OESON

PROJECT 3/4

EDA AND MACHINE LEARNING ANALYSIS

STUDENT ACADEMIC RECORDS

Rohit Sunku

INTRODUCTION

- **Situation:**

We are provided with a dataset featuring different academic backgrounds, their academic records and economical statistics of those within their area.

- **Task**

Utilise Machine Learning Models to predict student graduation outcomes based on the data, so the university can use this within their student support campaigns.

- **Action**

Understand the data, conduct detailed pre-processing steps removing noise. Prepare data for ML Analysis.

- **Result**

We compare and analyse different models and assess them on performance using Binary Classification performance metrics (precision, recall, f1-score, accuracy), lift or gain, KS Statistics, confusion matrices and ROC - AUC Curves

- **Impact**

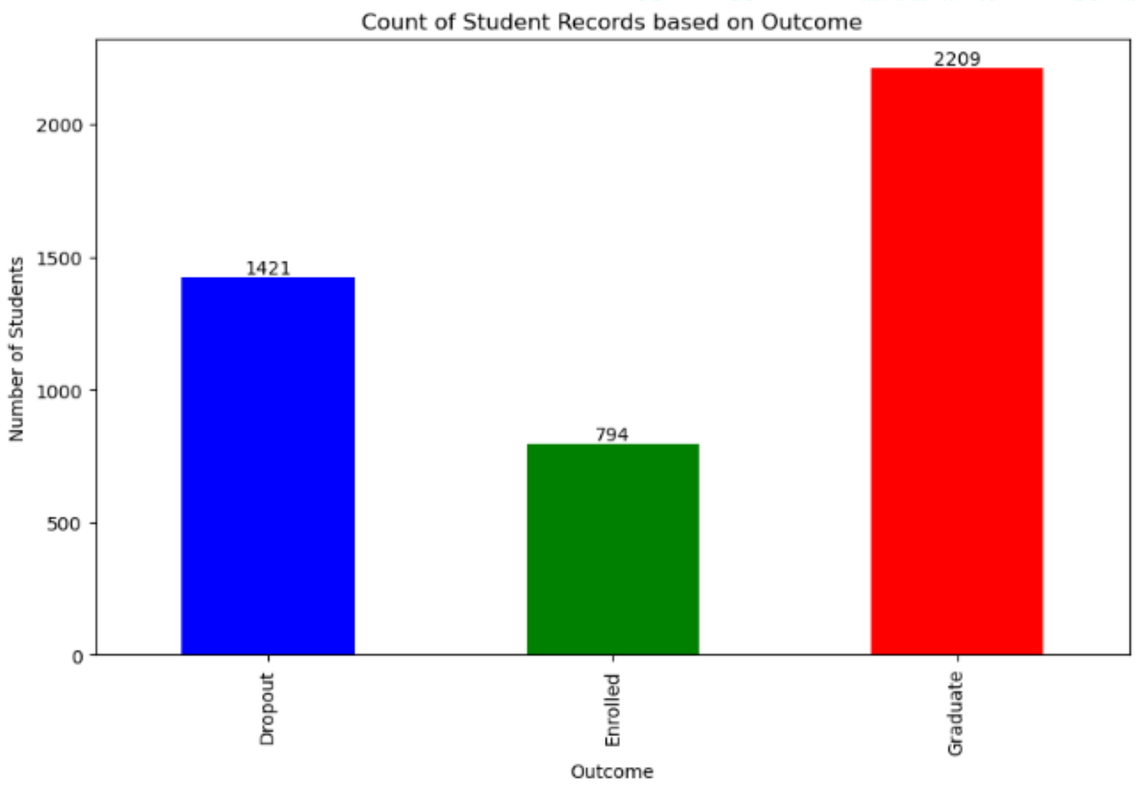
Data-driven decisions can be made to improve the current learning environment, providing aid to students who are statistically less likely to graduate, groups of students in need of bursaries and funding can be identified, track record of university can improve since the dropout rate is lower.

DATA DESCRIPTION

- **Overview**: Dataset containing records of different students profiles and their academic backgrounds. There are 35 variables and 4424 observations.
- **Features:**
  - **Independent Variable:** Target
  - **Dependant Variables (Profile):** Marital Status, Course, Nationality, Debtor, Tuition Fees Up to Date, Application Mode, Application Order, Daytime/ Evening Attendance, Previous Qualification, Mother's Qualification, Father's Qualification, Mother's Occupation, Father's Occupation, Displaced, Educational Special Needs, Gender, Scholarship holder, Age at enrolment, International
  - **Dependant Variables (Curriculum):** Curricular units 1st sem (credited), Curricular units 1st sem (enrolled) + 9 more (11 curriculum features overall)
  - **Dependant Variables (Global):** Unemployment rate, Inflation rate, GDP
- In total, there are 26 numerical variables and 9 categorical variables.

EDA

As we can see from the Target Variable plot, there are 1421 people who have dropped out, 794 people currently enrolled and 2209 students who have graduated. We exclude those students who have enrolled within this data for training the model, however we use the enrolled data to test the accuracy of our models.



EDA

Age at Enrollment

We can see that most college students who are older tend to be dropout students since they are not motivated to pass and start working.

Application Mode

Application Modes are distributed fairly evenly for graduate and dropout students, however a higher proportion of graduates apply through two specific application modes.

Scholarship Holder

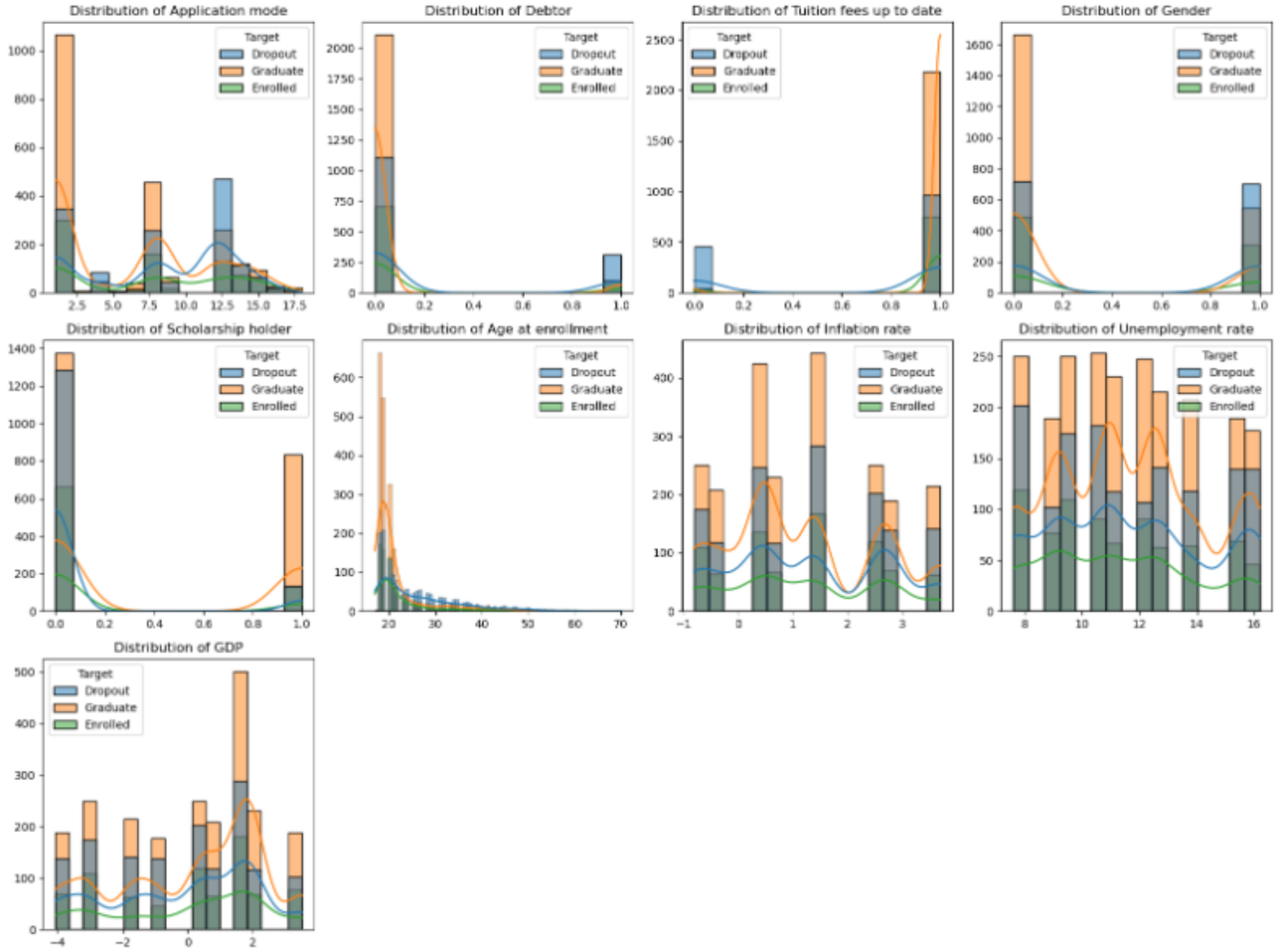
Most scholarship holders graduate, there are only a few dropouts who have scholarships.

Inflation, Unemployment and GDP

There is no impact on student outcome for these global variables; these economic indicators have no correlation.

Debtor and Tuition Fees

Most graduate students have paid their tuition fees and are not in debt however dropout students have not paid and are in debt.



STEPS FOR DATA PRE-PROCESSING

- **Duplicate/ Null Value Removal:** There are no duplicates/ null values within this dataset. Imputation techniques are invalid.
- **Outlier Removal:** Check to see if there are any outliers within the data, and remove them. We do this using the interquartile range/ z-score method.
- **Feature Selection:** Correlation Analysis using a Heatmap and using PCA to select important features with the highest correlation to the target variable. Remove features below threshold.
- **Feature Scaling:**
  - **Summaries:** Represent features in diagnostic plots and summaries the data, to measure skewness and kurtosis.
  - **Standardisation:** Features with a skewness higher than 1 should be standardised, preventing larger values dominating the modelling process.
  - **Label Encoding:** We also encode the target variable within our model

```
abs(correlation_matrix_percentage["Target"])
```

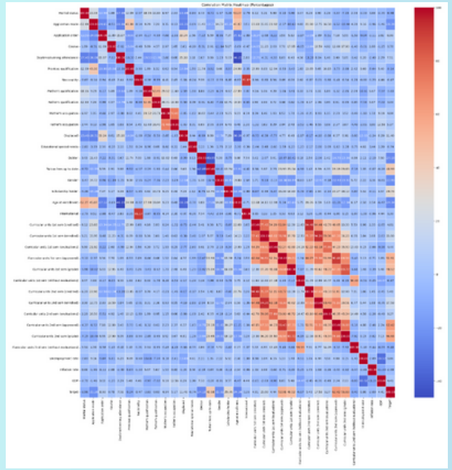
```
Marital status                                    10.047907
Application mode                                  23.388814
Application order                                  9.435463
Course                                             0.681430
Daytime/evening attendance                         8.449594
Previous qualification                            10.279451
Nacionality                                        0.382283
Mother's qualification                             4.845856
Father's qualification                             0.384991
Mother's occupation                                6.419503
Father's occupation                                7.323826
Displaced                                         12.611304
Educational special needs                          0.725365
Debtor                                            26.720720
Tuition fees up to date                           44.213758
Gender                                            25.195481
Scholarship holder                                31.301766
Age at enrollment                                 26.722938
International                                      0.618126
Curricular units 1st sem (credited)                4.690002
Curricular units 1st sem (enrolled)               16.107352
Curricular units 1st sem (evaluations)             5.978626
Curricular units 1st sem (approved)               55.488086
Curricular units 1st sem (grade)                  51.992709
Curricular units 1st sem (without evaluations)     7.464226
Curricular units 2nd sem (credited)                5.240197
Curricular units 2nd sem (enrolled)               18.289654
Curricular units 2nd sem (evaluations)            11.923877
Curricular units 2nd sem (approved)               65.399525
Curricular units 2nd sem (grade)                  60.535013
Curricular units 2nd sem (without evaluations)    10.268683
Unemployment rate                                  0.419811
Inflation rate                                     3.032587
GDP                                                5.026015
Target                                           100.000000
Name: Target, dtype: float64
```

FEATURE SELECTION

From our first correlation matrix heatmap, we isolate the Target column and assess feature correlation strength.

The threshold for our initial models is 20% or above, so we include the following variables in our models: Application mode, Debtor, Tuition fees up to date, Gender, Scholarship holder, Age at enrollment, Curricular units 1st sem (approved), Curricular units 1st sem (grade), Curricular units 2nd sem (approved) and Curricular units 2nd sem (grade).

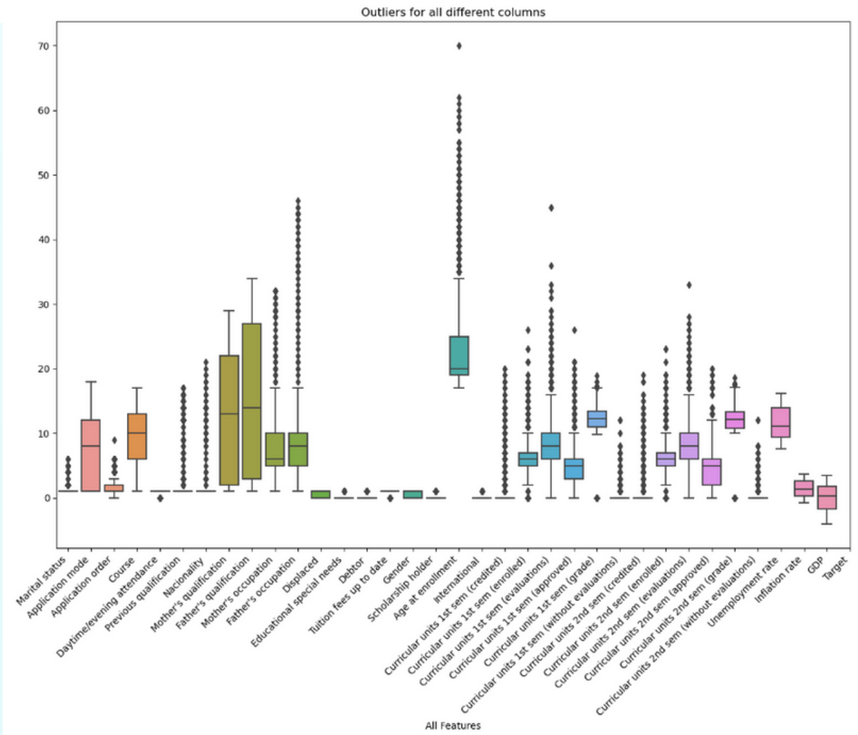
We will use K-Fold cross validation to experiment model accuracy with different thresholds (see this later)

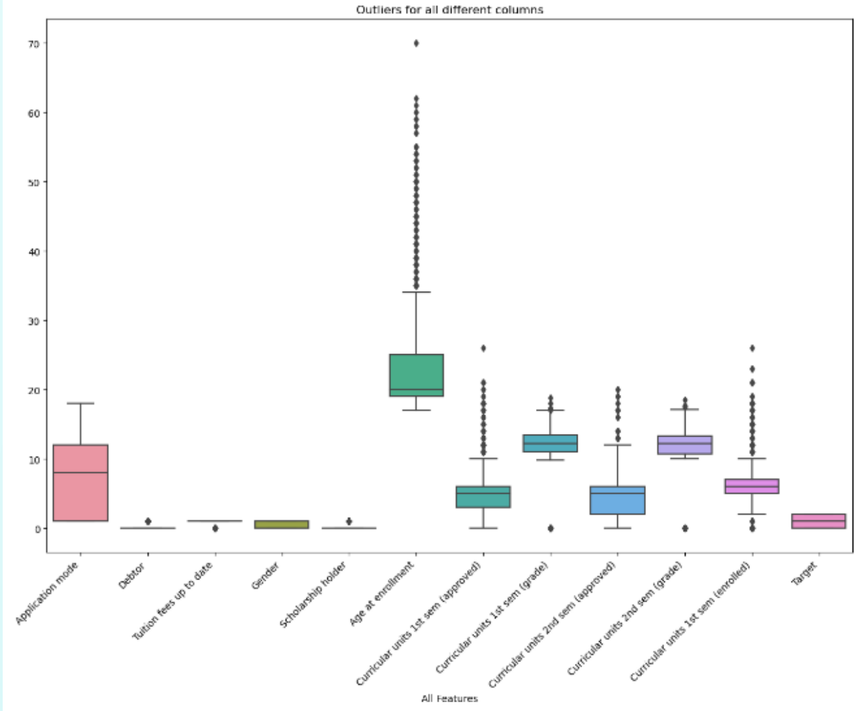


OUTLIER REMOVAL

We can see all the outliers within the numerical columns:

- **Outliers we keep:** Encoded categorical variables such as Marital Status, Nationality, Previous Qualification, Application Order, Debtor, Tuition Fees up to data are not removed, since this will reduce the quality of the data distribution. When considering outliers, it is important to focus on continuous numerical variables and exclude encoded ones.
- **Outliers:** Post feature selection, we remove the outliers from index 6-11 on the x axis of the boxplot graph.





STANDARDISATION OF DATA

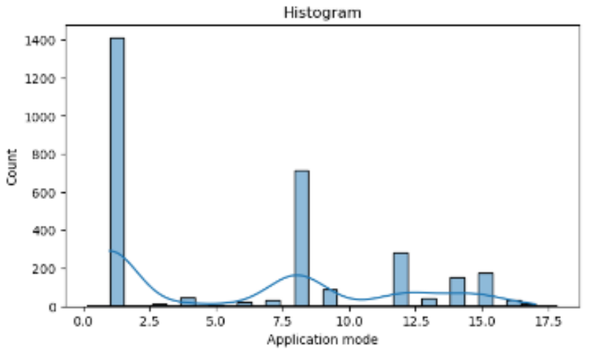
Variables with low skewness

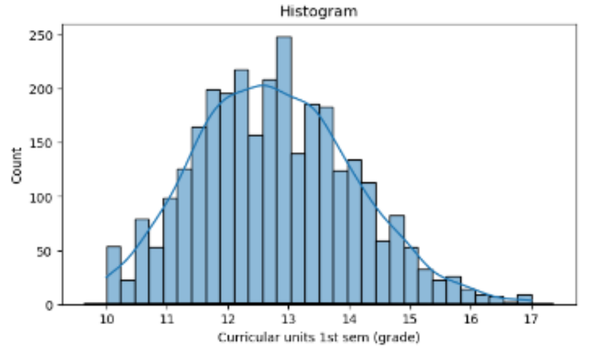
Look at some of the variables on the right and their distributions, we can see that Application Mode (Skewness = 0.12) and Curricular units 1st sem (grade) after removing outliers has a skewness of has a distribution of results around the mean which can be represented like a normal distribution.

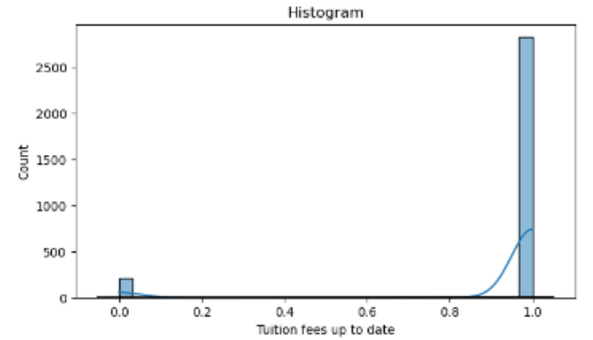
Variables with outliers

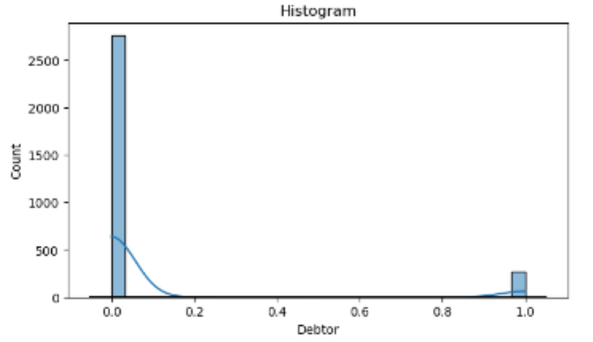
On the other hand, Curricular Units 1st Sem is an example of a variable with outliers, that need to be removed. The bottom two histograms on the right feature variables which are categorical. There is a need for standardisation, as these distributions cannot be represented in a bell-shaped curve, however we do not consider outliers for this.

We employ a minimum maximum scaler to transform the data.



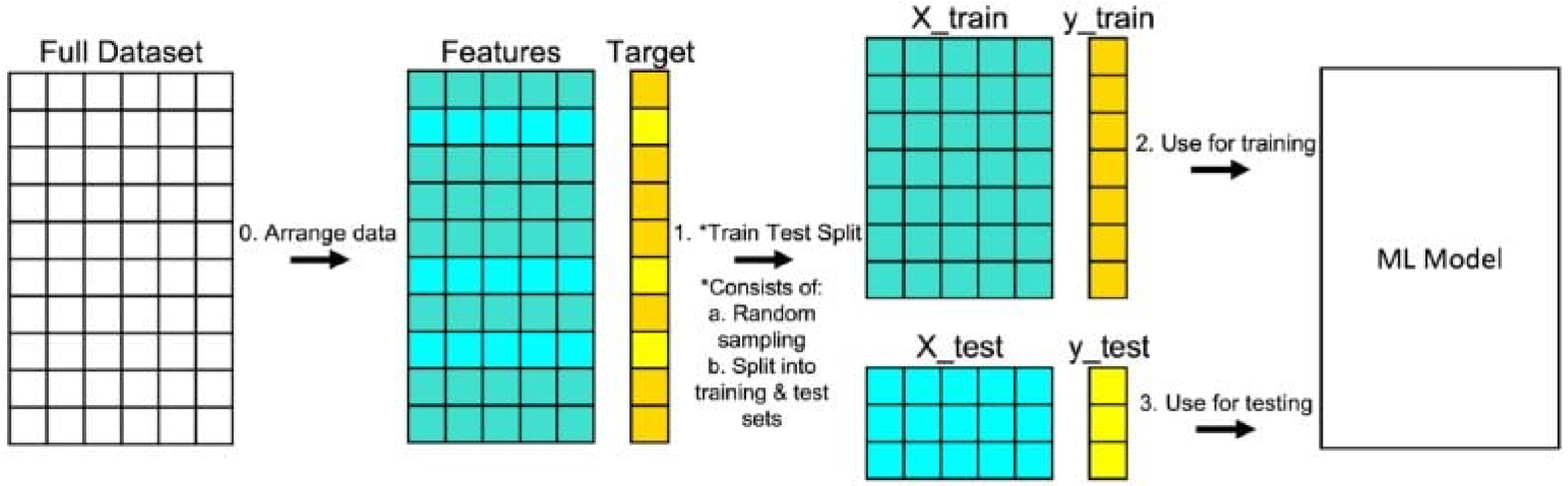






TRAIN TEST SPLIT

After standardising all the features, we perform train-test split on the data.



OPTIMISATION + HYPER-PARAMETER TUNING

Models	Optimal Hyper-parameters
K-Nearest Neighbours (KNN)	**K** = 7
Support Vector Machine (SVM)	**C** = 10, **Gamma** = 0.1, **Kernel** = Linear
Decision Tree (DT)	**Criterion** = Entropy, **Maximum Depth** = 5, **Minimum Samples Leaf** = 1, **Minimum Samples Split** = 5
Random Forest (RF)	**Criterion** = Entropy, **Maximum Depth** = 20, **Minimum Samples Leaf** = 1, **Min Samples Split** = 2, **Number of estimators** = 100
Multinomial Naïve Bayes	**Alpha** = 1.5, **Fit Prior** = True
Logistic Regression	**C** = 1.0, **Penalty** = L1, **Solver** = Liblinear

We tune the hyper-parameters using “GridSearchCV” to find our optimal models for each ML algorithm.

MODEL PERFORMANCE METRICS (CLASSIFICATION REPORTS)

Model	Precision	Recall	F1 Score	Accuracy
K-Nearest Neighbours (KNN)	0.876	0.977	0.924	0.875
Support Vector Machine (SVM)	0.885	0.983	0.932	0.888
Decision Tree (DT)	0.875	0.968	0.919	0.868
Random Forest (RF)	0.884	0.987	0.933	0.890
Multinomial Naïve Bayes	0.820	0.987	0.896	0.822
Logistic Regression	0.890	0.983	0.934	0.888

- In terms of overall performance based on the metrics provided, the **Random Forest** has the highest scores for recall and accuracy.
- Nevertheless, **Logistic Regression** are better at predicting the positive class, having a higher f1-score and precision.
- In this case, the positive class is treated as students who have dropped out, since we want a model to predict dropout students more accurately to protect university reputation and target support schemes to those students. So a Logistic Regression model would be desired in this case.

PLAYING AROUND WITH DIFFERENT CORRELATION THRESHOLDS

Models	Threshold	Mean Accuracies
K-Nearest Neighbours (KNN)	30%	0.96
Support Vector Machine (SVM)	5%	1.00
Decision Tree (DT)	5%	1.00
Random Forest (RF)	5%	1.00
Multinomial Naïve Bayes	20%	0.93
Logistic Regression	5%	1.00

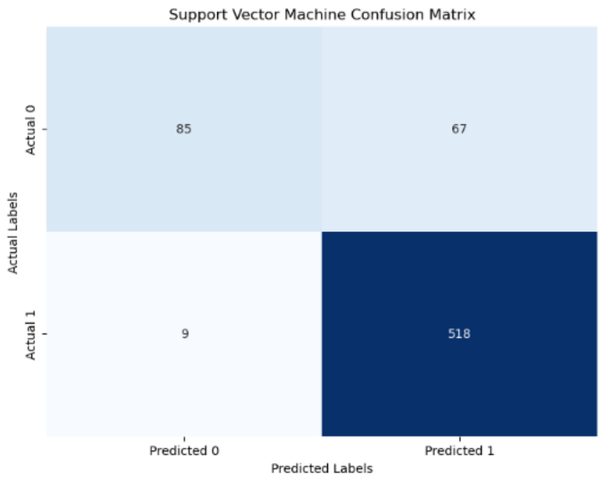
Using the K-Fold cross validation to play around with different percentage thresholds for choosing features to predict graduation outcomes we experiment with 5%, 10%, 15%, 20%, 25% and 30% and observe that:

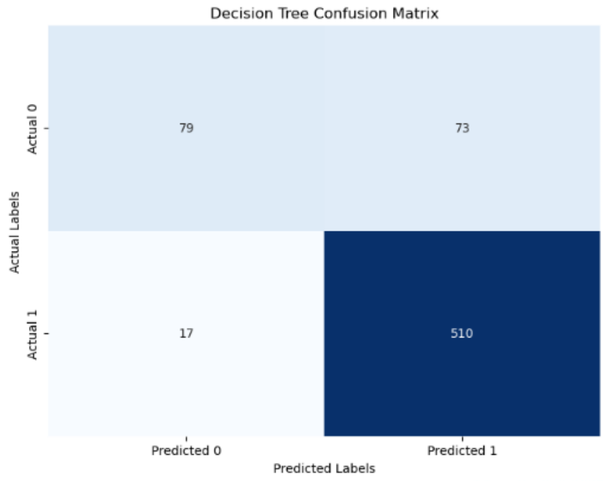
- KNN would favour a higher threshold due to the nature of the algorithm, as it is more efficient in grouping data points on a smaller number of features
- SVM, Decision Trees, Random Forest Models and Logistic Regression models should be trained on all the features, choosing specific features does not improve their accuracy.
- Multinomial NB has a more concrete percentage threshold range of 20%

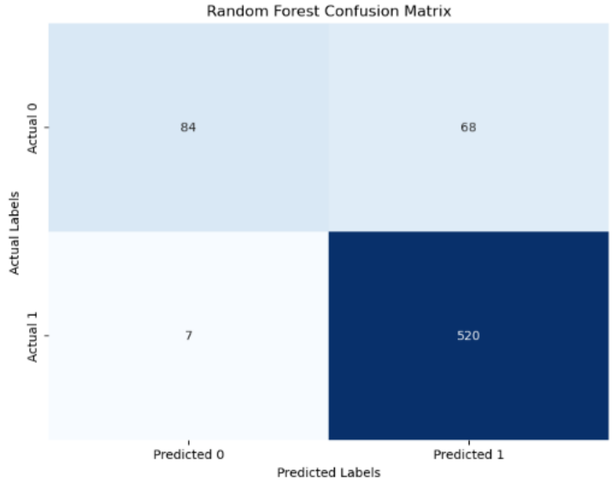
CONFUSION MATRICS

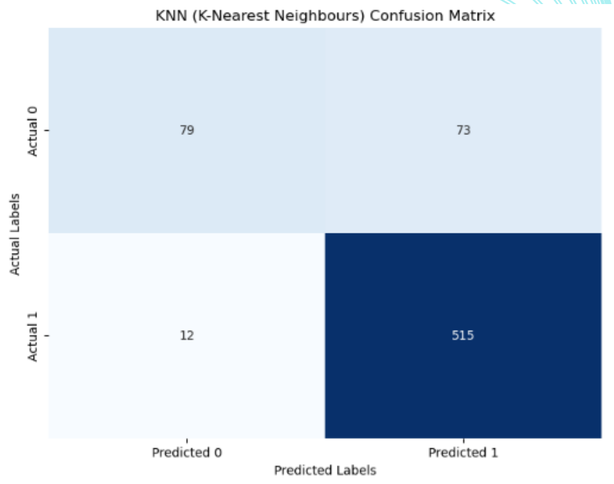
We can see all the outliers within the numerical columns:

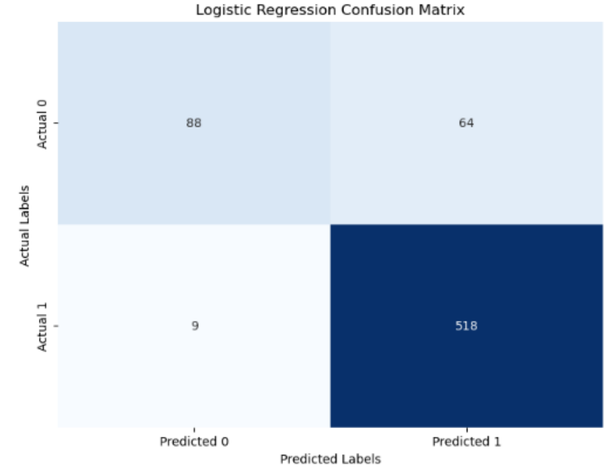
- Logistic Regressions are able to predict the positive class better, however Multinomial Bayes and Random Forests give better indications of the negative class
- Random Forest models are better at predicting students who have graduated, but the Logistic Regression Model is better at predicting dropout students.

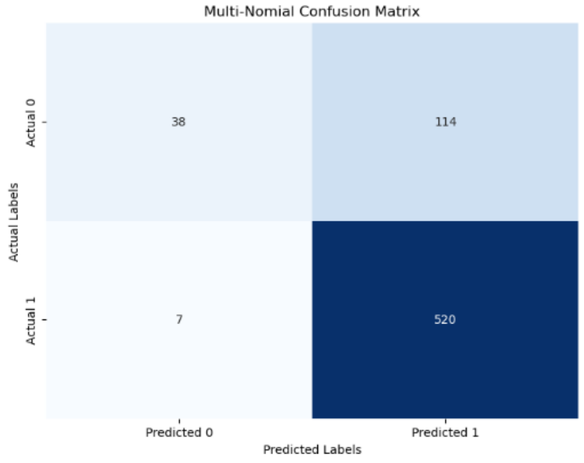




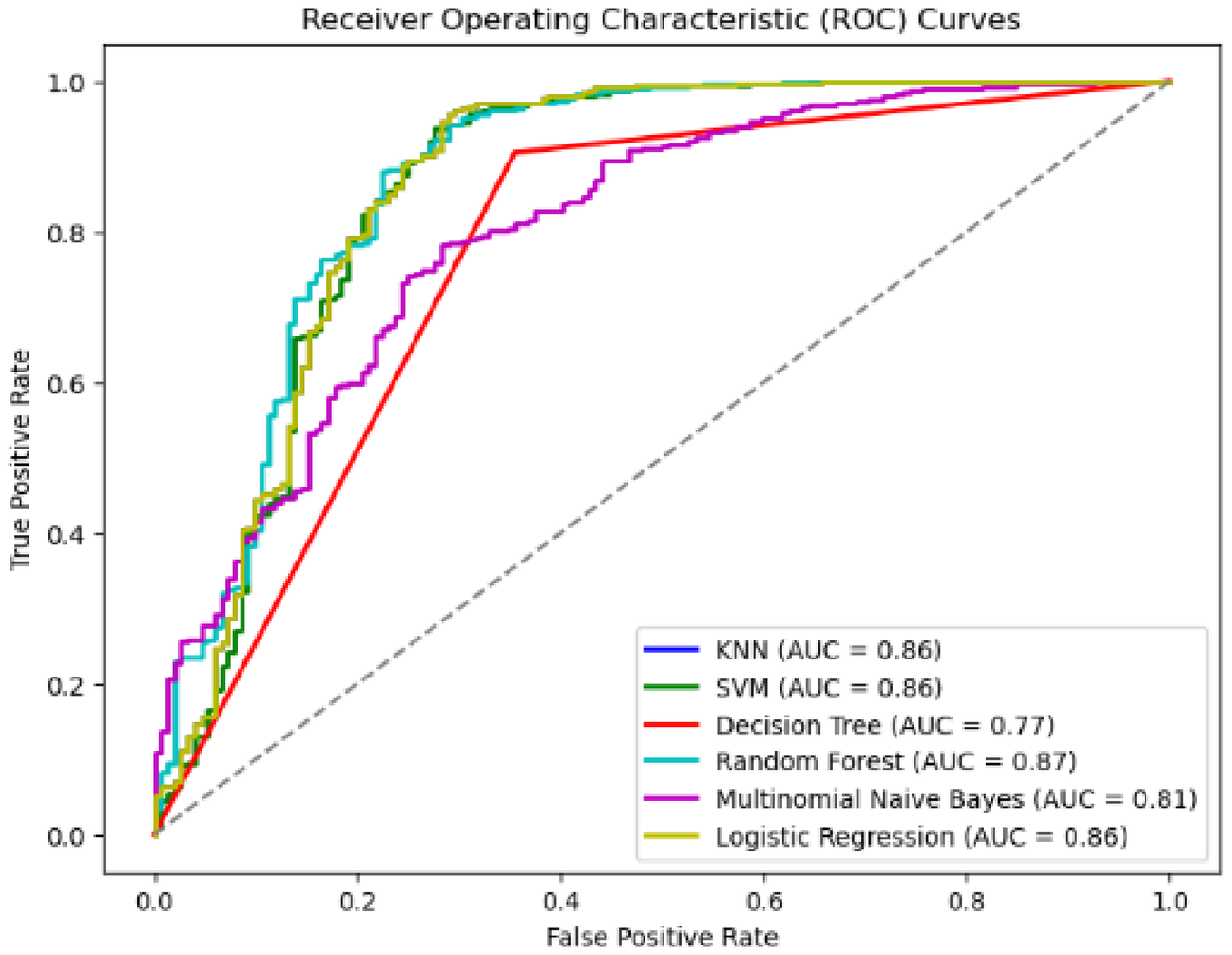








ROC - AUC CURVES



KS STATISTICS

Models	K-Statistics
K-Nearest Neighbours (KNN)	0.548
Support Vector Machine (SVM)	0.661
Decision Tree (DT)	0.550
Random Forest (RF)	0.657
Multinomial Naïve Bayes	0.501
Logistic Regression	0.667

The Kolmogorov-Smirnov test is a non-parametric tests (do not require a normal distribution) which tests for differences between the two distributions (two classes). A higher KS Statistic indicates the model has good discrimination and is better at separating the two classes.

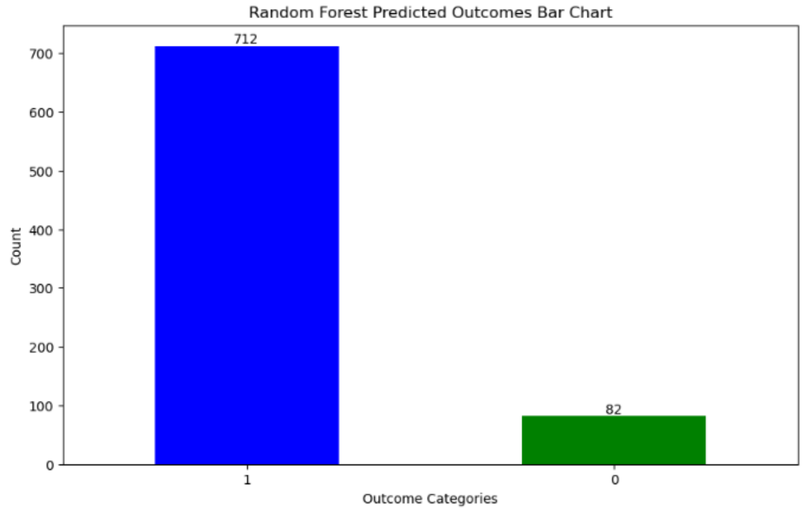
Hierarchy of Models with the best KS Statistics:

1. Logistic Regression
2. Support Vector Machine
3. Random Forest

MODEL PREDICTIONS

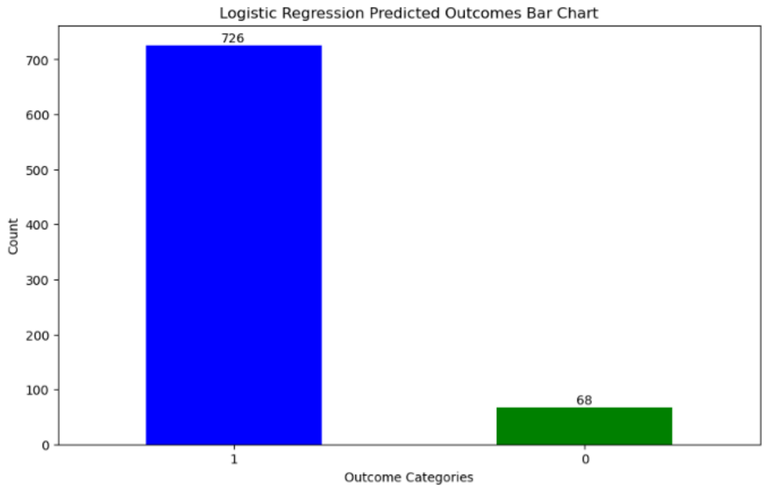
Logistic Regression Model Predictions:

1	712
0	82



Random Forest Model Predictions:

1	726
0	68



We use our final Random Forest model and make predictions on the number of enrolled students which we separate from the Graduate and Dropout students during data preparation. The Random Forest predicts a higher number of dropouts but a lower number of Graduates than the Logistic Regression.

CONCLUSIONS AND FURTHER WORK

Missing Values: There are no missing or duplicate values within this dataset, however there are a lot of outliers which need to be removed.

Outlier Removal: Removing outliers increases the accuracy of the model. We remove outliers for all variables.

Feature Selection: Feature Selection has a minimal effect on the model accuracy of SVM, LR, DT and RF models, but it does affect Multinomial Naïve Bayes and K-Nearest Neighbours.

Feature Importance: We standardise all features, since there is a large number of them with a high skewness to reduce bias and overfitting of the model.

Model Performance Metrics: The Random Forest model with optimised hyperparameters achieved the highest accuracy (0.890) but the Logistic Regression model had the highest precision (0.890) on the test set. Multinomial Bayes was the worst model overall, but had a very high recall, having a strong ability to predict the negative class. LR also has the highest KS Statistic value, so it is the best at being able to differentiate between the positive and negative class.

Future Work: More training and testing has to be done on more data points to get a better indication of model performance before selecting our final model. We can train optimised models with the ideal percentage threshold features again, removing outliers and oversampling data to account for the small but relative class imbalance between graduated students and dropouts.

THANK YOU