Report on Data Analysis Using Pandas

Name: Aditi Singh

Reg no: 24BCE1 4488

GitHub link:
https://github.com/

Aim:- Doing Linear Regression analysis, Logistic Regression, KNN, Classification,

Prediction ,Clustering ,Seeding, with 5 scores.

Dataset Overview:-

This dataset provides a detailed view of student lifestyle patterns and their correlation with academic performance, represented by GPA. It contains 5 records of students' daily habits across study, extracurriculars, sleep, socializing, and physical activities. Each student's stress level is derived based on study and sleep hours, offering insights into how lifestyle factors may impact academic outcomes.

Program:

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression, LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import mean_squared_error, accuracy_score
from sklearn.cluster import KMeans
from sklearn.preprocessing import LabelEncoder

# Manually inputting the dataset based on the provided data
data_dict = {
    "Study_Hours_Per_Day": [6.9, 5.3, 5.1, 6.5, 8.1],
    "Extracurricular_Hours_Per_Day": [3.8, 3.5, 3.9, 2.1, 0.6],
    "Sleep_Hours_Per_Day": [8.7, 8.0, 9.2, 7.2, 6.5],
    "Social_Hours_Per_Day": [2.8, 4.2, 1.2, 1.7, 2.2],
    "Physical_Activity_Hours_Per_Day": [1.8, 3.0, 4.6, 6.5, 6.6],
    "GPA": [2.99, 2.75, 2.67, 2.88, 3.51],
    "Stress_Level": ["Moderate", "Low", "Low", "Moderate", "High"]
}
df = pd.DataFrame(data_dict)

# Splitting data into features (X) and targets (y)
X = df[["Study_Hours_Per_Day", "Extracurricular_Hours_Per_Day", "Sleep_Hours_Per_Day",
        "Social_Hours_Per_Day", "Physical_Activity_Hours_Per_Day"]]
y_gpa = df["GPA"]
y_stress = df["Stress_Level"]
```

```
# Encoding Stress_Level for classification
le = LabelEncoder()
y_stress_encoded = le.fit_transform(y_stress)

# Splitting data for regression and classification tasks
X_train, X_test, y_gpa_train, y_gpa_test = train_test_split(X, y_gpa, test_size=0.4, random_state=42)
_, _, y_stress_train, y_stress_test = train_test_split(X, y_stress_encoded, test_size=0.4, random_state=42)

# 1. Linear Regression (Predict GPA)
lr = LinearRegression()
lr.fit(X_train, y_gpa_train)
y_gpa_pred = lr.predict(X_test)
linear_regression_mse = mean_squared_error(y_gpa_test, y_gpa_pred)

# 2. Logistic Regression (Classify Stress_Level)
log_reg = LogisticRegression(max_iter=1000)
log_reg.fit(X_train, y_stress_train)
y_stress_pred_logistic = log_reg.predict(X_test)
logistic_regression_accuracy = accuracy_score(y_stress_test, y_stress_pred_logistic)

# 3. KNN (Classify Stress_Level)
knn = KNeighborsClassifier(n_neighbors=3)
knn.fit(X_train, y_stress_train)
y_stress_pred_knn = knn.predict(X_test)
knn_accuracy = accuracy_score(y_stress_test, y_stress_pred_knn)
```

```
# 4. Clustering (KMeans for grouping)
kmeans = KMeans(n_clusters=3, random_state=42)
clusters = kmeans.fit_predict(X)
cluster_centers = kmeans.cluster_centers_

# Print the results
print("Linear Regression Mean Squared Error (MSE):", linear_regression_mse)
print("Logistic Regression Accuracy:", logistic_regression_accuracy)
print("KNN Accuracy:", knn_accuracy)
print("KMeans Clusters:", clusters)
print("KMeans Cluster Centers:\n", cluster_centers)
```

Output:

```
Linear Regression Mean Squared Error (MSE): 0.2553622349693233
Logistic Regression Accuracy: 0.0
KNN Accuracy: 0.0
KMeans Clusters: [2 0 0 1 1]
KMeans Cluster Centers:
 [[5.2  3.7  8.6  2.7  3.8 ]
 [7.3  1.35 6.85 1.95 6.55]
 [6.9  3.8  8.7  2.8  1.8 ]]
C:\Users\poona\anaconda3\Lib\site-packages\sklearn\cluster\_kmeans.py:1446: UserWarning:
KMeans is known to have a memory leak on Windows with MKL, when there are less chunks
than available threads. You can avoid it by setting the environment variable
OMP_NUM_THREADS=1.
  warnings.warn(
```

Explanation:

Analysis Report

1. Linear Regression Analysis

Objective:
To predict the GPA of students using daily habits (e.g., study hours, extracurricular hours, sleep, social hours, and physical activity).

Model:
Linear Regression

Performance Metric:

- Mean Squared Error (MSE): *Value from your output*

Interpretation:
The MSE indicates the average squared difference between the predicted GPA values and the actual GPA

values. A lower MSE represents better accuracy. Given the small dataset, this result provides preliminary insights into how daily habits influence GPA.

---

2. Logistic Regression Analysis

Objective:
To classify the Stress Level of students into categories (Low, Moderate, High) based on their daily habits.

Model:
Logistic Regression

Performance Metric:

- Accuracy: *Value from your output*

Interpretation:
Accuracy measures the proportion of correctly predicted stress levels compared to the total number of predictions. A high accuracy value indicates that the model effectively identifies stress levels based on the input features. However, due to the small dataset, results may be biased and should be validated with more data.

---

3. K-Nearest Neighbors (KNN) Analysis

Objective:
To classify the Stress Level of students similarly to Logistic Regression but using a different approach (KNN).

Model:
K-Nearest Neighbors (k=3)

Performance Metric:

- Accuracy: *Value from your output*

Interpretation:
KNN identifies stress levels based on the similarity between students' habits. The accuracy score shows how well this approach works compared to Logistic Regression. Any differences in performance highlight the strengths and weaknesses of the methods for this dataset.

---

4. Clustering Analysis

Objective:
To group students into clusters based on their daily habits, identifying patterns without predefined categories.

Model:
KMeans Clustering (k=3)

Results:

- Cluster Assignments: [Cluster numbers from your output, e.g., 0, 1, 2 for each student]
- ClusterCenters:
  Values from your output representing the average features of students in each cluster.

Interpretation:
The clustering groups students with similar daily habits into the same cluster. For example, one cluster might represent students with high physical activity and low stress, while another could represent students with moderate study hours and high stress levels. This unsupervised learning helps identify trends in student behaviors.

Conclusion

This analysis demonstrates the relationships between daily habits and GPA, stress levels, and potential behavior patterns.

- Linear Regression indicates how well daily habits predict academic performance (GPA).

- Logistic Regression and KNN classify stress levels, with varying performance metrics.
- KMeans Clustering reveals distinct behavioral groups among students, offering insights into lifestyle patterns.