تمرین 1. مقایسه Linear Regression، Elastic Net و تأثیر کاهش ابعاد با PCA

مقدمه

دیتاست **California Housing** شامل 20,640 نمونه با 8 ویژگی عددی است که از سرشماری 1990 ایالات متحده گرفته شده. هدف پیش‌بینی قیمت متوسط مسکن (به صد هزار دلار) در مناطق مختلف کالیفرنیاست.

در این تمرین جامع، سه رویکرد مختلف را مقایسه می‌کنیم:

1. **Linear Regression** (مدل پایه)
2. **Elastic Net Regression** (با regularization)
3. **Linear Regression + PCA** (کاهش ابعاد)

ویژگی‌های دیتاست

دیتاست شامل ویژگی‌های زیر است:

- MedInc: متوسط درآمد در منطقه
- HouseAge: متوسط سن خانه‌ها
- AveRooms: تعداد متوسط اتاق‌ها در هر خانوار
- AveBedrms: تعداد متوسط اتاق‌خواب‌ها
- Population: جمعیت منطقه
- AveOccup: متوسط تعداد افراد در هر خانوار
- Latitude: عرض جغرافیایی
- Longitude: طول جغرافیایی

هدف: MedHouseVal (قیمت متوسط خانه به صد هزار دلار)

بخش 1: بارگذاری و آشنایی با داده

1-1. با استفاده از کد زیر داده را بارگذاری کنید:

```
from sklearn.datasets import fetch_california_housing
```

بخش 2: آماده‌سازی داده

2-1. داده را به ویژگی‌ها (X) و متغیر هدف (y) تقسیم کنید.

2-2. با نسبت 80-20 داده را به train و test تقسیم کنید. از random_state=42 استفاده کنید.

2-3. با استفاده از StandardScaler، داده train را fit کرده و هر دو مجموعه train و test را transform کنید.

بخش 3:مدل Linear Regression پایه

3-1. یک مدل Linear Regression بسازید و روی داده train آموزش دهید.

3-2. برای این مدل موارد زیر را محاسبه و گزارش کنید:

- R2_score روی train
- R2_score روی test
- اختلاف بین اعداد بالا را بررسی کنید (شاخص overfiting)

3-3. ضرایب (coefficients) مدل را به همراه نام ویژگی‌ها نمایش دهید و به ترتیب نزولی مرتب کنید. کدام ویژگی بیشترین تأثیر مثبت (بزرگترین ضریب مثبت) روی قیمت مسکن دارد؟

بخش 4: مدل Elastic Net اولیه

4-1. یک مدل ElasticNet با پارامترهای alpha=0.1 و l1_ratio=0.5 بسازید و آموزش دهید. (برای فهم هایپر پارامترهای داده شده به بخش 6 مراجعه کنید)

4-2. همان معیارهای بخش 3 را برای این مدل محاسبه کنید:

- R2_scoreروی train
- R2_score روی test
- اختلاف بین اعداد بالا را بررسی کنید (شاخص overfiting)

4-3. ضرایب Elastic Net را با Linear Regression مقایسه کنید. یک جدول بسازید که شامل نام ویژگی، ضریب Linear، ضریب Elastic Net، و اختلاف آن‌ها باشد. آیا Elastic Net ضرایب را کوچک‌تر (shrink) کرده است؟

3-5. تحلیل کنید:

- کدام مدل overfitting کمتری دارد؟ چگونه تشخیص دادید؟
- Elastic Net چطور ضرایب را تغییر داده است؟ (مفهوم shrinkage)
- کدام مدل برای generalization (عملکرد روی داده جدید) بهتر است؟

بخش 6: توضیحات و سوالات تحلیلی

6-1- توضیح هایپرپارامتر ها (این بخش برای مطالعه است که درک بهتری پیدا کنید)

Alpha یک **هایپرپارامتر** است که مشخص می‌کند **چقدر پنالتی (جریمه)** به ضرایب بزرگ اعمال شود:

- **Alpha کوچک (مثل 0.01)**: پنالتی کم → مدل آزادتر است → ضرایب بزرگ‌تر → خطر overfitting

- **Alpha بزرگ (مثل 10)**: پنالتی زیاد → مدل محدودتر است → ضرایب کوچک‌تر → خطر underfitting
- **Alpha بهینه**: تعادل بین پیچیدگی و generalization، معمولاً با **Cross-Validation** پیدا می‌شود در این سوال هدف پیدا کردن بهترین مقدار alpha نبوده صرفا هدف آشنایی شما با این هایپر پارامتر است
- **پنالتی یعنی**: مدل برای هر ضریب بزرگ، یک **جریمه** به تابع هزینه اضافه می‌کند، پس مجبور می‌شود ضرایب کوچک‌تری انتخاب کند تا overfitting نشود.
- **l1_ratio** مشخص می‌کنه **چه درصدی از پنالتی L1 و چه درصدی L2 باشه**.

وقتی:

l1_ratio = 0.5

یعنی:

- **۵۰٪ L1 (Lasso)**
- **۵۰٪ L2 (Ridge)**

6-2. سوال تحلیلی:

•Regularization چیست؟

تعریف Regularization در مدل‌های یادگیری ماشین چیست و چرا بدون آن، مدل‌ها مستعد **overfitting** می‌شوند؟

- **انواع Regularization**

به‌طور مشخص، دو روش زیر را توضیح دهید:

- **L1 Regularization (Lasso)**
- **L2 Regularization (Ridge)**

و تفاوت‌های آن‌ها را از نظر:

- نحوه جریمه‌کردن ضرایب
- تأثیر بر مقدار ضرایب
- حذف یا باقی‌ماندن ویژگی‌ها (feature selection)

- **Elastic Net Regularization**

توضیح دهید Elastic Net چگونه ترکیبی از L1 و L2 است و چرا در برخی مسائل (به‌ویژه زمانی که ویژگی‌ها با یکدیگر هم‌بستگی دارند) نسبت به Lasso یا Ridge خالص عملکرد بهتری دارد.

بخش 7: کاهش ابعاد با PCA

در این بخش می‌خواهیم ببینیم اگر تعداد ویژگی‌ها را با **PCA** کاهش دهیم، چه تأثیری روی عملکرد مدل دارد.

7-1. کاهش ابعاد با PCA

یک مدل PCA با 5 مؤلفه اصلی بسازید و روی X_train_scaled آموزش دهید. سپس هر دو مجموعه train و test را transform کنید.

7-2. آموزش Linear Regression روی داده کاهش‌یافته

یک مدل LinearRegression روی داده تبدیل‌شده PCA آموزش دهید و معیارهای زیر را محاسبه کنید:

- Train R^2 و Test R^2
- اختلاف بین اعداد بالا را بررسی کنید (شاخص overfiting)

8. بخش تحلیلی: مقایسه سه رویکرد

در این بخش باید **سه مدل** را با هم مقایسه کنید:

1. **Linear Regression** (بدون regularization)
2. **Elastic Net** (با regularization بهینه)
3. **Linear Regression + PCA** (با داده کاهش‌یافته)

تحلیل کنید:

- کدام مدل بهترین Test R^2 را دارد؟
- آیا کاهش ابعاد با PCA باعث بهبود generalization شد؟
- چرا Elastic Net نسبت به Linear Regression ساده بهتر/بدتر عمل کرد؟

تمرین 2. بررسی Random Forest برای پیش‌بینی بیماری قلبی (Heart Disease)

یک مرکز پزشکی می‌خواهد با استفاده از الگوریتم **Random Forest**، احتمال ابتلا به بیماری قلبی را در بیماران پیش‌بینی کند. داده‌های بیماران کلینیک کلیولند در اختیار شماست.

لینک دیتا :

https://drive.google.com/file/d/1quJH3KqmUd7ZBVaU4f4GtgoX0NWtAda8/view?usp=sharing

ویژگی‌های دیتاست:

ویژگی‌های ورودی:

- **age**: سن بیمار
- **sex**: جنسیت (1 = مرد، 0 = زن)
- **cp**: نوع درد قفسه سینه (0–3)
- **trestbps**: فشار خون استراحت (mm Hg)
- **chol**: کلسترول سرم (mg/dl)
- **fbs**: قند خون ناشتا (1 = بله، 0 = خیر)
- **restecg**: نتایج الکتروکاردیوگرافی استراحت (0–2)
- **thalach**: حداکثر ضربان قلب
- **exang**: آنژین ناشی از ورزش (1 = بله، 0 = خیر)
- **oldpeak**: افسردگی ST ناشی از ورزش
- **slope**: شیب بخش ST در اوج ورزش
- **ca**: تعداد رگ‌های اصلی
- **thal**: تالاسمی

متغیر هدف:

- **condition**: وجود بیماری قلبی (0 = سالم، 1 = بیمار)

بخش 1.پیش‌نیاز: مطالعه Gini Impurity و Information Gain

قبل از شروع، **الزامی است** با مفاهیم زیر آشنا شوید:

1. **Gini Impurity چیست و چگونه محاسبه می‌شود؟**
2. **Information Gain چیست و چگونه محاسبه می‌شود؟**

بخش 2: بارگذاری و بررسی داده

2-1. داده را به X (ویژگی‌ها) و y (هدف) تقسیم کنید. سپس با نسبت **70-30** به train و test تقسیم کنید (random_state=42). تعداد نمونه‌های train و test را گزارش دهید.

بخش 3: ساخت و آموزش Random Forest

3-1. یک مدل **RandomForestClassifier** با پارامترهای زیر بسازید و آموزش دهید:

- n_estimators=100
- criterion='gini'
- max_depth=4
- random_state=42

3-2. Accuracy مدل Random Forest را روی مجموعه train و test محاسبه کنید. آیا مدل دچار overfitting است؟ (اختلاف بین Train و Test Accuracy را تحلیل کنید)

3-3. Feature Importance را محاسبه کنید و نمودار نواری رسم کنید. کدام ویژگی‌ها مهم‌ترین نقش را در تشخیص بیماری قلبی دارند؟

3-4. تحلیل کنید: چرا Random Forest معمولاً overfitting کمتری نسبت به یک Decision Tree دارد؟

بخش 4: ارزیابی پیشرفته - Confusion Matrix و تحلیل خطاها

قبل از این بخش، الزامی است مفاهیم Precision, Recall, F1-Score را مطالعه کنید:

- https://www.geeksforgeeks.org/machine-learning/f1-score-in-machine-learning/

4-1. رسم و تحلیل Confusion Matrix

الف) برای مدل Random Forest (با Gini)، **Confusion Matrix** رسم کنید و از آن موارد زیر را استخراج کنید:

- **TP (True Positive)**: بیماران صحیح تشخیص داده شده
- **TN (True Negative)**: سالم‌های صحیح تشخیص داده شده
- **FP (False Positive)**: سالم‌هایی که اشتباه بیمار تشخیص داده شدند
- **FN (False Negative)**: بیمارانی که اشتباه سالم تشخیص داده شدند

ب) تعداد هر کدام از موارد بالا را از Confusion Matrix در هنگام ارایه خود استخراج و گزارش کنید.

ج) با توجه به تحلیل بالا، در سیستم تشخیص پزشکی قلب، از دست دادن یک بیمار (FN)خطرناک‌تر است یا تشخیص اشتباه سالم به عنوان بیمار (FP)؟