجة الخطورة المحسوبة إلى السجل بشكل دائم
<code>(s >= 126).astype(int)</code>	تحويل شرط منطقي إلى راية 0/1	راية سكري من الغلوكوز الصائم في سطر واحد
<code>np.where(cond, a, b)</code>	قاعدة موجّهة ثنائية الفرع	تصنيف WBC إلى High مقابل Normal دون أي <code>if</code>
<code>s.map({"Y": 1, "N": 0})</code>	بحث قاموسي لكل عنصر	تحويل ترميز الذبحة الجهدية إلى قيم رقمية
<code>pd.cut(s, bins, labels=...)</code>	تحويل القيم الرقمية إلى نطاقات سريرية	HbA1c إلى طبيعي / ما قبل السكري / سكري
<code>pd.qcut(s, 4)</code>	تقسيم طبقي كمي (أرباع)	تقسيم المرضى إلى أرباع كولسترول لأغراض التدقيق
<code>pd.get_dummies(df, columns=...)</code>	ترميز أحادي للفئات	جعل نوع ألم الصدر مستساغاً للنموذج

جدول A.5: اشتقاق متغيرات سريرية جاهزة للنمذجة (تُدرّس في اليوم 6).

الملاحق B

مرجع سريع لمكتبة scikit-learn

أعظم ما تقدّمه مكتبة scikit-learn للأطباء السريريين هو التوحيد: كل نموذج، من الانحدار اللوجستي (Logistic Regression) إلى الغابة العشوائية (Random Forest)، يُدار بالأفعال الأربعة نفسها. تعلّم هذه القواعد مرة واحدة، وكل مقبّر مستقبلي — بما في ذلك ما لم يُدرّس في هذه الدورة — لن يكلفك شيئاً جديداً.

قواعد لغة المقدّرات

`model.fit(X_train, y_train)` يتعلم من مرضى موسومين؛ `model.predict(X_new)` يعيد فئة صارمة (0/1) للمرضى الجدد؛ `model.predict_proba(X_new)[:, 1]` يعيد احتمال المرض، وهو ما تحتاجه عتبات القرار التشخيصي؛ `model.score(X_test, y_test)` يبلغ عن الدقة. الميزات X دائماً جدول من أعمدة رقمية؛ والهدف y عمود واحد من النتائج. لا شيء آخر يتغير بين الخوارزميات.

الأداة	الغرض والملاحظة السريرية
<code>train_test_split(X, y, test_size=0.25, stratify=y, random_state=42)</code>	حجز صادق واحد لمرضى غير مرئيين؛ <code>stratify=y</code> يحافظ على معدل انتشار المرض في النصفين؛ <code>random_state</code> يجعل التقسيم قابلاً لإعادة الإنتاج للفوج كله
<code>StandardScaler()</code>	يعيد معايرة كل ميزة إلى متوسط 0 وانحراف معياري 1 حتى لا يطغى الكوليسترول (مئات الملغ/دل) على العمر (عشرات)؛ <code>fit_transform</code> على بيانات التدريب، و <code>transform</code> فقط على بيانات الاختبار
<code>LogisticRegression(max_iter=1000)</code>	نموذج خطي مُنتج لاحتمالات بمعاملات قابلة للتفسير؛ المصنّف التشخيصي من الخط الأول
<code>RandomForestClassifier(n_estimators=200)</code>	حشد من أشجار القرار؛ يلتقط التأثيرات اللاخطية ويبلغ عن أهميات الميزات لفحوص الواجهة السريرية
<code>SVC(probability=True)</code>	مصنّف قائم على الهامش؛ قوي على البيانات المعايرة، و <code>probability=True</code> مطلوب لتحليل ROC
<code>Pipeline([("scaler", ...), ("clf", ...)])</code>	يلحم المعالجة المسبقة والنموذج في كائن واحد بعملية <code>predict/fit</code> واحدة؛ يمنع بنويواً تسرب مجموعة الاختبار
<code>joblib.dump(model, "m.joblib")</code> <code>joblib.load(...)</code>	حفظ مسار عمل مدرب وإعادة تحميله، لكي تعيد عيادة الغد استخدام نموذج اليوم المُتحقّق منه دون إعادة تدريب

جدول B.1: الواجهة البرمجية الأساسية للنمذجة (تدرّس في الأيام 7-8 و10).

دالة المقياس	ما تخبر به الطبيب السريري
<code>confusion_matrix(y_test, y_pred)</code>	تعدادات الجدول 2×2 (الصفوف = الحقيقة، والأعمدة = التنبؤ) التي يُشتق منها كل مقياس في الملحق C
<code>recall_score(y_test, y_pred)</code>	الحساسية (Sensitivity): من بين المصابين فعلاً، كم واحداً التقطنا؟
<code>precision_score(y_test, y_pred)</code>	القيمة التنبؤية الإيجابية (PPV): من بين من أشرنا عليهم، كم واحداً مصاب فعلاً؟
<code>f1_score(y_test, y_pred)</code>	المتوسط التوافقي للحساسية وPPV: رقم واحد عندما يهيم الاثنان معاً
<code>accuracy_score(y_test, y_pred)</code>	النسبة الإجمالية للتصنيف الصحيح؛ مضللة عند اختلال توازن الفئات — اذكرها أخيراً
<code>classification_report(y_test, y_pred)</code>	كل ما سبق، لكل فئة، في كتلة واحدة قابلة للطباعة
<code>roc_auc_score(y_test, y_prob)</code>	قدرة تمييز مستقلة عن العتبة؛ أعطها احتمالات، لا تسميات صارمة 0/1 أبداً
<code>roc_curve(y_test, y_prob)</code>	المصفوفات (FPR, TPR, threshold) الكامنة وراء رسم ROC والعتبة المثلى بمؤشر يودن (Youden)

جدول B.2: دوال التقييم (تُدْرَس في اليوم 9).

✖ خطأ شائع — التسرب الوحيد الذي يبطل كل شيء

ملاءمة المعايير — أو أي تعويض للقيم المفقودة — على مجموعة البيانات الكاملة قبل التقسيم تسمح لإحصاءات مجموعة الاختبار بتلويث التدريب. قيمة AUC الناتجة خيالٌ لن يصمد أمام مرضى حقيقيين. وُجد Pipeline تحديداً لجعل هذا الخطأ مستحيلاً: فهو يعيد ملاءمة المعالجة المسبقة داخل طية التدريب وحدها.

B.1 سكربت مصغّر متكامل من البداية إلى النهاية

السكربت الكامل التالي — الخالي من الحلقات التكرارية ومن if — يأخذ سجل القلب من ملف CSV خام إلى نموذج تشخيصي محفوظ ومُقيّم. إنه الهيكل العظمي لكل نموذج بُني في هذه الدورة.

```

1 import pandas as pd
2 from sklearn.model_selection import train_test_split
3 from sklearn.preprocessing import StandardScaler
4 from sklearn.linear_model import LogisticRegression
5 from sklearn.pipeline import Pipeline
6 from sklearn.metrics import (accuracy_score, recall_score,
7                               roc_auc_score, confusion_matrix)
8 import joblib
9
10 df = pd.read_csv("heart_risk.csv")
11 df["sex_male"] = (df["sex"] == "M").astype(int)
12 df["angina"] = (df["exercise_angina"] == "Y").astype(int)
13 features = ["age", "sex_male", "resting_bp", "cholesterol",
14             "max_hr", "st_depression", "angina"]
15 X = df[features]
16 y = df["target"]
17
18 X_train, X_test, y_train, y_test = train_test_split(
19     X, y, test_size=0.25, stratify=y, random_state=42)
20
21 model = Pipeline([("scaler", StandardScaler()),
22                  ("clf", LogisticRegression(max_iter=1000))])
23 model.fit(X_train, y_train)
24
25 y_pred = model.predict(X_test)
26 y_prob = model.predict_proba(X_test)[: , 1]
```

```

27
28 print("Test patients :", len(y_test))
29 print("Sensitivity  :", round(recall_score(y_test, y_pred), 3))
30 print("Accuracy     :", round(accuracy_score(y_test, y_pred), 3))
31 print("ROC AUC      :", round(roc_auc_score(y_test, y_prob), 3))
32 print("Confusion matrix (rows=truth, cols=prediction):")
33 print(confusion_matrix(y_test, y_pred))
34 joblib.dump(model, "heart_model.joblib")
35 print("Saved: heart_model.joblib")

```

شيفرة B.1: الدورة كلها في شاشة واحدة: تحميل، هندسة، تقسيم، معايرة، تدريب، تقييم، حفظ.

```

Test patients : 225
Sensitivity    : 0.594
Accuracy      : 0.724
ROC AUC       : 0.794
Confusion matrix (rows=truth, cols=prediction):
[[106  23]
 [ 39  57]]
Saved: heart_model.joblib

```

لاحظ القراءة السريرية: قيمة AUC البالغة 0.794 تُظهر قدرة تمييزية حقيقية، ومع ذلك يفوّت النموذج عند العتبة الافتراضية 0.50 تسعة وثلاثين مريضاً من أصل 96 مريضاً مصاباً (حساسية 0.594). فالتمييز واختيار العتبة قراران منفصلان — وهذا بالضبط سبب وجود الملحق C.

الملاحق C

دليل المقاييس التشخيصية

كل سؤال عن الأداء التشخيصي يؤول إلى مصفوفة الارتباك (Confusion Matrix) 2×2 التي تقارن حكم الاختبار (أو النموذج) بالحقبة التي أرساها معيار ذهبي.

تنبؤ سلبى	تنبؤ إيجابى	
سلبى كاذب (FN)	إيجابى حقيقى (TP)	المرض موجود
سلبى حقيقى (TN)	إيجابى كاذب (FP)	المرض غائب

جدول C.1: مصفوفة الارتباك 2×2 . حيث $N = TP + FN + FP + TN$.

C.1 التعاريف والصيغ

Sensitivity (recall) = $\frac{TP}{TP + FN}$	diseased patients correctly detected
Specificity = $\frac{TN}{TN + FP}$	healthy patients correctly cleared
PPV (precision) = $\frac{TP}{TP + FP}$	probability of disease given a positive
NPV = $\frac{TN}{TN + FN}$	probability of health given a negative
Accuracy = $\frac{TP + TN}{N}$	overall fraction correct
Prevalence = $\frac{TP + FN}{N}$	true disease burden in the tested group
Youden $J = \text{Sens} + \text{Spec} - 1$	net gain over a coin flip; picks cut-offs
$LR^+ = \frac{\text{Sens}}{1 - \text{Spec}}$	how much a positive raises disease odds
$LR^- = \frac{1 - \text{Sens}}{\text{Spec}}$	how much a negative lowers disease odds

منحنى ROC والمساحة تحت المنحنى (AUC)

كل عتبة احتمال تعطي زوجاً واحداً (حساسية، نوعية). رسم الحساسية مقابل $1 - \text{Specificity}$ عبر كل العتبات يخط منحنى خاصية تشغيل المستقبل (ROC): القائمة الكاملة لنقاط التشغيل التي يستطيع الاختبار تقديمها. والمساحة تحت هذا المنحنى (AUC) تساوي

احتمال أن يحصل مريض مصاب مختار عشوائياً على درجة أعلى من شخص سليم مختار عشوائياً. وكدليل تقريبي: 0.5 عديمة القيمة، و0.8-0.7 مقبولة، و0.9-0.8 ممتازة، وما فوق 0.9 استثنائية. الحساسية والنوعية خاصيتان للاختبار نفسه؛ أما القيمتان التنبؤيتان الإيجابية والسلبية فليستا كذلك — فهما تتحركان مع معدل الانتشار، كما يبين المثال المحلول أدناه.

C.2 مثال محلول بالكامل

يفحص برنامج تحريّ $N = 1000$ من السكان بحثاً عن مرض معدل انتشاره الحقيقي 5%: 50 مصاباً و950 سليماً. وكانت النتائج: $TP = 45$ و $TN = 860$ و $FP = 90$ و $FN = 5$

$$\text{Prevalence} = \frac{50}{1000} = 0.050 \text{ (5.0\%)}$$

$$\text{Sensitivity} = \frac{45}{45+5} = \frac{45}{50} = 0.900 \text{ (90.0\%)}$$

$$\text{Specificity} = \frac{860}{860+90} = \frac{860}{950} = 0.905 \text{ (90.5\%)}$$

$$\text{PPV} = \frac{45}{45+90} = \frac{45}{135} = 0.333 \text{ (33.3\%)}$$

$$\text{NPV} = \frac{860}{860+5} = \frac{860}{865} = 0.994 \text{ (99.4\%)}$$

$$\text{Accuracy} = \frac{45+860}{1000} = 0.905$$

$$J = 0.900 + 0.905 - 1 = 0.805$$

$$\text{LR}^+ = \frac{0.900}{1-0.905} = 9.5 \quad \text{LR}^- = \frac{1-0.900}{0.905} = 0.11$$

★ لؤلؤة سريرية — درس القيمة التنبؤية الإيجابية المنخفضة

يكشف هذا الاختبار 90% من المصابين ويبرئ 90% من الأصحاء — ومع ذلك اثنان من كل ثلاثة تقارير إيجابية إنذاران كاذبان (قيمة تنبؤية إيجابية 33.3%). الجاني هو معدل الانتشار: فعند 5% يولد السكان الأصحاء البالغ عددهم 950 تسعين إيجابية كاذبة تغمر الإيجابيات الحقيقية الخمس والأربعين. هذه هي الرياضيات الكامنة وراء كل سياسة اختبار تأكيدية: التحري الإيجابي عند انتشار منخفض إشارة إلى إعادة الاختبار، لا إلى العلاج. والاختبار نفسه إذا نُشر في عيادة عالية الانتشار (40% مثلاً) سيُظهر قيمة تنبؤية إيجابية أعلى بكثير مع حساسية ونوعية متطابقتين. لا تذكر أبداً القيمة التنبؤية الإيجابية دون ذكر معدل الانتشار.

C.3 المجالات المرجعية المختبرية للبالغين المستخدمة في هذه الدورة

المجال المرجعي للبالغين	الوحدات	تحليل صورة الدم (CBC)
4.0-11.0	$\times 10^9/L$	كريات الدم البيض (WBC)
4.0-5.2 إناث؛ 4.5-5.9 ذكور	$\times 10^{12}/L$	كريات الدم الحمر (RBC)
12-15 إناث؛ 13-17 ذكور	غ/دل	الهيموغلوبين (Hgb)
36-46 إناث؛ 40-50 ذكور	%	الهيماتوكريت (Hct)
150-450	$\times 10^9/L$	الصفائح الدموية (Plt)
40-70	WBC من %	العدلات
20-40	WBC من %	اللمفاويات

جدول C.2: المجالات المرجعية الدموية (cbc_results.csv).

تحليل الكيمياء السريرية	الوحدات	القيم المرجعية / قيم القرار
الغلوكوز الصائم	ملغ/دل	70-99؛ ما قبل السكري 100-125؛ السكري ≥ 126
السكر التراكمي HbA1c	%	≥ 5.7 ؛ ما قبل السكري 5.7-6.4؛ السكري ≥ 6.5
الكرياتينين	ملغ/دل	ذكور 0.7-1.3؛ إناث 0.6-1.1
اليوريا	ملغ/دل	15-40
ALT	وحدة/لتر	7-56
AST	وحدة/لتر	10-40
الكوليسترول الكلي	ملغ/دل	المرغوب < 200 ؛ المرتفع ≥ 240
الدهون الثلاثية	ملغ/دل	الطبيعي < 150
كوليسترول HDL	ملغ/دل	ذكور ≥ 40 ؛ إناث ≥ 50
كوليسترول LDL	ملغ/دل	الأمثل < 100 ؛ المرتفع ≥ 160
الصبوديوم	ملي مول/لتر	135-145
البوتاسيوم	ملي مول/لتر	3.5-5.1

جدول C.3: المجالات المرجعية للكيمياء السريرية وعتبات القرار التشخيصي (chemistry_panel.csv).

★ لؤلؤة سريرية — تفسير قيم Ct في RT-PCR

قيمة عتبة الدورة (Ct) هي دورة التضخيم التي يعبر عندها التآلق الفلوري خط الكشف؛ وهي مرتبطة عكسياً بالحمل الفيروسي. تكون النتيجة إيجابية عندما يتضخم جين هدف دون عتبة الفحص (40 دورة عادة). وكدليل عملي: $Ct < 25$ يشير إلى حمل فيروسي مرتفع، و30-25 متوسط، و35-30 منخفض، و40-35 حمل منخفض جداً قرب حد الكشف — ففكر في إعادة الاختبار. تضخيم جين واحد فقط (مثل N دون (ORF1ab يُبلغ عنه بأنه غير حاسم. وقيم Ct غير قابلة للمقارنة المباشرة بين المنصات أو المختبرات دون موافقة — وهو تنبيه مهم عند تحليل مصادر المختبرات الثلاثة في rtpcr_results.csv.

الملاحق D

مسرد المصطلحات

تعريف عملية من جملة واحدة، بلغة سريرية، لكل مصطلح يجب أن يمتلكه المتدرب بحلول اليوم العاشر.

جدول D.1: مسرد مصطلحات الدورة (مرتب أبجدياً بحسب المصطلح الإنكليزي).

المصطلح	التعريف
الدقة (Accuracy)	حصة جميع المرضى، مصابين وأصحاء على السواء، الذين يصنفهم الاختبار تصنيفاً صحيحاً؛ مضللة كلما كانت إحدى الفئتين نادرة.
المساحة تحت المنحنى (AUC)	المساحة تحت منحنى ROC: احتمال أن يحصل مريض مصاب مختار عشوائياً على درجة خطورة أعلى من شخص سليم مختار عشوائياً.
القناع المنطقي (Boolean Mask)	عمود من قيم True/False، قيمة لكل مريض، يُستخدم لاختيار السجلات التي تحقق معياراً سريرياً بدقة.
البث (Broadcasting)	قاعدة NumPy لتطبيق عملية واحدة على مصفوفة كاملة دفعة واحدة، مثل تحويل كل قيم الغلوكوز إلى ملي مول/لتر في تعبير واحد.
اختلال توازن الفئات (Class Imbalance)	مجموعة بيانات تكون فيها إحدى النتيجتين (المرض عادة) أندر بكثير من الأخرى، فتتضخم الدقة ويصبح التقييم القائم على الحساسية واجباً.
مصفوفة الارتباك (Confusion Matrix)	جدول 2×2 للإيجابيات والسلبيات الحقيقية والكاذبة الذي تُشتق منه كل المقاييس التشخيصية.
القيم المفصولة بفواصل (CSV)	صيغة جداول البيانات النصية البسيطة التي تصدر بها أنظمة معلومات المختبرات نتائجها.
تسرب البيانات (Data Leakage)	أي وضع تؤثر فيه معلومات مجموعة الاختبار على التدريب، فينتج تقديرات أداء متفائلة زائفة.
إطار البيانات (DataFrame)	البنية المركزية في Pandas: جدول موسوم صفوفه مرضى أو عينات وأعمدته متغيرات.
عتبة القرار (Decision Threshold)	عتبة الاحتمال (افتراضياً 0.50) التي يبلغ النموذج فوقها عن المرض؛ وتحريكها يقاوض الحساسية النوعية.
نوع البيانات (dtype)	نوع البيانات المخزن للعمود (عدد صحيح، عدد عشري، نص): العمليات الحسابية على القيم المختبرية تتطلب نوعاً رقمياً.
التحليل الاستكشافي للبيانات (EDA)	الفحص الأول المنهجي لمجموعة بيانات بالملخصات والرسوم قبل أي نمذجة.
المقدّر (Estimator)	أي كائن في scikit-learn يتعلم من البيانات عبر <code>.fit()</code> ؛ وكل المقدّرات تشترك في القواعد اللغوية نفسها.
الميزة (Feature)	متغير إدخال يُعرض على النموذج، مثل العمر أو HbA1c أو انخفاض قطعة ST.
هندسة الميزات (Feature Engineering)	إنشاء مدخلات جديدة ذات معنى سريري من الأعمدة الخام، مثل مؤشر كتلة الجسم من الوزن والطول، أو فئة عمرية من العمر.
التجميع (groupby)	عملية Pandas التي تقسم السجلات إلى مجموعات سريرية، وتطبق إحصاءاً على كل منها، ثم تعيد تجميع النتائج.

يتبع في الصفحة التالية

المصطلح	التعريف (تابع)
المعامل الفائق (Hyperparameter)	إعداد يختاره المحلل قبل التدريب، مثل عدد الأشجار في الغابة العشوائية.
التعويض (Imputation)	ملء القيم المختبرية المفقودة ببديل معقول، وهو في الغالب وسيط العمود.
المدى الربيعي (IQR)	المدى بين الربيعين ($Q_3 - Q_1$)؛ القيم الواقعة خارج الربيعين بأكثر من $1.5 \times IQR$ تُؤشّر كقيم شاذة إحصائياً.
التسمية أو الهدف (Label / Target)	عمود النتيجة الذي يتعلم النموذج الخاضع للإشراف التنبؤ به، مثل وجود السكري أو غيابه.
الانحدار اللوجستي (-Logistic Regression)	نموذج خطي ينتج احتمال مرض بين 0 و1؛ حصان العمل القابل للتفسير في التنبؤ التشخيصي.
الدمج (Merge)	ربط جدولين على مفتاح مشترك مثل patient_id، كإلحاق لوحات الكيمياء بالبيانات الديموغرافية.
القيمة المفقودة (NaN)	خلية فارغة؛ تنتشر عبر العمليات الحسابية ويجب عدّها أو تعويضها أو حذفها عن قصد.
الترميز الأحادي (One-hot Encoding)	تحويل فئة مثل المحافظة إلى أعمدة 0/1 منفصلة ليتمكن النموذج من استهلاكها.
القيمة الشاذة (Outlier)	قيمة بعيدة عن المجال المعقول، مثل غلوكوز 9000 ملغ/دل — خطأ في إدخال البيانات عادة، لا حالة مرضية.
فرط التخصيص (Overfitting)	نموذج يحفظ مرضى تدريبه عن ظهر قلب، فيحقق درجات باهرة عليهم ويفشل مع المرضى الجدد.
مسار العمل (Pipeline)	كائن scikit-learn يسلسل المعالجة المسبقة والنموذج في وحدة واحدة، فتُلاءم المعايير دائماً على بيانات التدريب وحدها.
القيمة التنبؤية الإيجابية (PPV)	من بين المرضى المُبلّغ عنهم إيجابيين، نسبة المصابين فعلاً؛ تنهار عندما يكون معدل الانتشار منخفضاً.
معدل الانتشار (Prevalence)	نسبة السكان المفحوصين المصابين فعلاً بالمرض؛ وهو ما يحرك القيمتين التنبؤيتين الإيجابية والسلبية.
الغابة العشوائية (Random Forest)	حشد من أشجار قرار كثيرة تُجمّع أصواتها، فيقل فرط التخصيص وتُلتقط التأثيرات اللاخطية.
البذرة العشوائية (random_state)	البذرة الثابتة التي تجعل التقسيم العشوائي أو النموذج قابلاً لإعادة الإنتاج، فيحصل كل متدرب على الأرقام نفسها.
منحنى خاصية تشغيل المستقبل (ROC)	رسم الحساسية مقابل 1 - Specificity عبر كل العتبات: كل نقطة تشغيل يستطيع الاختبار تقديمها.
الحساسية (/ Sensitivity Recall)	من بين المرضى المصابين فعلاً، النسبة التي يكشفها الاختبار؛ المقياس الذي يجب أن يكون عالياً لنفي المرض.
السلسلة (Series)	عمود واحد موسوم من إطار البيانات، مثل قيم الهيموغلوبين لجميع المرضى.
النوعية (Specificity)	من بين الأصحاء فعلاً، النسبة التي يبرّئها الاختبار بصورة صحيحة؛ النوعية العالية تحدّ من الإنذارات الكاذبة.
المعايرة القياسية (Standardization)	إعادة معايرة كل ميزة إلى متوسط 0 وانحراف معياري 1 حتى لا يهيمن الكوليسترول على العمر بمجرد ضخامة أرقامه.
التوزيع الطبقي (Stratification)	تقسيم البيانات بحيث تكون نسبة المرض متطابقة في مجموعتي التدريب والاختبار (stratify=y).
آلة المتجهات الداعمة (SVM)	مصنّف يبحث عن الحد الفاصل ذي هامش الأمان الأعرض بين الفئتين.
مجموعة الاختبار (Test Set)	المرضى المحجوزون خارج التدريب والمستخدمون مرة واحدة بالضبط للتقييم النهائي الصادق.
مجموعة التدريب (Training Set)	المرضى الذين يتعلم منهم النموذج؛ والأداء عليهم وحدهم ليس دليلاً يصح الاستشهاد به أبداً.
قصور التخصيص (Underfitting)	نموذج أبسط من أن يلتقط النمط السريري الحقيقي، فيؤدي أداءً ضعيفاً على بيانات التدريب والاختبار على السواء.
المعالجة الموجبة (Vectorization)	العمل على أعمدة كاملة دفعة واحدة بدلاً من المرور على الصفوف؛ الانضباط الجوهرية لهذه الدورة.

يتبع في الصفحة التالية

المصطلح	التعريف (تابع)
مؤشر يودن (J)	الحساسية + النوعية - 1؛ والعتبة التي تعظم J ، خيار قياسي للعتبة المثلى.

د. حسام صلاح مهدي
 رئيس قسم الذكاء الاصطناعي — جامعة المصطفى
 Dr. Husam Salah Mahdi — Head of AI Department, Al-Mustafa University

الملاحق E

إعداد البيئة ومجموعات البيانات

E.1 البداية السريعة مع Colab Google (الخيار الموصى به)

لا يتطلب Colab أي تثبيت ولا أي صلاحيات مسؤول، ويمنح كل متدرب بيئة متطابقة مثبتة فيها كل مكتبات الدورة مسبقاً.

1. افتح <https://colab.research.google.com> في متصفح حديث وسجل الدخول بأي حساب Google.
2. اختر *New notebook* → *File*. يعمل الدفتر على جهاز سحابي مجاني — لا شيء يمس الحاسوب المحلي.
3. ارفع ملفات CSV السبعة الخاصة بالدورة: انقر أيقونة المجلد في الشريط الجانبي الأيسر ثم زر الرفع — أو شغل الخلية أدناه.
4. تحقق من نجاح الرفع بتحميل ملف واحد (القائمة البرمجية الثانية أدناه).
5. لاحظ أن أجهزة Colab تُعاد تهيئتها بعد الانقطاع الناجم عن الخمول: أعد رفع ملفات CSV في بداية كل جلسة، أو اربط Google Drive واحفظها هناك.

```
1 from google.colab import files
2 uploaded = files.upload() # select the seven CSV files
```

شيفرة E.1: رفع مجموعات بيانات الدورة داخل Colab.

```
1 import pandas as pd
2 pd.set_option("display.width", 100)
3 df = pd.read_csv("cbc_results.csv")
4 print("Loaded:", df.shape[0], "rows x", df.shape[1], "columns")
5 print(df[["patient_id", "wbc", "hgb", "plt"]].head(3))
```

شيفرة E.2: فحص البيئة في ثلاثين ثانية.

```
Loaded: 1200 rows x 9 columns
  patient_id  wbc  hgb  plt
0 MOH-00001   6.7  16.3  292
1 MOH-00002  12.7  13.6  170
2 MOH-00003   8.2  13.9  234
```

✖ خطأ شائع — بيانات المرضى الحقيقية والخدمات السحابية

Colab معتمد هنا فقط لمجموعات بيانات الدورة الاصطناعية. أما تصديرات بيانات مرضى الوزارة الفعلية فيجب ألا تُرفع أبداً إلى خدمات سحابية عامة؛ فالتحليلات الإنتاجية تعمل على التثبيت دون اتصال الموضح أدناه، داخل شبكة المديرية، وفق سياسة حماية البيانات في الوزارة.

E.2 التثبيت دون اتصال (محطات عمل المديرية)

للأجهزة التي لا تتوفر فيها شبكة إنترنت موثوقة، ثبتت **Anaconda** (<https://www.anaconda.com/download>) الذي يجمع بايثون و **Jupyter** والحزمة العلمية كلها في مثبت واحد. وبدلاً من ذلك، مع تثبيت بايثون عادي بإصدار 3.11+ فما فوق، أنشئ البيئة من ملف المتطلبات:

```
pip install -r requirements.txt
jupyter notebook
```

النسخة الدقيقة من ملف requirements.txt المرفقة بالدورة:

```
numpy>=2.0
pandas>=2.2
matplotlib>=3.9
seaborn>=0.13
scikit-learn>=1.5
scipy>=1.13
openpyxl>=3.1
jupyter
```

E.3 إعادة توليد مجموعات البيانات

تُنَجّ ملفات CSV السبعة كلها بسكريت واحد ذي بذرة عشوائية ثابتة، لذا تعطي إعادة التوليد الملفات المتطابقة نفسها دائماً — ويبقى كل رقم في هذا الكتاب قابلاً لإعادة الإنتاج على أي جهاز:

```
cd datasets
python generate_datasets.py
```

الأيام الرئيسية	الغرض	الصفوف	الملف
2, 4-5	سجل الإدخالات: البيانات الديموغرافية والقسم ومدة الإقامة وإعادة الإدخال	1200	patients_demographics.csv
2, 4-5	لوحة صورة الدم الكاملة مفهرسة بالمفتاح patient_id	1200	cbc_results.csv
4-6	الغلوكوز وHbA1c وتحاليل الكلى والكبد والدهون	1200	chemistry_panel.csv
4-5	نتائج جزيئية بقيم Ct من ثلاثة مختبرات	800	rtPCR_results.csv
3	تصدير مُفسد عمداً: قيم مفقودة وقيم شاذة وتكرارات	600	messy_lab_data.csv
2, 6-9	سجل سكري على نمط بيما لأولى نماذج تعلم الآلة	768	diabetes_dataset.csv
6, 8-10	سجل الخطورة القلبية للمصنفات والتقييم والمشروع الختامي	900	heart_risk.csv

جدول E.1: مجموعات بيانات الدورة السبع ومواضع الحمل التعليقي لكل منها.

♣ القيمة لمديرية التشخيص — أصل تدريبي مكتفٍ بذاته

كل شيء في هذا الدليل — البيئة ومجموعات البيانات والشفيرة — يُعاد توليده من ملفات تملكها المديرية. يمكن تدريب أفواج جديدة، وإعادة تدقيق كل رقم منشور، دون أي اعتماد خارجي يتجاوز متصفحاً مجانياً أو تثبيتاً واحداً دون اتصال بالإنترنت.