Project Members:

Taha Abbas Ali

Muhammad Talha

Date: 25-12-2023

Section: 7A BCS

Course: Data Science

Course Lecturer: *Dr. Musadaq Mansoor*

Let us describe our actions we perform in this project

Data Loading and Initial Inspection

Imported necessary libraries for data manipulation and visualization.
Read the dataset 'edu_dataset.csv' into a Pandas DataFrame (df).
Displayed the first 10 rows of the dataset to get an overview of the data.

Data Exploration and Insights

Provided a detailed description of the columns in the dataset, explaining the meaning of each feature.
For ease here it go,

1.Gender – Male, Female.
2.Home Location – Rural, Urban
3.Level of Education – Post Graduate, School, Under Graduate
4.Age – Years
5.Number of Subjects – 1- 20
6.Device type used to attend classes – Desktop, Laptop, Mobile
7.Economic status – Middle Class, Poor, Rich
8.Family size – 1 -10
9.Internet facility in your locality – Number scale (Very Bad to Very Good)
10.Are you involved in any sports? – Yes, No
11.Do elderly people monitor you? – Yes, No
12.Study time – Hours
13.Sleep time – Hours
14.Time spent on social media – Hours
15.Interested in Gaming? – Yes, No
16.Have separate room for studying? – Yes, No
17.Engaged in group studies? – Yes, No
18.Average marks scored before pandemic in traditional classroom – range
19.Your interaction in online mode - Number scale (Very Bad to Very Good)
20.Clearing doubts with faculties in online mode - Number scale (Very Bad to Very Good)
21.Interested in? – Practical, Theory, Both
22.Performance in online - Number scale (Very Bad to Very Good)
23.Your level of satisfaction in Online Education – Average, Bad, Good

Checked the shape of the dataset (number of rows and columns).
Checked the size of the dataset (total number of values).
Checked for duplicate values in the dataset.
Checked for null values in each column.
Displayed general information about the dataset using df.info().
Displayed descriptive statistics of the numerical columns using df.describe().
Displayed descriptive statistics of the categorical columns using df.describe(include='object').

Data Preprocessing
Removed unimportant columns ('Home Location', 'Economic status', 'Do elderly people monitor you?', 'Engaged in group studies?', 'Have separate room for studying?').
Selected only numeric columns for correlation analysis.
Computed and visualized the correlation matrix using a heatmap.

Data Profiling
Created a function (data_proflileing) to generate a profile of the dataset, including data types, unique values, null values, maximum and minimum values, and duplicates for each column.

Data Visualization
Visualized the distribution of numerical features using boxplots, revealing potential outliers.

Outlier Handling
Detected and handled outliers in the 'Age(Years)' and 'Number of Subjects' columns using the Interquartile Range (IQR) method.
Visualized outliers before and after the outlier removal process using boxplots.

Exploratory Analysis & Visualizations - EDA

1. Gender Impact on Online Education
Analyzed the distribution of genders in the dataset using a pie chart.
Visualized the percentage of male and female students.
Insight: Male students exhibit higher adaptability to online education (59.4%) compared to females (40.6%).

2. Age Group Impact on Online Education
Explored the age distribution of students using a count plot.
Insights:
Students aged 17-21 show a higher preference for online education.
Interest in online education decreases for people aged 22 and above.

3. Education Level Impact on Online Education
Investigated the distribution of education levels using a histogram.
Insight: Undergraduate students comprise a larger population compared to postgraduate and school students.

4. Device Type Impact on Online Education
Examined the distribution of device types used for attending classes using a count plot.
Insight: Laptops and mobile devices are more preferred over desktops.

5. Average Marks Scored Before Pandemic
Visualized the distribution of average marks scored before the pandemic using a count plot.

6. Time Spent on Study
Examined the distribution of study time using a count plot.
Insight: The average time spent on studies is 3-4 hours.

7. Students' Interest in Practical vs. Theory
Used a count plot to visualize students' preferences for practical, theory, or both.

8. Level of Satisfaction in Online Education
Visualized the distribution of satisfaction levels using a pie chart.

Common Visualizations
Created distribution plots, scatter plots, line plots, and bar plots to visualize various aspects of the dataset.
Displayed distributions of device types, scatter plots for age vs. study time, and time spent on social media vs. study time.
Visualized the line plot for the interaction in online mode over the level of satisfaction.

Data Preprocessing for Machine Learning
Applied label encoding to categorical columns using LabelEncoder.
Split the dataset into features (X) and the target feature (y).
Performed standard scaling on the features using StandardScaler.
Split the data into training and testing sets using train_test_split.

Supervised Machine Learning Models (Random Forest, K-Nearest Neighbors)
Trained a Random Forest Classifier and a K-Nearest Neighbors Classifier.
Evaluated the models using accuracy, precision, recall, and F1-score.
Visualized the model evaluation metrics using bar plots.

Un-Supervised K-Means Clustering Model
Utilized the elbow method to determine the optimal number of clusters (k).
Applied K-Means clustering with the optimal k.
Added cluster information to the original DataFrame.
Visualized the clusters using a scatter plot.