سؤال ۱ – تشخیص نوع مسئله و انتخاب مدل مناسب

الف) برای هر یک از سناریوهای زیر مشخص کنید:

1. مسئله از نوع **Regression** است یا **Classification** ؟
2. دلیل خود را در یک یا دو جمله توضیح دهید.

سناریوها:

الف) پیش‌بینی قیمت یک خانه بر اساس متراژ، تعداد اتاق و سن ساختمان.

ب) تشخیص اینکه یک ایمیل Spam است یا خیر.

ج) پیش‌بینی میزان مصرف برق یک ساختمان در روز آینده.

د) تشخیص گونه یک گل بر اساس طول و عرض گلبرگ.

هـ) پیش‌بینی مدت زمان رسیدن تاکسی به مقصد.

و) پیش‌بینی اینکه یک مشتری وام خود را بازپرداخت خواهد کرد یا خیر.

ب) برای هر یک از سناریوهای بالا، مناسب‌ترین الگوریتم از بین گزینه‌های زیر را انتخاب کنید.

- Linear Regression
- Logistic Regression
- KNN Regressor
- KNN Classifier
- Decision Tree Classifier
- K-Means

در صورت وجود بیش از یک انتخاب مناسب، دلیل انتخاب خود را بیان کنید.

ج) برای هر سناریو، مشخص کنید کدام معیار ارزیابی مناسب‌تر است.

از میان معیارهای زیر انتخاب کنید.

- Accuracy
- Confusion Matrix
- MSE
- MAE
- R^2

در صورت نیاز می‌توانید بیش از یک معیار انتخاب کنید.

د) جدول زیر را کامل کنید.

الگوریتم	نوع یادگیری	نوع خروجی
Linear Regression	؟	؟
Logistic Regression	؟	؟
KNN Regressor	؟	؟
KNN Classifier	؟	؟
Decision Tree	؟	؟
K-Means	؟	؟

هـ) به پرسش‌های زیر پاسخ کوتاه دهید.

1. چرا Accuracy معیار مناسبی برای Linear Regression نیست؟
2. چرا R^2 برای طبقه‌بندی استفاده نمی‌شود؟
3. مهم‌ترین تفاوت خروجی Regression و Classification چیست؟

سؤال ۲ — آموزش مدل Regression

در این سؤال قصد داریم یک مدل **Linear Regression** برای پیش‌بینی قیمت خانه آموزش دهیم.

```
import pandas as pd

house_data = pd.DataFrame({

"Area": [ 60,65,70,75,80,85,90,95,100,105,110,115,120,125,
130,135,140,145,150,155,160,165,170,175,180,185, 190,195,200,205,210,215,220,225,230,235,240,245,
250,255,260,265,270,275,280,285,290,295 ],

"Bedrooms": [ 1,1,2,2,2,2,2,3,3,3,3,3,3,3, 4,4,4,4,4,4,4,4,4,5,5,5, 5,5,5,5,5,5,6,6,6,6,6,6, 6,6,7,7,7,7,7,7,7,7 ],

"Age": [ 30,28,26,24,22,20,18,17,16,15,14,13,12,11, 10,9,8,8,7,7,6,6,5,5,4,4, 3,3,2,2,2,1,1,1,1,1,1,1,
1,1,1,1,1,1,1,1,1,1 ], ”

Price": [ 130,145,158,170,185,198,210,228,245,258,270,285,300,315,
328,340,355,365,375,392,400,410,425,438,450,462, 475,488,500,515,528,540,555,568,580,595,610,625,
640,652,665,680,695,708,720,735,750,765 ]

 })
```

هدف، پیش‌بینی ستون **Price** است.

الف) ابتدا داده‌ها را بررسی کنید.

1. پنج سطر اول را نمایش دهید.
2. تعداد نمونه‌ها و تعداد ویژگی‌ها را مشخص کنید.
3. ویژگی‌ها (Features) و متغیر هدف (Target) را تعیین کنید.

ب) داده‌ها را به نسبت 80٪ آموزش و 20٪ آزمون تقسیم کنید.

از random_state=42 استفاده کنید.

ج) یک مدل Linear Regression آموزش دهید.

د) برای داده‌های آزمون، قیمت خانه‌ها را پیش‌بینی کنید.

هـ) جدولی مشابه جدول زیر ایجاد کنید.

Actual Price	Predicted Price
...	...

و) برای یک خانه با مشخصات زیر قیمت را پیش‌بینی کنید.

- Area = 140
- Bedrooms = 4
- Age = 5

ز) به سؤال‌های زیر پاسخ دهید.

نکته: در این قسمت فقط استدلال کنید و نیازی به اجرای مجدد مدل نیست.

1. آیا همه ویژگی‌ها برای پیش‌بینی قیمت مفید به نظر می‌رسند؟
2. انتظار دارید کدام ویژگی بیشترین تأثیر را داشته باشد؟ چرا؟
3. اگر ویژگی **Age** را حذف کنیم، آیا انتظار دارید دقت مدل افزایش یابد یا کاهش؟ دلیل خود را بیان کنید.

سؤال ۳ — ارزیابی و تحلیل مدل Linear Regression

در این سؤال از مدل آموزش‌داده‌شده در سؤال قبل استفاده کنید.

الف) معیارهای زیر را برای داده‌های آزمون محاسبه کنید.

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- R^2 Score

ب) نمودار Actual vs Predicted را رسم کنید.

سپس به پرسش‌های زیر پاسخ دهید.

1. آیا نقاط نزدیک خط ایده‌آل قرار گرفته‌اند؟
2. اگر چند نقطه فاصله زیادی از خط داشته باشند، چه برداشتی می‌توان داشت؟
3. اگر تمام نقاط دقیقاً روی خط باشند، مقدار R^2 تقریباً چند خواهد بود؟

ج) فرض کنید خروجی مدل به صورت زیر باشد.

Actual	Predicted
210	205
270	280
340	335
410	430

بدون استفاده از کد، مشخص کنید:

1. کدام نمونه بیشترین خطا را دارد؟
2. کدام نمونه کمترین خطا را دارد؟
3. آیا مدل در مجموع عملکرد قابل قبولی دارد؟ دلیل خود را بیان کنید.

د) یکی نفر ادعا می‌کند:

«مدل من ۹۵٪ Accuracy دارد، بنابراین مدل Regression بسیار خوبی است.»

نظر ایشان را نقد کنید.

در پاسخ خود به موارد زیر اشاره کنید.

- چرا Accuracy برای Regression مناسب نیست؟
- به جای آن از چه معیارهایی باید استفاده کرد؟

- هر معیار چه اطلاعاتی درباره عملکرد مدل ارائه می‌دهد؟

هـ) فرض کنید مقدار $R^2 = 0.94$ به دست آمده است.

به پرسش‌های زیر پاسخ دهید.

1. این مقدار چه برداشتی از کیفیت مدل به ما می‌دهد؟
2. آیا صرفاً با داشتن R^2 بالا می‌توان نتیجه گرفت که مدل همیشه خوب است؟ چرا؟
3. چه عواملی ممکن است باعث شوند مدلی با R^2 بالا در عمل عملکرد مطلوبی نداشته باشد؟

سؤال ۴ — آموزش مدل طبقه‌بندی با Logistic Regression

موضوع: تشخیص تراکنش‌های بانکی مشکوک (Fraud Detection)

Dataset:

```
import pandas as pd

transaction_data = pd.DataFrame({

 "Amount": [ 15,20,22,25,28,30,35,40,45,50,55,60,65,70,75,
80,90,100,110,120,130,140,150,160,180,200,220,250,280,300,
320,350,380,400,450,500,550,600,650,700,750,800,850,900,950,
1000,1050,1100,1150,1200,1250,1300,1350,1400,1450,1500,1600,1700,1800 ],

"TransactionCount": [ 1,1,1,1,1,2,2,2,2,2,2,2,3,3,3, 3,3,3,3,4,4,4,4,4,5,5,5,5,6,6,
6,6,6,7,7,8,8,8,8,9,9,9,9,10,10, 10,10,11,11,11,12,12,12,13,13,14,14,15,15 ],

"International": [ 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0, 0,0,1,0,1,0,1,1,0,1,1,0,1,1,1,
0,1,1,0,1,1,1,0,1,1,0,1,1,0,1, 1,0,1,1,0,1,1,0,1,1,1,0,1,1 ],

"Fraud": [ 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,0,0,0,1,0,1,1,1,1,1,1,1, 1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
1,1,1,1,1,1,1,1,1,1,1,1,1,1 ]
})
```

ستون Fraud متغیر هدف است.

الف) ابتدا داده‌ها را بررسی کنید.

- پنج سطر اول را نمایش دهید.
- تعداد نمونه‌های هر کلاس (Fraud=0 و Fraud=1) را محاسبه کنید.
- آیا داده‌ها متوازن (Balanced) هستند یا خیر؟

ب) داده‌ها را به دو بخش آموزش و آزمون تقسیم کنید.

شرایط:

- نسبت داده‌های آموزش و آزمون: 80/20
- مقدار random_state=42

ج) یک مدل Logistic Regression برای تشخیص تراکنش‌های تقلبی آموزش دهید.

د) برای داده‌های آزمون:

- کلاس هر تراکنش را پیش‌بینی کنید.
- سعی کنید خروجی را در قالب جدول زیر یا پرینت ساده نمایش دهید.

Amount	TransactionCount	International	Actual	Predicted

هـ) برای تراکنشی با مشخصات زیر، نتیجه را پیش‌بینی کنید:

Amount = 400
TransactionCount = 5
International = 1

هم موارد زیر را گزارش کنید:

- کلاس پیش‌بینی‌شده
- احتمال تقلبی بودن تراکنش

و) به پرسش‌های زیر پاسخ دهید.

1. اگر مبلغ تراکنش افزایش پیدا کند، آیا همیشه احتمال تقلبی بودن بیشتر می‌شود؟ چرا؟
2. چرا در مسائل تشخیص تقلب، داشتن احتمال پیش‌بینی مدل اهمیت زیادی دارد؟
3. اگر تعداد نمونه‌های تقلبی بسیار کمتر از نمونه‌های سالم باشد، چه مشکلی برای مدل ایجاد می‌شود؟

سؤال ۵ — تحلیل خروجی Logistic Regression

موضوع: ارزیابی مدل تشخیص تقلب بانکی

در این سؤال از مدل آموزش‌داده‌شده در سؤال قبل استفاده کنید.

الف) برای داده‌های آزمون موارد زیر را محاسبه کنید:

- Accuracy
- Confusion Matrix
- Classification Report

ب) ماتریس درهم‌ریختگی را رسم یا نمایش دهید.

سپس مشخص کنید:

- تعداد True Positive (TP)
- تعداد True Negative (TN)
- تعداد False Positive (FP)
- تعداد False Negative (FN)

ج) با استفاده از تابع:

```
predict_proba()
```

احتمال تعلق هر تراکنش به کلاس‌های 0 و 1 را محاسبه کنید.

خروجی را در قالب جدول زیر یا پرینت ساده نمایش دهید:

Transaction	Probability(No Fraud)	Probability(Fraud)

د) به خروجی predict() و predict_proba() توجه کنید و پاسخ دهید:

1. تفاوت این دو تابع چیست؟
2. چرا در سیستم‌های تشخیص تقلب، احتمال مدل می‌تواند از کلاس نهایی مهم‌تر باشد؟
3. چرا بانک ممکن است بخواهد تراکنشی که احتمال تقلب آن 0.6 است را نیز بررسی کند؟

هـ) فرض کنید خروجی predict_proba() برای چهار تراکنش به شکل زیر باشد:

Transaction	Fraud Probability
A	0.95
B	0.52
C	0.15
D	0.75

پاسخ دهید:

1. مدل برای هر تراکنش چه کلاسی پیش‌بینی می‌کند؟
2. کدام پیش‌بینی با اطمینان بیشتری انجام شده است؟
3. کدام تراکنش نزدیک مرز تصمیم‌گیری قرار دارد؟
4. اگر آستانه تصمیم‌گیری از 0.5 به 0.7 تغییر کند، کدام پیش‌بینی‌ها تغییر می‌کنند؟

و) فرض کنید خروجی مدل روی داده‌های آزمون به شکل زیر باشد:

```
Accuracy = 95%

:Confusion Matrix

        Predicted
            1   0

Actual 0    95   2
Actual 1     3   0
```

به پرسش‌های زیر پاسخ دهید:

1. آیا Accuracy بالا نشان‌دهنده عملکرد مناسب مدل است؟
2. مشکل اصلی مدل در تشخیص تراکنش‌های تقلبی چیست؟
3. چرا Recall برای کلاس Fraud اهمیت زیادی دارد؟
4. اگر هدف بانک جلوگیری از عبور تراکنش‌های تقلبی باشد، کدام خطا مهم‌تر است؟ False Positive یا False Negative؟ چرا؟

ز) یکی از برنامه‌نویسان می‌گوید:

«مدل من Accuracy برابر 95٪ دارد، پس مدل کاملاً قابل اعتماد است.»

این جمله را نقد کنید و توضیح دهید:

- چرا Accuracy ممکن است در مسائل Fraud Detection گمراه‌کننده باشد؟
- Confusion Matrix چه اطلاعاتی را نشان می‌دهد که Accuracy نشان نمی‌دهد؟
- چرا Recall و Precision برای کلاس تقلب اهمیت زیادی دارند؟

سؤال ۶ — آموزش مدل KNN Regressor

در این سؤال قصد داریم با استفاده از الگوریتم **K-Nearest Neighbors Regressor**، قیمت خودروهای دست دوم را پیش‌بینی کنیم.

Dataset

```
import pandas as pd

car_data = pd.DataFrame({
"Age": [ 1,1,2,2,2,3,3,4,4,5,5,6, 6,7,7,8,8,9,9,10,10,11,12,12, 13,14,15,16,17,18,19,20,21,22,
23,24,25,26,27,28,29,30,31,32, 33,34,35,36 ],

"Mileage": [ 10000,12000,15000,18000,22000,25000,30000,32000,
38000,42000,47000,50000,55000,62000,65000,70000,
75000,82000,87000,93000,98000,105000,115000,130000,
135000,145000,155000,165000,175000,185000,195000,210000,
220000,235000,250000,265000,280000,295000,310000,325000,
340000,355000,370000,385000,400000,420000,440000,460000 ],

"EngineSize": [ 1.2,1.2,1.4,1.4,1.4,1.6,1.6,1.6, 1.8,1.8,1.8,1.8,2.0,2.0,2.0,2.0,
2.2,2.2,2.2,2.2,2.4,2.4,2.5,2.5, 2.5,2.7,2.7,2.8,2.8,3.0,3.0,3.0, 3.2,3.2,3.5,3.5,3.5,3.8,3.8,4.0,
4.0,4.2,4.2,4.5,4.5,4.8,5.0,5.0 ],

"Price": [ 23500,23000,22000,21500,21000,20000,19800,19000,
18500,18000,17200,16800,16000,15500,15000,14500,
13800,13200,12800,12300,11800,11200,10500,9800,
9200,8600,8000,7500,7000,6500,6000,5500, 5000,4600,4200,3800,3500,3200,2900,2600,
2400,2200,2000,1800,1600,1400,1200,1000 ]
})
```

هدف، پیش‌بینی ستون **Price** است.

الف) داده‌ها را بررسی کنید.

1. پنج سطر اول را نمایش دهید.
2. تعداد نمونه‌ها و ویژگی‌ها را مشخص کنید.
3. ویژگی‌ها (Features) و متغیر هدف (Target) را تعیین کنید.

ب) داده‌ها را به نسبت 80٪ آموزش و 20٪ آزمون تقسیم کنید.

از random_state = 42 استفاده کنید.

ج) قبل از آموزش مدل، ویژگی‌ها را با استفاده از StandardScaler استانداردسازی کنید. سپس یک مدل KNeighborsRegressor با مقدار زیر آموزش دهید.

```
n_neighbors = 3
```

د) برای داده‌های آزمون قیمت خودروها را پیش‌بینی کنید.

جدولی مشابه جدول زیر ایجاد کنید.

Actual Price	Predicted Price
...	...

هـ) برای خودروی زیر قیمت را پیش‌بینی کنید.

- Age = 4 سال
- Mileage = 45000
- EngineSize = 1.8

و) به پرسش‌های زیر پاسخ دهید.

1. چرا قبل از آموزش مدل، ویژگی‌ها را استانداردسازی کردید؟
2. اگر ویژگی **Mileage** بر حسب کیلومتر و ویژگی **EngineSize** بر حسب لیتر باشد، آیا بدون استانداردسازی فاصله‌ها به درستی محاسبه می‌شوند؟ توضیح دهید.
3. آیا KNN در مرحله آموزش واقعاً مدلی یاد می‌گیرد؟ دلیل خود را بیان کنید.

سؤال ۷ — تحلیل KNN Regressor و اثر مقدار K

در این سؤال از مدل آموزش‌داده‌شده در سؤال قبل استفاده کنید.

الف) معیارهای زیر را برای داده‌های آزمون محاسبه کنید.

- MAE
- MSE
- R^2

ب) مقدار K را برابر مقادیر زیر قرار دهید.

K = 1
K = 3
K = 5
K = 7
K = 9

برای هر مقدار مدل را آموزش داده و مقدار R^2 را محاسبه کنید.

جدول زیر را کامل کنید.

K	MAE	MSE	R^2
1	...	...	...
3	...	...	...
5	...	...	...
7	...	...	...
9	...	...	...

ج) نموداری رسم کنید که در آن:

- محور افقی مقدار **K**
- محور عمودی مقدار **R^2**

باشد.

سپس به پرسش‌های زیر پاسخ دهید.

1. بهترین مقدار K کدام است؟
2. آیا افزایش K همیشه باعث افزایش دقت مدل می‌شود؟
3. اگر K بسیار بزرگ شود، مدل چه رفتاری خواهد داشت؟

د) فرض کنید دو مدل زیر آموزش داده شده‌اند.

مدل	R^2
Linear Regression	0.91
KNN Regressor	0.89

به پرسش‌های زیر پاسخ دهید.

1. آیا صرفاً با نگاه کردن به R^2 می‌توان بهترین مدل را انتخاب کرد؟
2. چه معیارهای دیگری را بررسی می‌کنید؟
3. در چه شرایطی ممکن است KNN عملکرد بهتری نسبت به Linear Regression داشته باشد؟

هـ) به پرسش‌های مفهومی زیر پاسخ دهید.

1. چرا KNN را یک **Lazy Learning Algorithm** می‌نامند؟
2. تفاوت اصلی مرحله آموزش در KNN و Linear Regression چیست؟
3. اگر تعداد نمونه‌های Dataset ده برابر شود، انتظار دارید زمان پیش‌بینی KNN چگونه تغییر کند؟ چرا؟

و) فرض کنید یکی از همکلاسی‌های شما ویژگی‌ها را استانداردسازی نکرده است اما مدل را آموزش داده و مقدار $R^2 = 0.48$ به دست آورده است.

به نظر شما:

1. یکی از دلایل احتمالی این عملکرد ضعیف چیست؟
2. چگونه می‌توان این فرضیه را بررسی کرد؟
3. آیا همیشه استانداردسازی باعث بهبود عملکرد KNN می‌شود؟ توضیح دهید.

سؤال ۸ — آموزش مدل KNN Classification

در این سؤال قصد داریم با استفاده از **KNN Classifier**، وضعیت سلامت افراد را پیش‌بینی کنیم.

Dataset

```
import pandas as pd

health_data = pd.DataFrame({
"BMI": [ 18,19,20,21,22,23,24,24,25,26,27,28,29,30,
31,32,33,34,35,36,37,38,39,40,41,42,43,44, 20,22,23,25,27,28,30,32,34,36,38,40,42,44 ],

"BloodPressure": [ 110,111,112,115,118,119,120,122,123,126,128,130,132,135,
138,140,142,143,145,148,150,152,154,157,160,162,164,166,
113,117,121,124,129,131,136,141,144,149,153,158,163,167 ],

"HighRisk": [ 0,0,0,0,0,0,0,0,0,0,0,0, 1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1, 0,0,0,0,1,1,1,1,1,1,1,1,1,1 ]
})
```

هدف، پیش‌بینی ستون **HighRisk** است.

الف) ابتدا داده‌ها را بررسی کنید.

1. پنج سطر اول را نمایش دهید.
2. تعداد نمونه‌های هر کلاس را محاسبه کنید.
3. آیا داده‌ها متوازن هستند؟

ب) داده‌ها را به نسبت 80٪ آموزش و 20٪ آزمون تقسیم کنید.

از random_state=42 استفاده کنید.

ج) ویژگی‌ها را با استفاده از StandardScaler استانداردسازی کنید. سپس یک مدل KNeighborsClassifier با مقدار زیر آموزش دهید.

```
n_neighbors = 3
```

د) برای داده‌های آزمون:

1. کلاس هر نمونه را پیش‌بینی کنید.
2. Accuracy مدل را محاسبه کنید.

هـ) جدولی مشابه جدول زیر ایجاد کنید.

Actual	Predicted
...	...

و) برای بیمار زیر پیش‌بینی انجام دهید.

- BMI = 28
- BloodPressure = 130

هم **کلاس پیش‌بینی‌شده** و هم **احتمال تعلق به هر کلاس** را گزارش کنید.

سؤال ۹ — تحلیل KNN Classification و اثر مقدار K

در این سؤال از مدل آموزش‌داده‌شده در سؤال قبل استفاده کنید.

الف) برای مقادیر زیر مدل را آموزش دهید.

K = 1
K = 3
K = 5
K = 7
K = 9

برای هر مقدار، معیارهای زیر را محاسبه کنید.

- Accuracy
- Precision
- Recall
- F1-score

جدول زیر را کامل کنید.

K	Accuracy	Precision	Recall	F1
1	...	...	...	...
3	...	...	...	...
5	...	...	...	...
7	...	...	...	...
9	...	...	...	...

ب) برای بهترین مقدار K، موارد زیر را نمایش دهید.

1. Confusion Matrix
2. Classification Report

سپس مشخص کنید.

- TP
- TN
- FP
- FN

ج) فرض کنید خروجی predict_proba() برای سه بیمار به صورت زیر باشد.

بیمار	احتمال سالم	احتمال پرخطر
A	0.49	0.51
B	0.05	0.95
C	0.52	0.48

به پرسش‌های زیر پاسخ دهید.

1. مدل هر بیمار را در کدام کلاس قرار می‌دهد؟
2. کدام بیمار نزدیک مرز تصمیم‌گیری قرار دارد؟
3. کدام پیش‌بینی از اطمینان بیشتری برخوردار است؟

د)نتایج مقادیر مختلف K را تحلیل کنید.

1. آیا بهترین مقدار K همان کوچک‌ترین مقدار بود؟ چرا؟
2. اگر K برابر تعداد کل نمونه‌ها باشد، مدل چگونه تصمیم‌گیری می‌کند؟
3. اگر K برابر ۱ باشد، مدل چه مزایا و چه معایبی خواهد داشت؟
4. چگونه می‌توان مقدار مناسب K را انتخاب کرد؟

هـ) فرض کنید دو مدل زیر را آموزش داده‌ایم.

مدل	Accuracy
Logistic Regression	91%
KNN Classifier	91%

با وجود یکسان بودن Accuracy، آیا می‌توان نتیجه گرفت که عملکرد دو مدل کاملاً مشابه است؟

سؤال ۱۰ — آموزش مدل Decision Tree

در این سؤال قصد داریم با استفاده از **Decision Tree Classifier**، وضعیت تأیید یا رد درخواست وام مشتریان را پیش‌بینی کنیم.

Dataset:

```
import pandas as pd

loan_data = pd.DataFrame({

"Income": [ 30,32,34,35,37,39,40,42,44,45,47,49,50,52,
54,55,57,59,60,62,64,65,67,69,70,72,74,75, 77,79,80,82,84,86,88,90,92,94,96,98,100,102,
104,106,108,110,112,114,116,118,120,122,124,126 ],

"CreditScore": [ 580,585,590,600,605,608,610,615,618,620,625,630,640,645,
650,660,665,670,680,685,690,700,705,710,720,725,730,740,
745,750,760,765,770,775,778,780,785,790,795,798,800,803,
806,810,812,820,822,825,828,830,832,835,838,840 ],

"LoanAmount": [ 120,118,115,110,108,107,105,103,102,100,98,97,95,93,
92,90,89,88,85,84,83,80,79,78,75,74,73,70, 69,68,65,64,63,62,61,60,59,58,57,56,55,54,
53,52,51,50,49,48,47,46,45,44,43,42 ],

"Approved": [ 0,0,0,0,0,0,0,0,0,0,0,0, 0,0,0,1,1,1,1,1,1,1,1,1, 1,1,1,1,1,1,1,1,1,1,1,1,
1,1,1,1,1,1,1,1,1,1,1,1, 1,1,1,1,1,1,1,1 ]
 })
```

هدف، پیش‌بینی ستون **Approved** است.

الف) ابتدا داده‌ها را بررسی کنید.

پنج سطر اول را نمایش دهید.
تعداد نمونه‌ها و ویژگی‌ها را مشخص کنید.
تعداد نمونه‌های هر کلاس را محاسبه کنید.

ب) داده‌ها را به نسبت 80٪ آموزش و 20٪ آزمون تقسیم کنید.

از random_state=42 استفاده کنید.

ج) یک مدل DecisionTreeClassifier با تنظیمات پیش‌فرض آموزش دهید.

د) برای داده‌های آزمون:

کلاس هر نمونه را پیش‌بینی کنید.
مقدار Accuracy را محاسبه کنید.
Confusion Matrix را نمایش دهید.

هـ) درخت تصمیم را با استفاده از تابع plot_tree() رسم کنید.

و) به نمودار درخت نگاه کنید و به پرسش‌های زیر پاسخ دهید.

اولین ویژگی که در ریشه (Root Node) قرار گرفته است چیست؟
چرا فکر می‌کنید مدل این ویژگی را برای اولین تقسیم انتخاب کرده است؟
هر برگ (Leaf Node) چه مفهومی دارد؟

ز) برای مشتری زیر نتیجه درخواست وام را پیش‌بینی کنید.

Income = 68
CreditScore = 710
LoanAmount = 78

ح) به پرسش‌های مفهومی زیر پاسخ دهید.

آیا برای Decision Tree نیاز به StandardScaler وجود دارد؟ چرا؟
مهم‌ترین تفاوت Decision Tree و Logistic Regression در نحوه تصمیم‌گیری چیست؟
آیا Decision Tree می‌تواند روابط غیرخطی را یاد بگیرد؟ توضیح دهید.

سؤال ۱۱ — تحلیل Decision Tree و مفهوم Overfitting

در این سؤال از مدل آموزش‌داده‌شده در سؤال قبل استفاده کنید.

الف) Accuracy مدل را روی هر دو مجموعه زیر محاسبه کنید.

داده‌های آموزش (Training)
داده‌های آزمون (Testing)
جدول زیر را کامل کنید.

Dataset	Accuracy
Train	...
Test	...

ب) فرض کنید نتایج زیر به دست آمده است.

Dataset	Accuracy
Train	1.00
Test	0.73

به پرسش‌های زیر پاسخ دهید.

آیا مدل دچار Overfitting شده است؟
دلیل پاسخ خود را توضیح دهید.
چرا عملکرد مدل روی داده‌های آموزش و آزمون متفاوت است؟

ج) برای کاهش Overfitting، مدل را دوباره آموزش دهید؛ اما این بار از پارامترهای زیر استفاده کنید.

```
max_depth=3
random_state=42
```

دوباره Accuracy مجموعه آموزش و آزمون را محاسبه کنید.

د) نتایج دو مدل را با هم مقایسه کنید.

مدل	Train Accuracy	Test Accuracy
بدون محدودیت	...	...
max_depth=3	...	...

هـ) به پرسش‌های زیر پاسخ دهید.

آیا محدود کردن عمق درخت باعث کاهش Overfitting شد؟
آیا همیشه Accuracy آموزش باید بیشترین مقدار ممکن باشد؟
اگر عمق درخت بیش از حد کوچک باشد، چه مشکلی ایجاد می‌شود؟

و) یکی از کاروندها می‌گوید:

«اگر Accuracy آموزش برابر ۱۰۰٪ باشد، یعنی بهترین مدل ممکن را ساخته‌ایم.»

نظر این کاروند را نقد کنید و در پاسخ خود به موارد زیر اشاره کنید:

- مفهوم Generalization
- Overfitting
- اهمیت داده‌های آزمون

سؤال ۱۲ — آموزش K-Means Clustering

در این سؤال قصد داریم مشتریان یک فروشگاه را بدون داشتن برچسب، به چند گروه تقسیم کنیم.

Dataset:

```
import pandas as pd

customer_data = pd.DataFrame({
"AnnualIncome": [ 15,18,20,22,25,28,30,32,35,38,40,42, 45,48,50,52,55,58,60,62,65,68,70,
72,75,78,80,82,85,88,90,95,100,105,110, 115,120,125,130,135,140,145,150,160,170 ],

"SpendingScore": [ 90,85,80,82,78,75,72,70,68,65,62,60, 58,55,52,50,48,45,42,40,38,35,32,
30,28,25,23,20,18,15,12,10,8,15,18, 20,22,25,28,30,32,35,38,40,42,45 ]
})
```

توجه: این Dataset هیچ برچسبی ندارد.

الف) داده‌ها را بررسی کنید.

پنج سطر اول را نمایش دهید.
تعداد نمونه‌ها را مشخص کنید.
آیا این Dataset دارای متغیر هدف (Target) است؟ توضیح دهید.

ب)ویژگی‌ها را با استفاده از StandardScaler استانداردسازی کنید.

ج) یک مدل KMeans با تنظیم زیر آموزش دهید.

```
n_clusters=3
random_state=42
```

د) برای هر مشتری شماره خوشه (Cluster Label) را به دست آورید و آن را به Dataset اضافه کنید.

نمونه‌ای مشابه جدول زیر ایجاد کنید.

AnnualIncome	SpendingScore	Cluster
...	...	...

هـ) نمودار پراکندگی (Scatter Plot) داده‌ها را رسم کنید.

رنگ هر نقطه بر اساس خوشه آن باشد.
مراکز خوشه‌ها (Centroids) را نیز روی نمودار نمایش دهید.

و) مختصات مراکز خوشه‌ها را نمایش دهید.

ز) به پرسش‌های زیر پاسخ دهید.

آیا K-Means برای آموزش خود به برچسب کلاس نیاز دارد؟ چرا؟
هر عدد موجود در ستون **Cluster** چه مفهومی دارد؟
آیا شماره خوشه (مثلاً ۰، ۱ یا ۲) نشان‌دهنده ترتیب یا کیفیت خوشه‌هاست؟ توضیح دهید.

ح) با مشاهده نمودار، برای هر خوشه یک نام یا توصیف مناسب انتخاب کنید.

برای مثال:

مشتریان کم‌درآمد و پرخرج
مشتریان پردرآمد و کم‌خرج
مشتریان با رفتار خرید متعادل
(نام‌ها صرفاً نمونه هستند و باید بر اساس خروجی واقعی مدل انتخاب شوند.)

ط) K-Means را با مقادیر زیر نیز اجرا کنید.

```
n_clusters = 2
n_clusters = 4
```

سپس به پرسش‌های زیر پاسخ دهید.

با تغییر تعداد خوشه‌ها چه تغییری در نتیجه مشاهده می‌کنید؟
آیا همیشه افزایش تعداد خوشه‌ها باعث بهتر شدن خوشه‌بندی می‌شود؟
اگر مقدار n_clusters را خیلی بزرگ انتخاب کنیم، چه مشکلی ایجاد خواهد شد؟

پروژه ترکیبی — تحلیل داده‌های یک پلتفرم پخش پادکست

معرفی پروژه

یک استارتاپ پخش پادکست دو مجموعه داده از سیستم خود استخراج کرده و از شما خواسته با ترکیب مباحث (Regression، Clustering، Classification و ارزیابی صحیح مدل)، یک تحلیل کامل ارائه دهید.

این پروژه از **۵ مرحله پیوسته** تشکیل شده و هر مرحله روی خروجی مرحله قبل بنا می‌شود.

دیتاست ۱ — podcast_episodes.csv (۲۲۰ اپیزود) لینک دسترسی

ستون	توضیح
episode_id	شناسه اپیزود
episode_length_min	طول اپیزود (دقیقه)
genre	ژانر (Comedy, True Crime, Tech, Health, Business, Fiction)
num_ads	تعداد تبلیغات در اپیزود
guest_rating	امتیاز میهمان برنامه (۱ تا ۵)
host_experience_years	سابقه مجری (سال)
avg_listen_minutes	هدف **Regression** — میانگین دقایقی که شنوندگان این اپیزود را گوش داده‌اند

دیتاست ۲ — podcast_listeners.csv (۳۵۰ شنونده) لینک دسترسی

ستون	توضیح
listener_id	شناسه شنونده
hours_per_week	ساعت گوش‌دادن در هفته
num_genres_followed	تعداد ژانرهای دنبال‌شده
num_episodes_completed	تعداد اپیزودهای کامل گوش‌داده‌شده
skip_rate_percent	درصد اپیزودهایی که نیمه‌کاره رد شده‌اند
days_since_signup	تعداد روز از ثبت‌نام
is_premium	هدف **Classification** — آیا اشتراک پولی خریده (۱) یا خیر (۰)

مرحله ۱ — پیش‌بینی محبوبیت اپیزود (Regression)

Implement:

با استفاده از podcast_episodes.csv، یک مدل **Linear Regression** برای پیش‌بینی avg_listen_minutes بر اساس ویژگی‌های عددی (episode_length_min, num_ads, guest_rating, host_experience_years) بسازید. داده را ۸۰/۲۰ به Train/Test تقسیم کنید.

Observe:
MAE، RMSE و R^2 مدل را روی Test گزارش کنید.

Reason:
بر اساس ضرایب مدل (model.coef_)، کدام ویژگی بیشترین تأثیر مثبت و کدام بیشترین تأثیر منفی را روی مدت‌زمان گوش‌دادن دارد؟ آیا این نتیجه با شهود شما (مثلاً تأثیر منفی تعداد تبلیغات) هم‌خوانی دارد؟

مرحله ۲ — پیش‌بینی خرید اشتراک (Premium (Classification

Implement:
با استفاده از podcast_listeners.csv، یک مدل **Logistic Regression** برای پیش‌بینی is_premium بسازید (پیشنهاد: ابتدا ویژگی‌ها را با StandardScaler مقیاس‌بندی کنید).

Observe:
Confusion Matrix، Accuracy، Precision، Recall و F1-score را روی Test گزارش کنید. همچنین درصد شنوندگان Premium در کل داده را محاسبه کنید.

Reason:
با توجه به نسبت کلاس‌ها (آیا داده متوازن است یا نه؟)، آیا Accuracy به‌تنهایی معیار قابل‌اعتمادی برای این مدل است؟ توضیح دهید.

مرحله ۳ — کشف الگوی رفتاری شنوندگان (Clustering)

Implement:
با استفاده از دو ویژگی hours_per_week و num_genres_followed از دیتاست شنوندگان، یک مدل **K-Means** با ۳ خوشه اجرا کنید و نتیجه را با نمودار پراکندگی (رنگ‌بندی‌شده بر اساس خوشه) نمایش دهید.

Reason:
یک نام معنادار به هر یک از سه خوشه بدهید (مثلاً «شنونده گاه‌به‌گاه»، «شنونده پرمصرف»).

چرا برای این بخش نمی‌توانیم از Accuracy یا معیارهای Classification برای سنجش کیفیت خوشه‌بندی استفاده کنیم؟

مرحله ۴ — بررسی Overfitting در مدل Classification

Implement:
به‌جای Logistic Regression، یک **Decision Tree** برای پیش‌بینی is_premium آموزش دهید؛ یک‌بار با max_depth=None و یک‌بار با max_depth=3. Accuracy هر دو مدل را روی Train و Test گزارش کنید.

Reason:
کدام مدل علائم Overfitting نشان می‌دهد؟ چرا محدودکردن عمق درخت معمولاً باعث بهبود عملکرد روی داده جدید (Test) می‌شود؟

مرحله ۵ — ارزیابی نهایی (Cross Validation)

Implement:
روی مدل Decision Tree با max_depth=3، یک **5-Fold Cross Validation** با معیار F1-score اجرا کنید و میانگین و انحراف معیار نمرات را گزارش کنید.

Reason:
مقایسه کنید:

نتیجه تک Train/Test Split در مرحله ۴ با میانگین ۵-Fold در این مرحله چه تفاوتی دارد؟ کدام‌یک را برای گزارش نهایی به تیم محصول پیشنهاد می‌کنید و چرا؟

سؤال چالشی – پیش‌بینی خرابی ماشین‌آلات در خط تولید (اختیاری!)

یک کارخانه تولیدی روی هر ماشین سنسورهایی نصب کرده که داده‌های عملکردی را ثبت می‌کنند. تیم مهندسی می‌خواهد مدلی بسازد که **پیش از وقوع خرابی**، احتمال آن را پیش‌بینی کند تا بتوان نگهداری پیشگیرانه انجام داد.

دیتاست: machine_failure.csv (۴۲۰ رکورد) لینک دسترسی

ستون	توضیح
machine_id	شناسه ماشین
temperature_c	دمای عملیاتی ماشین (درجه سانتی‌گراد)
vibration_mm_s	میزان لرزش (میلی‌متر بر ثانیه)
pressure_bar	فشار داخلی (بار)
run_hours_since_maintenance	ساعات کارکرد از آخرین سرویس
operator_shift	شیفت کاری (Morning / Evening / Night)
failure	**متغیر هدف** — ۱ = خراب شده، ۰ = سالم

تنها حدود **۹.۵٪** ماشین‌ها دچار خرابی شده‌اند (داده نامتوازن – عمداً طراحی شده).

الف) تقسیم داده

1. داده را با نسبت ۸۰/۲۰ به Train و Test تقسیم کنید طوری که **نسبت کلاس failure=1 در هر دو مجموعه حفظ شود.**
2. اگر این تقسیم بدون رعایت نسبت کلاس انجام می‌شد، چه ریسکی برای ارزیابی مدل روی Test وجود داشت؟

ب) مدل پایه (Implement → Observe → Reason) Accuracy

1. یک مدل ساختگی بسازید که برای همه نمونه‌ها پیش‌بینی «سالم است» (۰) را برمی‌گرداند.
2. Recall، Precision، Accuracy و F1-score آن را روی Test گزارش کنید.
3. **Reason:** عدد Accuracy این مدل بی‌فایده چقدر بالاست؟ کدام معیار نقص واقعی آن را آشکار می‌کند و چرا؟

ج) Overfitting در (Implement → Observe → Reason) Decision Tree

1. دو Decision Tree روی Train آموزش دهید: یکی با max_depth=None و دیگری با max_depth=4.
2. Accuracy ا هر دو مدل را روی Train و Test گزارش دهید.
3. کدام مدل Overfit شده؟ از روی چه شاخصی تشخیص دادید؟

د) ارزیابی قابل‌اعتماد با Stratified K-Fold (Implement → Observe → Reason)

1. روی مدل max_depth=4، یک Stratified 5-Fold Cross Validation با معیار **Recall** اجرا کنید؛ میانگین و انحراف معیار ۵ نمره را گزارش کنید.
2. همان کار را با KFold معمولی تکرار کنید و نسبت کلاس failure=1 را در هر Fold چاپ کنید.
3. تفاوت نسبت کلاس بین Foldهای دو روش را مقایسه کنید و توضیح دهید چرا در این داده نامتوازن، Stratified K-Fold ایمن‌تر است.

ه) بخش نهایی

فرض کنید نمرات Recall حاصل از ۵ Fold در بخش (د) این‌ها بوده‌اند:

[0.58, 0.67, 0.50, 0.63, 0.55] → میانگین ≈ ۰.۵۹، انحراف معیار ≈ ۰.۰۶

در همان زمان، یک تقسیم ساده Train/Test (بدون Cross Validation) عدد Recall ≈ ۰.۶۸ داده است.

1. چرا این دو عدد می‌توانند نتیجه‌گیری‌های متفاوتی درباره کیفیت مدل ایجاد کنند؟
2. برای گزارش به مدیریت کارخانه که «مدل قابل اعتماد است»، کدام‌یک از این دو نوع گزارش (تک Split در برابر میانگین+انحراف‌معیار CV) شواهد قوی‌تری است؟ چرا؟