Assignment No: 2

Name: Kush Parihar

Class: LY-ISA 1 Batch- A

Roll no: 2203023

Problem Statement:

To implement k-means algorithm and show the results graphically using R programming. (here the students are expected to use R programming and comment on the value of k selected for the application and analyze the output. (why and how the k value is selected)) Objective:

The main objective of this assignment is to apply big data analysis techniques using Python to derive valuable insights from the provided dataset. You will explore both regression and classification tasks, selecting appropriate models, analyzing their performance, and drawing meaningful conclusions.

Theory:

1. Discovery Phase: Identify the problem you want to solve (regression or classification). Understand the dataset, its features, and the target variable. Define the objectives and success criteria for your analysis.
2. Data Preparation Phase: Load the dataset using Pandas or other data loading libraries. Handle missing values, outliers, and perform data cleansing. Split the dataset into features (X) and target variable (y).
3. Model Planning Phase: Determine the regression or classification models you want to use (e.g., Linear Regression, Decision Trees, Random Forest, Logistic Regression, Support Vector Machines, etc.). Set up the evaluation metrics relevant to your task (e.g., Mean Absolute Error for regression, Accuracy for classification). Split the data into training and testing sets using libraries like train_test_split from scikit-learn.
4. Model Building Phase: Create instances of the chosen regression or classification models using scikit-learn. Fit the models to the training data using the .fit() method. For classification, consider handling class imbalances using techniques like oversampling or undersampling.
5. Communicate Results Phase: Evaluate the model's performance on the testing set using the chosen evaluation metrics. Generate performance reports, confusion matrices, ROC curves, etc. Visualize predictions and actual values for regression, or confusion matrices for classification.
6. Operationalize Phase: Once satisfied with model performance, use the trained model to make predictions on new data. Integrate the model into applications, websites, or services using appropriate APIs. Regularly monitor the model's performance and retrain as needed.

Assignment Phases:

Data Loading and Preprocessing: Load the dataset using Python and relevant libraries. Perform data preprocessing steps including handling missing values, data transformation, and normalization.

Exploratory Data Analysis (EDA): Visualize data distributions, correlations, and patterns using Python libraries like Matplotlib and Seaborn. Interpret the insights obtained from EDA.

Regression Analysis: Choose a specific regression model based on the dataset and problem statement. Split the dataset into training and testing sets. Train the selected regression model on the training data. Evaluate the model's performance using appropriate regression metrics (e.g., Mean Squared Error, R-squared).

Classification Analysis: Select a classification model suited for the dataset. Prepare the data for classification analysis. Train the selected classification model on the training data. Evaluate the model using classification metrics (e.g., accuracy, precision, recall).

Code:

```
import pandas as pd
from sklearn.model_selection import train_test_split from
sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import mean_absolute_error, accuracy_score import numpy as np
import matplotlib.pyplot as plt import
seaborn as sns
```

```
kush=pd.read_excel('/content/EastWestAirlines.xlsx')
```

```
kush.head(10)
```

	ID#	Balance	Qual_miles	cc1_miles	cc2_miles	cc3_miles	Bonus_miles	Bonus_trans	Flight_miles_12mo	Flight_trans_12	Days_since_enrol
0	1	28143	0	1	1	1	174	1	0	0	700
1	2	19244	0	1	1	1	215	2	0	0	696
2	3	41354	0	1	1	1	4123	4	0	0	703
3	4	14776	0	1	1	1	500	1	0	0	695
4	5	97752	0	4	1	1	43300	26	2077	4	693
5	6	16420	0	1	1	1	0	0	0	0	694
6	7	84914	0	3	1	1	27482	25	0	0	699
7	8	20856	0	1	1	1	5250	4	250	1	693
8	9	443003	0	3	2	1	1753	43	3850	12	694
9	10	104860	0	3	1	1	28426	28	1150	3	693

```
X=kush.drop('cc2_miles',axis=1) Y
=kush.cc2_miles

X_train, X_test, Y_train, Y_test =train_test_split(X, Y, test_size=0.2)

# model building - regression
RegModel = LinearRegression()
RegModel.fit(X_train,Y_train)
RegPred= RegModel.predict(X_test)
RegMAE= mean_absolute_error(Y_test,RegPred)

print(f'Regression MAE: {RegMAE}')
```

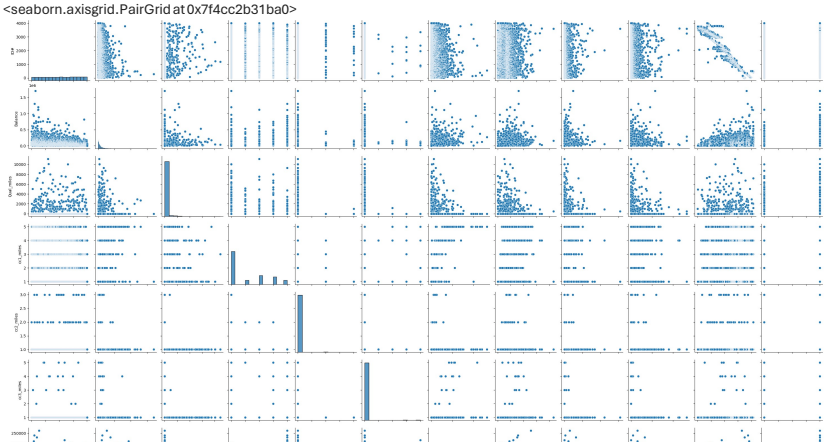
```
Regression MAE: 0.03834978953663382
```

```
x=kush.drop('cc2_miles', axis=1) y=
kush.cc2_miles

X_train, X_test, Y_train, Y_test =train_test_split(X, Y, test_size=0.2, random_state=42)

import pandas as pd
fromsklearn.model_selectionimporttrain_test_split from
sklearn.ensemble import RandomForestClassifier from
sklearn.metrics import  accuracy_score

clf_model=RandomForestClassifier() clf_model.fit(X_train,Y_train)
clf_pred = clf_model.predict(X_test)
clf_accuracy= accuracy_score(Y_test,clf_pred)

print(f"classificationAccuracy:{clf_accuracy}") classification
```

```
Accuracy:0.99375
```

```
importmatplotlib.pyplotasplt import
seaborn as sns

sns.pairplot(komal)
```

<seaborn.axisgrid.PairGrid at 0x7f4cc2b31ba0>



Conclusion:

In this assignment, you demonstrated your ability to perform comprehensive Big Data Analysis using Python. Through data preprocessing, exploratory analysis, and the application of Regression and Classification techniques, you gained insights from the dataset and made predictions based on the models' outcomes.