MACHINE LEARNING PROJECT

PROBLEM STATEMENT:-

The given dataset contains students' data, and we need to predict whether a given student has dropped out, graduated, or is still enrolled in the course.

DATASET DESCRIPTION AND PREPROCESSING:-

The given dataset has the following columns:-
['Marital_Status_Code', 'Application_Method', 'Application_Sequence', 'Attendance_Type', 'Prior_Qualification_Code', 'Prior_Qualification_Score', 'Nationality_Code', "Mother's_Education_Level", "Father's_Education_Level", "Mother's_Job_Category", "Father's_Job_Category", 'Admission_Score', 'Student_Displacement_Flag', 'Special_Educational_Needs', 'Outstanding_Debts_Flag', 'Tuition_Fees_UpToDate_Flag', 'Gender_Code', 'Scholarship_Recipient_Flag', 'Enrollment_Age', 'International_Status', 'Credits_1st_Semester', 'Enrolled_1st_Semester', 'Evaluations_1st_Semester', 'Passed_1st_Semester', 'Grade_1st_Semester', 'No_Evaluations_1st_Semester', 'Credits_2nd_Semester', 'Enrolled_2nd_Semester', 'Evaluations_2nd_Semester', 'Passed_2nd_Semester', 'Grade_2nd_Semester', 'No_Evaluations_2nd_Semester', 'Local_Unemployment_Rate', 'Inflation_Rate', 'Regional_GDP', 'Outcome']

In the given columns, ‘Outcome’ is our target column, and the rest are our features.

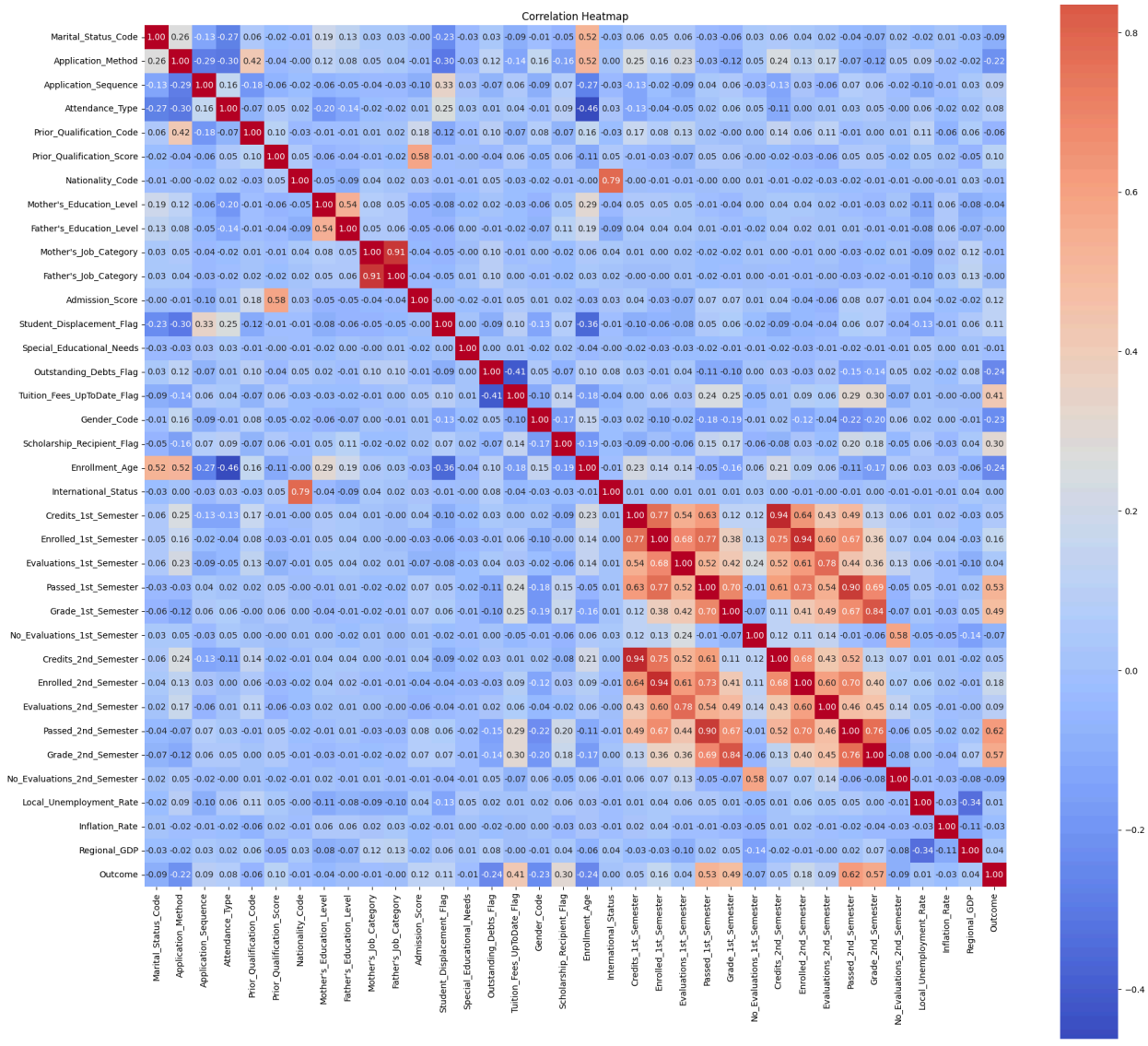
The given dataset has no null values for any of the features, but there is a slight imbalance for our target:-

Graduated - 49%
Dropped out - 32%
Enrolled - 18%

We use an Ordinal Encoder to encode the Outcome

Now with the target encoded, we plot the correlation matrix. The higher the correlation between a feature and the outcome, the better that feature contributes to the predictive power of our Machine Learning Model. The colour of a cell across a row and column depicts the value.

Correlation matrix Plot:-



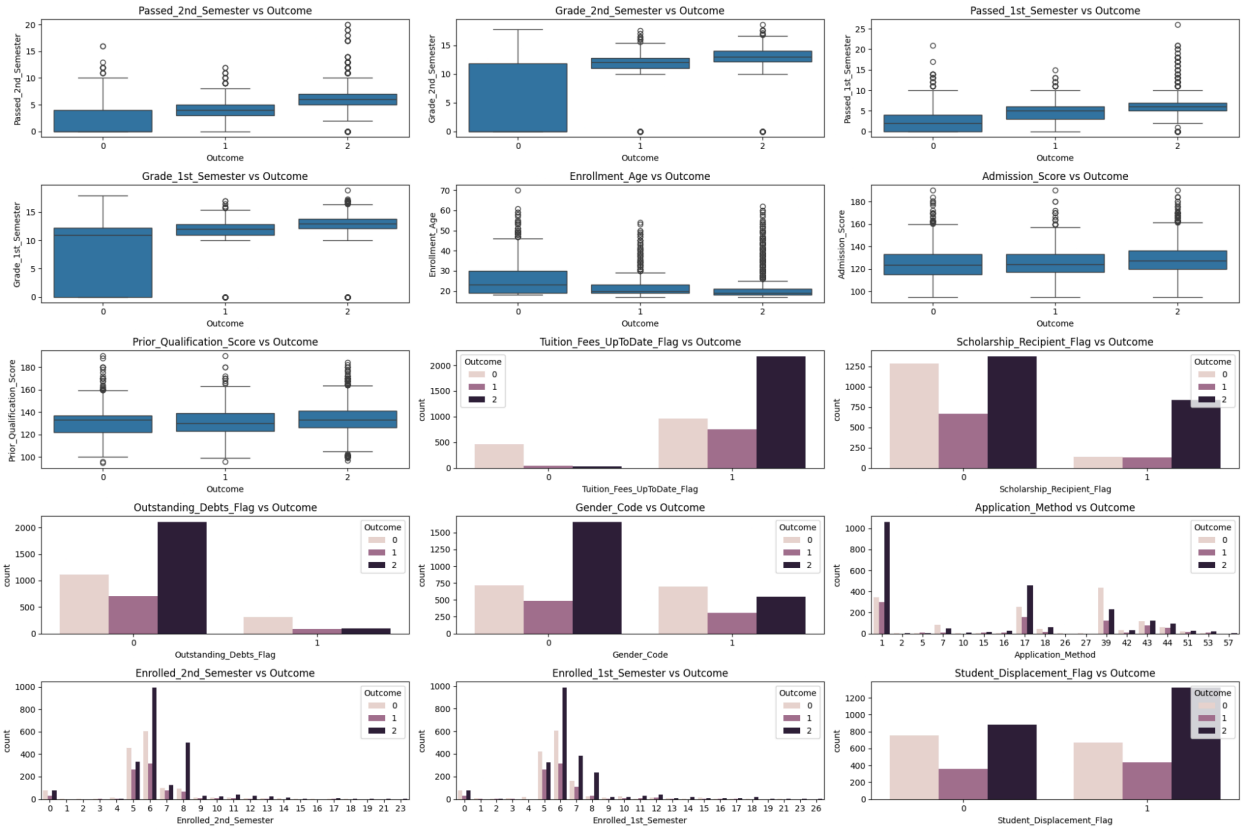
We extract the best features with a correlation value > 0.1 with the Outcome.

We are left with the following features:-

['Outcome', 'Passed_2nd_Semester', 'Grade_2nd_Semester',
'Passed_1st_Semester', 'Grade_1st_Semester',
'Tuition_Fees_UpToDate_Flag', 'Scholarship_Recipient_Flag',
'Enrollment_Age', 'Outstanding_Debts_Flag', 'Gender_Code',
'Application_Method', 'Enrolled_2nd_Semester', 'Enrolled_1st_Semester',
'Admission_Score', 'Student_Displacement_Flag',
'Prior_Qualification_Score']

We Visualise Numerical Features with Boxplots and Categorical Features with Countplots

Visualisation for Best Features:-



Feature Engineering:-

We add the following features from the best shortlisted features

- Overall_Grade = Average Grade Across both the Semesters
- Grade_Improvement = the trajectory of grades, whether upwards or downwards
- Sems_Passed = How many semesters passed successfully
- Total_Credits = Total credits enrolled in
- Age_Group = Age Grouped into categories
- Tuition_Fees_Flag = Financial Indicator

OUTLIER CLIPPING: We Clip Outliers with quantile

SCALING: With the Standard Scaler, We scale the data

MODEL TRAINING AND EVALUATION:-

First, we fit the data to a general Function for Model Training. To do this, we split the data with train_test_split.
Random_state is set to 18
We train the following ML models in this function:-

LogisticRegression, RandomForestClassifier, GradientBoostingClassifier, SVC, KNeighborsClassifier, XGBoostClassifier

We evaluate against Accuracy, Precision, Recall, F1-Score:-

	Model	Accuracy	Precision:	Recall	F1
1	RandomForest	0.800000	0.761671	0.705706	0.724078
2	GradientBoosting	0.790960	0.746389	0.693735	0.710891
5	XGBoost	0.770621	0.710625	0.674548	0.686714
0	LogisticRegression	0.769492	0.702616	0.663481	0.674100
3	SVC	0.697175	0.609768	0.542252	0.521951
4	KNN	0.655367	0.582655	0.559806	0.566952

Next, we create a new Function where we fine-tune the models using GridSearchCV to get the best models, where we limit the function to the following models because of time and resource constraints: Logistic Regressor, Random Forest, Gradient Boost and XGBoost

From the new function, we get the following results:-

	Model	Best Accuracy (CV)	Test Accuracy	Test Precision	Test Recall	Test F1-Score	Best Parameters
3	XGBoost	0.776207	0.796610	0.789688	0.796610	0.785986	{'learning_rate': 0.1, 'max_depth': 3, 'n_esti...
2	GradientBoosting	0.773100	0.792090	0.784038	0.792090	0.782493	{'learning_rate': 0.1, 'max_depth': 3, 'n_esti...
1	RandomForest	0.771122	0.786441	0.776671	0.786441	0.774001	{'max_depth': 20, 'min_samples_leaf': 1, 'n_es...
0	LogisticRegression	0.758692	0.763842	0.737957	0.763842	0.738251	{'C': 0.1, 'solver': 'liblinear'}

RESULT VISUALISATION:-

We visualise our Best results with a Bar Chart Plot and conclude with an approximate **80 percent Accuracy**