STUDENT DROPOUT RISK ANALYSIS

Objective: To identify the factors affecting students to drop out and predict target variable.

Key Responsibilities:

- To collect , clean, structure, validate and visualize the data.
- Find the factors affecting the target variable.
- To fit best machine learning model to predict the target variable.

Data Source:

The dataset was collected from Kaggle.

Importing Libraries:

Imported required libraries to google colab notebook.

Understanding Data and Cleaning Data:

Data Variables:

There are 36 variables and 44 instances.

With variables like Marital status, Age at enrollment, Gender, Father's and Mother's Occupation, Curricular Units in 1st sem (approved), Curricular Units in 1st sem (Grade), Scholarship holder etc. as independent variables and Target as dependent variable and predictable variable.

Refer
https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

Performed summary statistics to understand the data mean, median, quantiles, count.

Checked for datatypes in the data.

Checked for missing values and null values. There were no such values.

Transforming or Structuring Data:

Used heatmap of correlation matrix to understand the relationship between the variables. The range of correlation lies between -1 -to 1. If the value is nearer to '0' then it would not affect much for a target variable and if the value is closer to -1 or 1 then it would affect much, and we consider it as correlated variables.

Removed columns with less correlation value and constructed new dataset with these 'Target', 'Tuition fees up to date', 'Scholarship holder', 'Curricular units 1st sem (approved)', 'Curricular units 1st sem (grade)', 'Curricular units 2nd sem (approved)', 'Curricular units 2nd sem (grade)', 'Debtor', 'Gender', 'Age at enrolment ' variables.

Data Visualizations:

- Plotted boxplot to find the outliers and analyzed whether those outliers affect them or not by comparing the variable's mean and median. And concluded those outliers don't affect the data.
- Plotted pie chart for target variable and observed dropout, enrolled, and graduate percentages are 32.1, 17.9, 49.9 respectively.
- Pair plot to understand the distribution of variables.
- Histogram to find the count of age at enrollment.

These plots were employed by using python data visualization tools.

Building Machine Learning Models:

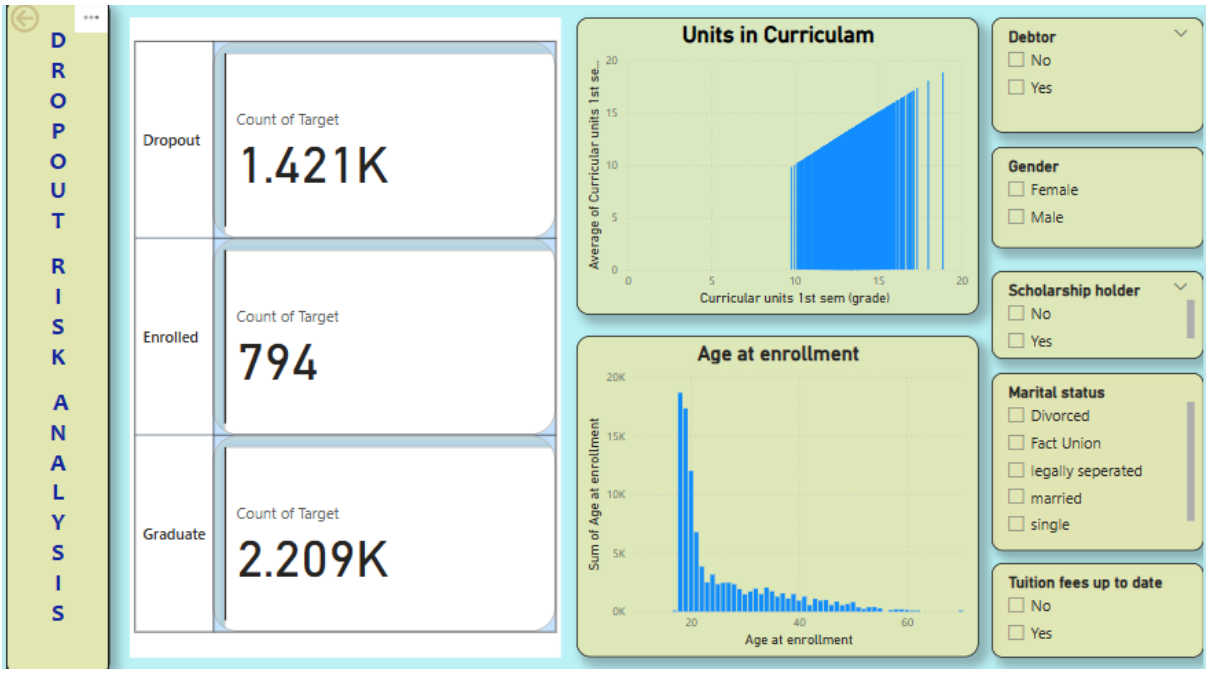
Splitted the data into train and test splitting the ratio 7:3 .

Accuracy Percentages of models:

- Logistic Regression: 73.04%
- Decision Tree Classifier: 66.64%
- Random Forest Classifier: 74.77%
- K-Nearest Neighbors Classifier: 72.62%
- Support Vector Machine Classifier: 74.1%
- Bagging Classifier: 70.26%
- Ada Boosting Classifier: 74.62%
- XG Boosting Classifier: 74.92%

Conclusion: XG boosting classifier has 74.92 % accuracy by models prediction.

Data Visualization through Power BI:



Transformed data and encoded numerical variables into categorical variables and build a dashboard**.**

Conclusion:

Age, scholarship holder, curriculum units affect students to dropout.

As the age at the time of enrollment increases, the dropout rate also increases.

Recommendations:

As there are a greater number of units(more than 10) in a curriculum in a single semester, it becomes hard for students to obtain good grades and to grasp the content. So, the units in a curriculum should be decreased.

Scholarships should be provided to students to decrease the dropout rate.