Python İle Veri Bilimi Mart 4. Hafta Ödevi

🔗 Görev:

Bir veri kümesi kullanarak (hazıır veri kümesi de kullanabilirsiniz) **basit** ve **çoklu lineer regresyon** modelleri kurun. Aşağıdaki adımları izleyin:

1. **Veri Kümesini Yükleyin:**
   sklearn.datasets.load_diabetes() veri kümesini kullanın.

Kodumuz:

```
import pandas as pd
from sklearn.metrics import mean_squared_error, mean_absolute_error,
r2_score
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.datasets import load_diabetes

# Veri kümesini yükleyelim
database = load_diabetes()

# Kümeyi DF'e çevirelim
df = pd.DataFrame(data=database.data, columns=database.feature_names)

print("\nLineer Regresyon Veri Seti:")
print(df.head())
```

Çıktısı:

```
Lineer Regresyon Veri Seti:
        age       sex       bmi  ...        s4        s5        s6
0  0.038076  0.050680  0.061696  ... -0.002592  0.019907 -0.017646
1 -0.001882 -0.044642 -0.051474  ... -0.039493 -0.068332 -0.092204
2  0.085299  0.050680  0.044451  ... -0.002592  0.002861 -0.025930
3 -0.089063 -0.044642 -0.011595  ...  0.034309  0.022688 -0.009362
4  0.005383 -0.044642 -0.036385  ... -0.002592 -0.031988 -0.046641

[5 rows x 10 columns]
```

2. Basit Lineer Regresyon:

- Hedef değişken: target
- Bağımsız değişken: Veri kümesinden bir sütun seçin (örneğin: BMI)
- Modeli eğitin ve R^2 skorunu yazdırın.

Hedef değişken: target

```
# Hedef değişkenimizi belirleyelim
df['target'] = database.target

print("\nLineer Regresyon Veri Seti:")
print(df.head())
```

Çıktısı:

```
Lineer Regresyon Veri Seti:
        age       sex       bmi        bp  ...        s4        s5        s6  target
0  0.038076  0.050680  0.061696  0.021872  ... -0.002592  0.019907 -0.017646   151.0
1 -0.001882 -0.044642 -0.051474 -0.026328  ... -0.039493 -0.068332 -0.092204    75.0
2  0.085299  0.050680  0.044451 -0.005670  ... -0.002592  0.002861 -0.025930   141.0
3 -0.089063 -0.044642 -0.011595 -0.036656  ...  0.034309  0.022688 -0.009362   206.0
4  0.005383 -0.044642 -0.036385  0.021872  ... -0.002592 -0.031988 -0.046641   135.0
```

Bağımsız değişken olarak veri kümesindeki 'bmi' sütununu seçiyoruz ve bağımlı değişkenimiz olan 'target' ile modelimizi eğitip R^2 skorunu yazdırıyoruz.

Kodumuz:

```
# Model Eğitme Basit Lineer Regresyon
x = df[['bmi']]  # 'bmi' sütununu bir DataFrame formatında seçiyoruz, çünkü model bir DataFrame bekliyor
y = df['target']

x_test, x_train, y_test, y_train = train_test_split(x, y, test_size=0.2, random_state=42)

print(f"Eğitim Seti Boyutu: {x_train.shape}")
print(f"Test Seti Boyutu: {x_test.shape}")

print("\nLineer Regresyon Modeli Eğitiliyor...")

model = LinearRegression()
model.fit(x_train, y_train)

print("\nLineer Regresyon Modeli Eğitildi!")
print("Katsayısal:")
print(model.coef_)
print(f"Intercept (b0) : {model.intercept_}")

# r2 Skoru
y_pred = model.predict(x_test)
```

Çıktısı:

```
Eğitim Seti Boyutu: (89, 1)
Test Seti Boyutu: (353, 1)

Basit Lineer Regresyon Modeli Eğitiliyor...
Basit Lineer Regresyon Modeli Eğitildi!

Katsayısal:
[764.23740072]
Intercept (b0) : 151.0363890069427

Basit Lineer Regresyon r2 skoru: 0.3453
```

---

3. Çoklu Lineer Regresyon:

- Tüm bağımsız değişkenleri kullanarak bir model kurun.
- R^2 skorunu yazdırın ve basit modelle karşılaştırın.

Şimdi, veri kümesindeki tüm sütunları kullanarak modeli eğiteceğiz. Veri kümesindeki bütün sütunlar (target hariç) bağımsız değişkenlerimizdir ve 'target' bağımlı değişkenimizdir. Modeli eğittikten sonra R^2 skorunu yazdıracağız.

Kodumuz:

```python
# Çoklu Regresyon
print("\nÇoklu Lineer Regresyon\n")

# Model Eğitme
new_x = df[['age', 'sex', 'bmi', 'bp', 's1', 's2', 's3', 's4', 's5', 's6']]
new_y = df['target']

new_x_test, new_x_train, new_y_test, new_y_train = train_test_split(new_x,
new_y, test_size=0.2, random_state=42)

print(f"Çoklu Lineer Regresyon Eğitim Seti Boyutu: {new_x_train.shape}")
print(f"Çoklu Lineer Regresyon Test Seti Boyutu: {new_x_test.shape}")

print("\nÇoklu Lineer Regresyon Modeli Eğitiliyor...")

model2 = LinearRegression()
model2.fit(new_x_train, new_y_train)
print("Çoklu Lineer Regresyon Modeli Eğitildi!\n")
print("Katsayısal:")
print(model2.coef_)
print(f"Intercept (b0) : {model2.intercept_}")

# r2 skoru
y_pred1 = model2.predict(new_x_test)
r2 = r2_score(new_y_test, y_pred1)
print(f"\nÇoklu Lineer Regresyon r2 skoru = {r2:.4f}")
```

Çıktısı:

```
Çoklu Lineer Regresyon

Çoklu Lineer Regresyon Eğitim Seti Boyutu: (89, 10)
Çoklu Lineer Regresyon Test Seti Boyutu: (353, 10)

Çoklu Lineer Regresyon Modeli Eğitiliyor...
Çoklu Lineer Regresyon Modeli Eğitildi!

Katsayısal:
[-135.25250072 -257.41126615  425.59094427  225.23812362  356.0659046
 -247.30397143 -418.2905857  -183.75571841  601.39789354  165.37404135]
Intercept (b0) : 154.3110080115359

Çoklu Lineer Regresyon r2 skoru = 0.4725
```

---

4. **Hata Metrikleri:**
   Her iki model için aşağıdaki metrikleri hesaplayın:
   - MAE (Ortalama mutlak hata)
   - MSE (Ortalama kare hata)

Hesaplayalım

Hata metrikleri kodumuz:

```python
# Hata Metrikleri
print("\nBasit ve Çoklu Lineer Regresyon Hata Metrikleri ")
mse = mean_squared_error(y_test, y_pred)
print(f"\nBasit Lineer Regresyon Mean Squared Error (MSE): {mse:.4f}")
mae = mean_absolute_error(y_test, y_pred)
print(f"Basit Lineer Regresyon Mean Absolute Error (MAE): {mae:4f}")

mse1 = mean_squared_error(new_y_test, y_pred1)
print(f"Çoklu Lineer Regresyon Mean Squared Error (MSE): {mse1:.4f}")
mae1 = mean_absolute_error(new_y_test, y_pred1)
print(f"Çoklu Lineer Regresyon Mean Absolute Error (MAE): {mae1:4f}")
```

Çıktısı:

```
Basit ve Çoklu Lineer Regresyon Hata Metrikleri

Basit Lineer Regresyon Mean Squared Error (MSE): 3978.3852
Basit Lineer Regresyon Mean Absolute Error (MAE): 53.206543
Çoklu Lineer Regresyon Mean Squared Error (MSE): 3205.5570
Çoklu Lineer Regresyon Mean Absolute Error (MAE): 45.507007
```

---

5. **Yorumlayın:**
    - **Hangi model daha başarılı? Neden?**

      Çoklu lineer regresyon modeli genellikle daha başarılıdır çünkü bu modelde daha fazla bağımsız değişken kullanılır. Basit lineer regresyonda ise yalnızca bir bağımsız değişken kullanılır. Basit regresyonda az sayıda bağımsız değişken olması, modelin sağlıklı veriyle çalışmamasına yol açabilir. Ancak, bu durum bazı özel koşullarda değişebilir; modelin başarısı genellikle veriyle ilişkilidir. Eğer bağımsız değişkenler birbirleriyle çok alakasızsa, çoklu regresyonun faydası az olur.

    - **R^2 değerleri ne ifade ediyor?**

      Basit lineer regresyon r2 değeri => 0.3453

      Çoklu lineer regresyon r2 değeri => 0.4725

      Yüksek bir R2 değeri, modelin iyi çalıştığını gösterir. Yukarıdaki rakamlara bakarsak, basit modelin R2 değeri daha düşük, çoklu modelin ise daha yüksek bir R2 değeri vardır. Bu da, çoklu modelin basit modele göre çok daha iyi bir **uyum sağladığını** gösterir, çünkü çoklu model daha fazla bağımsız değişken içerir ve veriye daha fazla uyum sağlar.
