Report Date: 21st June

LEARNINGS:

EXPLORATORY DATA ANALYSIS:

Basic understanding of EDA:

The eda is basically a preliminary step performed before feeding the data to the ml algorithms. it can be simply coined as a step in data preprocessing. The eda would be helping in normalizing the data by removing the outliers and also filling up the missing values. It would also help in getting the visualization of the data by using libraries such as matplotlib.

Eda also help in understanding the relationship between the variables, we retrieve the relationship understanding by using various ML algorithms such as linear regression and logistic regression.

Eda can be classified into three types:

univariate-only visualization.
bivariate- trying to establish the relationship between the two variables.
multivariate-trying to establish the relationship between the two or more variables.(dimensionality reduction is by pca is observed in multivariate).

The library that are used are:
pandas, matplotlib, seaborn ,ploty.

-->helps in uncovering the underlying patterns.
-->highlight relationship between data points
-->helps restructure data for modeling
-->helps in removing redundancies form the data

python-matplotlib,seaborn,numpy,pandas.
r-tidyr,deplyr,ggplot2.

-->Pandas Library:

The main function of the pandas is data manipulation ,it is mainly used in data cleaning and the helps in filling the missing data

Example code using pandas:

```
Import pandas as pd

#retrieving the dataset
df=pd.read._csv('data.csv')

#used to describe the features of the dataset ,in simple words the columns and their datatypes
print(df.describe())

Df.fillna(0, inplace=True)
```

—>Numpy Library:

The bumpy library is mainly used for the handling large and multiple dimensional datasets. It helps in handling large datasets and array manipulation.

Example code using numpy:

```
Import numpy as np


arr=np.array([1,2,3,4,5])

#performing the math matical operations scubas obtaining the means and the standard deviation.
mean=np.mean(arr)
std_dev=np.std(arr)
```

—>Seaborn Library:

The seaboard library plays thekeyrole in visualising the data it helps in creating the histograms, box plots, the graphs can also be customised as per the requierement.

—>Matplot Library:

The marplotlib is essential for creating the visualisations, the seaborn library is built upon matplotlib for advanced visualisations by establishing relationships between the variables such as line graphs, clustering graphs etc.

Example code using matplotlib:

```
import matplotlib.pyplot as plt

plt.plot([1, 2, 3, 4], [1, 4, 9, 16])
plt.xlabel('X-axis')
plt.ylabel('Y-axis')
plt.title('Simple Plot')
plt.show()
```

These libraries potentially help in data cleaning and manipulation, data visualisation and statical analysis.

Different Algorithms used Internally as part of EDA:

The algorithms part of Supervised Learning:

Linear regression:
The logistic regression has two tyes of variables
1)Dependent variable
2) Independent Variable

-->By analyzing the data we would be able to predict the and train the Ml model in such a way, so that when unknown value of x is given the value of the y can be presicted.

Working of the linear Regression:

-->The linear regression would try to find the mean of the points and establish the a line through the it actulaly then tries to create an functional approximation internally and when there is a new value given that functional approximation genrated formula works in place.

There are sevreal types of linear regression such as:
1) Single Linear Regression(Single independent variable)
2) multi Linear Regression(Multiple Independent Variables)
3) polynomial Linear Regression(used when the degree of x is equal to y.

Logistic Regression:
The logistic Regression basically works on the basis of the linear regression. from the linear regression graph a part is taken as a line of best fit. Also called as the sigmoid curve. The sigmoid curve is basiacally taken from 0 to 1 ,to use the concept of the probability.

The logistic regression just helps in dterming just yes or no. The logistic Regression heps in determing Whether True or False by considering the probability approach. If the probability is greather than 50% then it tends to be true or else it is considered as false.

Summarization of Today's Learning:

Today, I gained a basic understanding of libraries like Pandas, NumPy, Matplotlib, and Seaborn for data manipulation and visualization in Exploratory Data Analysis (EDA). I also learned about normalizing data, removing outliers, and filling missing values as essential steps in data preprocessing. Additionally, I explored the concepts of univariate, bivariate, and multivariate analysis to understand relationships between variables for machine learning preparation.