LM ASSIGNMENT
NAME : ANKITA PRAKASH
PRN NO : 23060641003

Problem Statement:

The project aimed at predicting the chance of admission to a university based on various features such as GRE score, TOEFL score, university rating, statement of purpose (SOP), letter of recommendation (LOR), CGPA, and research experience.

Data Set Link:

https://docs.google.com/spreadsheets/d/1jCf-alwxuwtUW-ogxbBVNn7a5A3zbT7y/edit?usp=sharing&ouid=106375008812993682809&rtpof=true&sd=true

Description For Code:
This code appears to be a data analysis and machine learning pipeline for predicting the chances of admission for students based on various features. Here's a breakdown of the code:

1. Data Loading and Overview:
 - The code starts by importing necessary libraries like pandas, numpy, seaborn, matplotlib, and scikit-learn.
 - A CSV file containing data about students' GRE scores, TOEFL scores, university ratings, etc., is loaded into a pandas DataFrame.
 - Basic information about the DataFrame such as its shape and data types of columns is displayed.

2. Data Preprocessing:
 - Serial numbers are dropped as they are not relevant features for prediction.
 - Some column names are renamed for clarity.
 - Categorical columns are converted to the categorical data type.
 - Missing values, outliers, and duplicates are checked and handled if necessary.

3. Exploratory Data Analysis (EDA):
 - Pairplots, correlation heatmaps, boxplots, histograms, and countplots are created to understand the relationships between different features and the target variable.

4. Label Encoding & Standardization:
 - Categorical variables are label encoded.
 - Numerical variables are standardized using Min-Max Scaling.

5. Model Building:
 - A Linear Regression model is chosen for prediction.
 - The data is split into training and testing sets.
 - The model is trained on the training data and evaluated on both training and testing data.

6. Model Evaluation:
 - Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R-squared (R2) score, and Adjusted R-squared (Adj R2) score are calculated to evaluate the model's performance.

Overall, the code performs data preprocessing, exploratory data analysis, model building, and evaluation steps for predicting admission chances based on student profiles.

Interpretation of Model:

1. Mean Absolute Error (MAE):
 - MAE measures the average absolute difference between the predicted and actual values.
 - In this model, the MAE on the training data is approximately 0.04 and on the test data is also approximately 0.04.
 - This means, on average, the model's predictions are off by around 0.04 in terms of the chance of admission.

2. Root Mean Squared Error (RMSE):
 - RMSE is the square root of the average squared differences between the predicted and actual values.
 - In this model, the RMSE on the training data is approximately 0.06 and on the test data is also approximately 0.06.
 - RMSE gives a similar interpretation to MAE but penalizes larger errors more heavily.

3. R-squared (R2) Score:
 - R-squared measures the proportion of the variance in the dependent variable (chance of admission) that is predictable from the independent variables.
 - In this model, the R2 score on the training data is approximately 0.82, indicating that around 82% of the variance in the chance of admission is explained by the independent variables included in the model.
 - Similarly, the R2 score on the test data is also approximately 0.82.

4. Adjusted R-squared (Adj R2) Score:
 - Adjusted R-squared adjusts the R2 score for the number of predictors in the model.
 - It penalizes the addition of unnecessary predictors that do not improve the model significantly.
 - In this model, the adjusted R2 score on both training and test data is approximately 0.81, indicating that the model is robust and the included predictors are relevant.

Overall, the Linear Regression model seems to perform well in predicting the chances of admission based on the given features, as indicated by the evaluation metrics. However, further analysis and model refinement might be needed for real-world deployment.

Model Adequacy test :
The model adequacy test evaluates how well the linear regression model fits the data and whether the assumptions of linear regression are met. Here are some common tests:

1. Residual Analysis: This involves examining the residuals (the differences between the observed and predicted values). Residual plots should be random and homoscedastic (equal variance).

2. Normality Test: It checks if the residuals are normally distributed. This can be done using statistical tests like the Shapiro-Wilk test or by visual inspection using a QQ plot.

3. Homoscedasticity Test: This tests whether the residuals have constant variance across different levels of the predictor variables. It can be checked by plotting residuals against fitted values or predictor variables.

4. Multicollinearity Test: This assesses if there are high correlations between predictor variables, which can affect the model's stability and interpretability. Variance inflation factor (VIF) is a common metric used for this purpose.

5. Independence of Errors: This assumption implies that the residuals are independent of each other. Durbin-Watson test is often used to check for autocorrelation in the residuals.

Model Summary:

- Model Type: Linear Regression
- Features: GRE Score, TOEFL Score, University Rating, SOP, LOR, CGPA, Research
- Target Variable: Chance of Admit
- Model Performance on Training Data:
  - Mean Absolute Error (MAE): 0.04
  - Root Mean Squared Error (RMSE): 0.06
  - R-squared Score (R2): 0.82
  - Adjusted R-squared: 0.82
- Model Performance on Test Data:
  - Mean Absolute Error (MAE): 0.04
  - Root Mean Squared Error (RMSE): 0.06
  - R-squared Score (R2): 0.82
  - Adjusted R-squared: 0.81

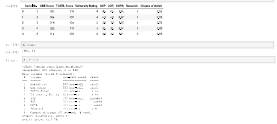
These metrics indicate that the linear regression model performs well in predicting the chance of admission based on the given features. However, further analysis such as residual diagnostics and assumption tests should be conducted to ensure the model's adequacy and reliability.

Conclusion:

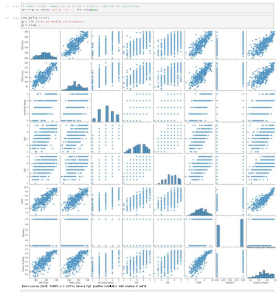
Since there is no difference in the loss scores of training and test data, we can conclude that there is no overfitting of the model

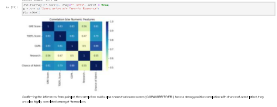
- Mean Absolute Error of 0.04 shows that on an average, the absolute difference between the actual and predicted values of chance of admit is 4%
- Root Mean Square Error of 0.06 means that on an average, the root of squared difference between the actual and predicted values is 6%
- R2 Score of 0.82 means that our model captures 82% variance in the data
- Adjusted R2 is an extension of R2 which shows how the number of features used changes the accuracy of the prediction

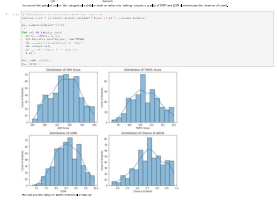
Python File :

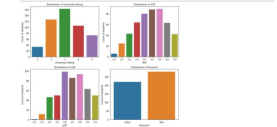


Exploratory Data Analysis









Data Preprocessing

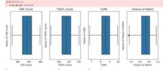
```
#Check for missing values in all columns
df.isna().sum()
```

```
GRE Score            0
TOEFL Score          0
University Rating    0
SOP                  0
LOR                  0
CGPA                 0
Research             0
Chance of Admit      0
dtype: int64
```

There are no missing values in the dataset

```
# Check for outliers in numerical columns
[illegible]
```

```
[illegible]
```



It can be observed that there are no outliers in the numeric columns (all the observations are within the whiskers which represent the minimum and maximum of the range of values)

```
# Check for Duplicate rows
[illegible]
```

```
(0, 8)
```

There are no duplicate rows in the dataset

Train-Test Split

```
[illegible]
```

```
# Separate predictor and target variables
[illegible]
```

```
X.head()
```

	GRE Score	TOEFL Score	CGPA	University Rating	SOP	LOR	Research
0	337	118	9.65	4	4.5	4.5	True
1	324	107	8.87	4	4.0	4.5	True
2	316	104	8.00	3	3.0	3.5	True
3	322	110	8.67	3	3.5	2.5	True
4	314	103	8.21	2	2.0	3.0	False

```
y.head()
```

	Chance of Admit
0	0.92
1	0.76
2	0.72
3	0.80
4	0.65

```
# Split the data into training and test sets
[illegible]
```

```
[illegible]
```

Label Encoding & Standardisation

```
[illegible]
```

```
[illegible]
```

```
[illegible]
```

```
[illegible]
```

	GRE Score	TOEFL Score	CGPA	University Rating	SOP	LOR	Research
249	321	111	8.83	2	6	8	1
433	316	111	8.54	3	6	8	0
19	303	102	8.50	2	6	4	0
322	314	107	8.27	1	5	6	0
332	308	106	8.21	2	6	5	1
56	316	102	7.40	2	2	7	0
301	319	108	8.76	1	5	4	0
229	324	111	9.01	3	4	4	1
331	311	105	8.12	1	4	2	1
132	309	105	8.56	4	5	5	0

```
[illegible]
```

```
[illegible]
```

```
MinMaxScaler()
```

```
[illegible]
```

```
[illegible]
```

```
[illegible]
```

	GRE Score	TOEFL Score	CGPA	University Rating	SOP	LOR	Research
361	0.86	0.857143	0.878205	0.75	0.750	0.625	1.0
73	0.48	0.571429	0.717949	0.75	0.875	0.750	1.0
374	0.50	0.464286	0.272436	0.75	0.250	0.375	0.0
155	0.44	0.607143	0.605769	0.50	0.500	0.500	0.0
104	0.72	0.714286	0.721154	0.50	0.625	0.500	1.0

Base Model: Linear Regression

```
[illegible]
```

```
[illegible]
```

```
LinearRegression()
```

```
[illegible]
```

```
[illegible]
```

```
[illegible]
```

```
[illegible]
```

```
[illegible]
```

```
[illegible]
```