**Predict Students’ Dropout and Academic Succuss**

**MODULE - DS7003**

**By**

**ASHIPA OLUWASEYI PAUL**

Table of Contents

[TABLE OF CONTENTS 3](#_Toc166264879)

[Abstract 4](#_Toc166264880)

[1.0 Introduction 4](#_Toc166264881)

[2.0 Literature review 5](#_Toc166264882)

[2.1 Classification Method Machine learning 5](#_Toc166264883)

[3.0 Methodology approach 6](#_Toc166264884)

[3.1 Data Acquisition / Cleaning 7](#_Toc166264885)

[3.2 Data description 7](#_Toc166264886)

[3.3 Meta Data Structure 7](#_Toc166264887)

[3.4 Summary 8](#_Toc166264888)

[3.5 Checking for missing values 9](#_Toc166264889)

[4.0 Correlation Matrix 10](#_Toc166264890)

[5.0 MIN & MAX Normalization 11](#_Toc166264891)

[6.0 Pie Chart Of The Dependent Variable 12](#_Toc166264892)

[7.0 Machine learning Model Building 13](#_Toc166264893)

[7.1 Random Forest 13](#_Toc166264894)

[7.2 Support Vector Machine 15](#_Toc166264895)

[7.3 K Nearest Neighbour 16](#_Toc166264896)

[7.4 Decision Tree 17](#_Toc166264897)

[7.5 Naive Bayes 18](#_Toc166264898)

[8.0 Conclusion 21](#_Toc166264899)

[Reference 22](#_Toc166264900)

[APPENDICS 23](#_Toc166264901)

# Abstract

Higher education establishments gather a great deal of data about their students, which offers a great deal of opportunity to produce understanding, information, insight, and monitoring. both academic achievement and school dropout rates. Dropouts in higher education have a direct influence on the lives of students and their families, higher education institutions, and society at large. They also impede economic growth, competitiveness, employment, and productivity. Academic information on enrolment and performance at the conclusion of the first and second semesters is included in the dataset that is being described. Machine learning models for forecasting academic achievement and dropout rates are constructed using the dataset.

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

The confusion matrix was used to evaluate the model's performance, and the outcomes for each student variety specifically, sensitivity, balanced accuracy, specificity, and F1\_score were compared. There was also a comparison of the models' general statistics. Random Forest was found to outperform the others by a small margin.

# 1.0 Introduction

Academic achievement in higher education is necessary for a deeper sense of purpose in life, as well as for improving economic growth, social justice, and job prospects. The fact that dropouts need to be addressed in order to increase their success is the most troublesome issue facing higher education institutions. (Dekker et al., 2009)There's no consensus definition for the term "dropout". Depending on the definition of dropout, the data source, and the computation method, the percentage of students who drop out differs amongst researches. Early vs. late dropout is a common basis for analysis in research literature. It is not possible to compare dropout rates amongst universities due to discrepancies in reporting. We define dropouts in this study from a micro-perspective, where changes in an institution or field are considered dropouts regardless of when they occur.

According to the European Commission's independent analysis, far too many students leave their higher education programs early . Over 40,000 undergraduate students in Denmark, the most successful nation, dropped out of UK universities, indicating that only about 80% of students finish their education; in Italy, the percentage is as low as 46% (Kocsis and Molnár, 2024). The socioeconomic environment is the primary reason for student dropouts, as this paper indicates.

# 2.0 Literature review

(Realinho et al., 2022)worked on a dataset of students gathered by the Polytechnic Institute of Portalegre. The dataset included 4424 records with 35 attributes, including demographics, socioeconomics, macroeconomic data, and academic performance at the end of the first and second semesters. The students were enrolled in a variety of undergraduate programs, including nursing, design, education, journalism, social service, management, and technologies. Four machine learning models were built: Random Forest, XGboost, LightGBM, and CatBoost. Within the four models, Curriculum Units 2nd Sem (approved) was shown as the most important factor that affects student performance the most. The random forest and XGboost tuition fees up to date were the second-most important factors. In total, there are 10 attributes that contribute the most.

(Nagy and Molontay, 2018) research and model were built on the undergraduate student dataset from the Budapest University of Technology and Economics. They were enrolled between 2010 and 2017 and finished their undergraduate studies either by graduation or by dropout. During the analysis of the dataset, they had a problem with missing data by imputation. After that, they proceeded to perform feature extraction and feature selection, then used them to build their model. Ten machine learning models were built: Decision Tree, Random Forest, Generalised Linear Model, Naive Bayes, Adeptive Boost, K-NN, Logistic Regression, Gradient Boosted Tree, and Deep Learning. We compared the models based on their accuracy, recall, consciousness, and area under the curve. Three models were picked based on their accuracy. Deep learning, gradient-boosted trees, and logistic regression. But more research is said to be conducted to improve the performance of the model by using semi-supervised learning with more attributes.

(Kiss Botond et al., 2019) research is said to predict students who are at risk of dropout. The dataset was accumulated from a Hungarian technical university and contains data on 10,196 students who finished their undergraduate studies (either by graduate or dropout) between 2013 and 2018. The machine learning techniques that were used were XG-boosted, Graident-boosted, and artificial neural networks.  
Two predictions were made: the first was with the data at the time of enrolment, but after that, they believe the performance can improve by adding more features. The second prediction is that XGBoost provides a very powerful classifier (AUC = 0.920).

## 2.1 Classification Method Machine learning

Support Vector Machines (SVM) is a kernel-based technique that effectively solves both regression and classification problems with a high processing capacity, according to (Ukil Abhisek, 2007) said. Support Vector Machines (SVM) have a better generalisation than other machine learning algorithms. In many applications, they are recognised to yield more accurate results than other algorithms and have a solid theoretical foundation.

A decision tree is a model or graphical representation that shows decisions and their potential results and is used as a tool for decision support. In decision analysis, decision trees are commonly utilised to determine the best course of action to follow. Decision trees are commonly used in classification models because of their high dependability, low implementation costs, simplicity of interpretation, and smooth connection with database systems (Aized Amin Soofi and Arshad Awan, 2017).It is easily comprehensible due to its visual and intuitive qualities.

A pattern recognition method called the k-Nearest Neighbour (KNN) algorithm groups items according to how close they are to other examples in the attribute space. To be more precise, it finds the k training examples that are closest to a given test example and then places the test example in the class that has the highest frequency among its k closest neighbours. (Zhang Zhongheng, 2016) Using the nearest neighbouring samples in the attribute space as a guide, the k-Nearest Neighbour (kNN) algorithm categorises objects. It makes use of the value of k, which establishes how many neighbours are taken into account for classification. A vector's class is assigned by the kNN method depending on the class of its closest neighbours. The algorithm examines the known vectors in the class in order to categorise a vector. The specimen under examination is then processed with each sample in the training set individually.

Building a breakfast of decision trees is how random forests, an ensemble learning technique, work for problems like regression, classification, and other stuff (Biau and Scornet, 2016). The random forest's output for classification problems is the class that the greatest number of trees chose. An extension algorithm was developed by who registered "Random Forests" as a trademark in 2006.). Tin Kam Ho developed the first algorithm of random decision utilising the random subspace approach in 1995.

# 3.0 Methodology approach

The method used to create the dataset is described in this section, along with the exploratory data analysis that highlights some pertinent issues that might make it easier for other researchers to access and utilize the dataset quickly. These issues include the imbalanced nature of the data, below you will see the process and step which was taken in analysing and building of the model.

**IMPORTING DATASET**

**TRAINING AND TESTING**

**EXPLORE THE DATA  
(factor, correlation, normalization)**

**CLEAN THE DATASET**  
**(DATA CLEANING)**

**BUILDING MODELS  
(SVM, RANDOM FOREST, DECISION TREE)**

**COMPARING OF MODELS**

## 3.1 Data Acquisition / Cleaning

The UCI Machine Learning repository provides a cleaned and pre-processed dataset on student drop out and academic performance. This dataset is well-suited for constructing machine learning models.

The dataset of student drop out and academic success, is available on the UCI Machine Learning repository, which has been cleaned and pre-processed to serve as a good dataset for building machine learning models.

## 3.2 Data description

This dataset provides a thorough overview of the students enrolled in various undergraduate programs that a university offers. The data contains information pertaining to academic achievement, socioeconomic status, and demographics, which can be utilized for the purpose of analysing and predicting academic success as well as student dropout rates. This dataset includes different databases with pertinent data that was accessible at the time of enrolment, including application mode, marital status, selected course, and more.

Furthermore, by assessing the curricular units that have been credited, enrolled, or approved, together with their respective grades, this information can be utilized to gauge the total academic achievement of students at the conclusion of each semester. An additional approach to comprehending the impact of economic factors on student attrition rates and academic achievement outcomes is to examine the gross domestic product (GDP), inflation rate, and unemployment rate of the region. This tool has powerful analytical features that will give you useful details about what makes students choose different job paths, such as agronomy, design, education, nursing, journalism, management, social service, technology, or dropping out of school.

## 3.3 Meta Data Structure

The structure below shows the columns of the data set and the types consisting of 4424 rows and 37 variable.

|  |
| --- |
| 'data.frame': 4424 obs. of 37 variables:  $ Marital.status : int 1 1 1 1 2 2 1 1 1 1 ...  $ Application.mode : int 17 15 1 17 39 39 1 18 1 1 ...  $ Application.order : int 5 1 5 2 1 1 1 4 3 1 ...  $ Course : int 171 9254 9070 9773 8014 9991 9500 9254 9238 9238 ...  $ Daytime.evening.attendance. : int 1 1 1 1 0 0 1 1 1 1 ...  $ Previous.qualification : int 1 1 1 1 1 19 1 1 1 1 ...  $ Previous.qualification..grade. : num 122 160 122 122 100 ...  $ Nacionality : int 1 1 1 1 1 1 1 1 62 1 ...  $ Mother.s.qualification : int 19 1 37 38 37 37 19 37 1 1 ...  $ Father.s.qualification : int 12 3 37 37 38 37 38 37 1 19 ...  $ Mother.s.occupation : int 5 3 9 5 9 9 7 9 9 4 ...  $ Father.s.occupation : int 9 3 9 3 9 7 10 9 9 7 ...  $ Admission.grade : num 127 142 125 120 142 ...  $ Displaced : int 1 1 1 1 0 0 1 1 0 1 ...  $ Educational.special.needs : int 0 0 0 0 0 0 0 0 0 0 ...  $ Debtor : int 0 0 0 0 0 1 0 0 0 1 ...  $ Tuition.fees.up.to.date : int 1 0 0 1 1 1 1 0 1 0 ...  $ Gender : int 1 1 1 0 0 1 0 1 0 0 ...  $ Scholarship.holder : int 0 0 0 0 0 0 1 0 1 0 ...  $ Age.at.enrollment : int 20 19 19 20 45 50 18 22 21 18 ...  $ International : int 0 0 0 0 0 0 0 0 1 0 ...  $ Curricular.units.1st.sem..credited. : int 0 0 0 0 0 0 0 0 0 0 ...  $ Curricular.units.1st.sem..enrolled. : int 0 6 6 6 6 5 7 5 6 6 ...  $ Curricular.units.1st.sem..evaluations. : int 0 6 0 8 9 10 9 5 8 9 ...  $ Curricular.units.1st.sem..approved. : int 0 6 0 6 5 5 7 0 6 5 ...  $ Curricular.units.1st.sem..grade. : num 0 14 0 13.4 12.3 ...  $ Curricular.units.1st.sem..without.evaluations.: int 0 0 0 0 0 0 0 0 0 0 ...  $ Curricular.units.2nd.sem..credited. : int 0 0 0 0 0 0 0 0 0 0 ...  $ Curricular.units.2nd.sem..enrolled. : int 0 6 6 6 6 5 8 5 6 6 ...  $ Curricular.units.2nd.sem..evaluations. : int 0 6 0 10 6 17 8 5 7 14 ...  $ Curricular.units.2nd.sem..approved. : int 0 6 0 5 6 5 8 0 6 2 ...  $ Curricular.units.2nd.sem..grade. : num 0 13.7 0 12.4 13 ...  $ Curricular.units.2nd.sem..without.evaluations.: int 0 0 0 0 0 5 0 0 0 0 ...  $ Unemployment.rate : num 10.8 13.9 10.8 9.4 13.9 16.2 15.5 15.5 16.2 8.9 ...  $ Inflation.rate : num 1.4 -0.3 1.4 -0.8 -0.3 0.3 2.8 2.8 0.3 1.4 ...  $ GDP : num 1.74 0.79 1.74 -3.12 0.79 -0.92 -4.06 -4.06 -0.92 3.51 ...  $ Target : chr "Dropout" "Graduate" "Dropout" "Graduate" ... |

## 3.4 Summary

A quick statistical analysis of each variable in the dataset is displayed in the table below, which calculates the third quartile, mean, median, and min/max for each column. This aids in providing a statistical sneak peek at the dataset's appearance.

|  |
| --- |
| Marital.status Application.mode Application.order Course Daytime.evening.attendance. Previous.qualification Previous.qualification..grade. Nacionality  Min. :1.000 Min. : 1.00 Min. :0.000 Min. : 33 Min. :0.0000 Min. : 1.000 Min. : 95.0 Min. : 1.000  1st Qu.:1.000 1st Qu.: 1.00 1st Qu.:1.000 1st Qu.:9085 1st Qu.:1.0000 1st Qu.: 1.000 1st Qu.:125.0 1st Qu.: 1.000  Median :1.000 Median :17.00 Median :1.000 Median :9238 Median :1.0000 Median : 1.000 Median :133.1 Median : 1.000  Mean :1.179 Mean :18.67 Mean :1.728 Mean :8857 Mean :0.8908 Mean : 4.578 Mean :132.6 Mean : 1.873  3rd Qu.:1.000 3rd Qu.:39.00 3rd Qu.:2.000 3rd Qu.:9556 3rd Qu.:1.0000 3rd Qu.: 1.000 3rd Qu.:140.0 3rd Qu.: 1.000  Max. :6.000 Max. :57.00 Max. :9.000 Max. :9991 Max. :1.0000 Max. :43.000 Max. :190.0 Max. :109.000  Mother.s.qualification Father.s.qualification Mother.s.occupation Father.s.occupation Admission.grade Displaced Educational.special.needs Debtor  Min. : 1.00 Min. : 1.00 Min. : 0.00 Min. : 0.00 Min. : 95.0 Min. :0.0000 Min. :0.00000 Min. :0.0000  1st Qu.: 2.00 1st Qu.: 3.00 1st Qu.: 4.00 1st Qu.: 4.00 1st Qu.:117.9 1st Qu.:0.0000 1st Qu.:0.00000 1st Qu.:0.0000  Median :19.00 Median :19.00 Median : 5.00 Median : 7.00 Median :126.1 Median :1.0000 Median :0.00000 Median :0.0000  Mean :19.56 Mean :22.28 Mean : 10.96 Mean : 11.03 Mean :127.0 Mean :0.5484 Mean :0.01153 Mean :0.1137  3rd Qu.:37.00 3rd Qu.:37.00 3rd Qu.: 9.00 3rd Qu.: 9.00 3rd Qu.:134.8 3rd Qu.:1.0000 3rd Qu.:0.00000 3rd Qu.:0.0000  Max. :44.00 Max. :44.00 Max. :194.00 Max. :195.00 Max. :190.0 Max. :1.0000 Max. :1.00000 Max. :1.0000  Tuition.fees.up.to.date Gender Scholarship.holder Age.at.enrollment International Curricular.units.1st.sem..credited. Curricular.units.1st.sem..enrolled.  Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :17.00 Min. :0.00000 Min. : 0.00 Min. : 0.000  1st Qu.:1.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:19.00 1st Qu.:0.00000 1st Qu.: 0.00 1st Qu.: 5.000  Median :1.0000 Median :0.0000 Median :0.0000 Median :20.00 Median :0.00000 Median : 0.00 Median : 6.000  Mean :0.8807 Mean :0.3517 Mean :0.2484 Mean :23.27 Mean :0.02486 Mean : 0.71 Mean : 6.271  3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:0.0000 3rd Qu.:25.00 3rd Qu.:0.00000 3rd Qu.: 0.00 3rd Qu.: 7.000  Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :70.00 Max. :1.00000 Max. :20.00 Max. :26.000 |

## 3.5 Checking for missing values

To check the missing values the Amelia library function was called. This highlights the locations of missing values throughout the dataset and creates a visual representation of the missingness map for the complete data frame. As every datapoint in this dataset was complete, there was no missing data.

A blue square with white text

Description automatically generated

# 4.0 Correlation Matrix

**A diagram showing the results of a statistical analysis

Description automatically generated with medium confidence**

The correlation matrix displays the variables that are negatively correlated with red circle dots and the positively correlated variables with blue circle dots. The variables curricular.units.1st.semeter.credited, curricular.units.1st.semeter.enrolled, and curricular.units.1st.semeters.approved are positively or negatively associated. But when I worked on the model, I chose to utilize them all because, with the exception of extent, they are all somewhat correlated with the dependent variable. I also decided to utilize all the variables since I discovered that if I only used some of them, there would be overfitting while testing the models.

# 5.0 MIN & MAX Normalization

Min/Max Normalization is a data pre-processing technique used to normalize the range of independent variables or features of data. It is also known as data normalization. The purpose is to ensure that all variables contribute equally to the model and to prevent features with large values.

Boxplot of the normalize data, displayed a lot of outliers and some of the dataset shows no normalization, we will processed with taking all the variable, because removing them I believe it may after the performance of the model.

A graph showing the number of people in the same direction

Description automatically generated with medium confidence

# 6.0 Pie Chart Of The Dependent Variable

This chart show the contribution of observation within the dependent variable “Target”

A diagram with numbers and a circle

Description automatically generated with medium confidence

# 7.0 Machine learning Model Building

The student dataset was split into 80% training and 20% testing. This was done to ensure the model gets enough data to work with to avoid overfitting and to evaluate the performance of the model accurately.

## 7.1 Random Forest

Random Forest model was use, showing and accuracy of 0.7774 and a kappa value of 0.6235 which is substantial. The cross-table below shows the amount of prediction the model got. It predicted 226 would drop out, which was correct, but E(36) and G(10) mean the Random Forest model predicted them wrongly. The same also goes Enrolled 54 were prided correctly, but D(21) and G(24) were false, and the graduate (G) got 408 correctly and D(37) and E(69) wrongly.

|  |
| --- |
| Total Observations in Table: 885    | test\_set$Target  predicted\_TargetRFM | D | E | G | Row Total |  --------------------|-----------|-----------|-----------|-----------|  D | 226 | 36 | 10 | 272 |  | 0.831 | 0.132 | 0.037 | 0.307 |  | 0.796 | 0.226 | 0.023 | |  | 0.255 | 0.041 | 0.011 | |  --------------------|-----------|-----------|-----------|-----------|  E | 21 | 54 | 24 | 99 |  | 0.212 | 0.545 | 0.242 | 0.112 |  | 0.074 | 0.340 | 0.054 | |  | 0.024 | 0.061 | 0.027 | |  --------------------|-----------|-----------|-----------|-----------|  G | 37 | 69 | 408 | 514 |  | 0.072 | 0.134 | 0.794 | 0.581 |  | 0.130 | 0.434 | 0.923 | |  | 0.042 | 0.078 | 0.461 | |  --------------------|-----------|-----------|-----------|-----------|  Column Total | 284 | 159 | 442 | 885 |  | 0.321 | 0.180 | 0.499 | |  --------------------|-----------|-----------|-----------|-----------| |

|  |
| --- |
| Overall Statistics    Accuracy : 0.7774  95% CI : (0.7485, 0.8044)  No Information Rate : 0.4994  P-Value [Acc > NIR] : < 2.2e-16    Kappa : 0.6235    Mcnemar's Test P-Value : 5.838e-09  Statistics by Class:  Class: D. Class: E Class: G  Sensitivity 0.7958 0.33962 0.9231  Specificity 0.9235 0.93802 0.7607  Pos Pred Value 0.8309 0.54545 0.7938  Neg Pred Value 0.9054 0.86641 0.9084  Precision 0.8309 0.54545 0.7938  Recall 0.7958 0.33962 0.9231  F1 0.8129 0.41860 0.8536  Prevalence 0.3209 0.17966 0.4994  Detection Rate 0.2554 0.06102 0.4610  Detection Prevalence 0.3073 0.11186 0.5808  Balanced Accuracy 0.8596 0.63882 0.8419 |

## 7.2 Support Vector Machine

SVM is a machine learning algorithm that finds the optimal hyperplane that best separates the data points of different classes in the feature space. When building the model, it is known that there are different types of kernels, e.g., linear, polynomial, radial, and sigmoid. All this was tested, but the best kernel was radial.

A support vector machine model was used, showing an accuracy of 0.7604 and a kappa value of 0.5945, which is moderate.

The cross-table below shows the amount of prediction the model got. It predicted 214 would drop out, which was correct, but E(32) and G(11) mean the Support Vector Machine model predicted them wrongly. The same also goes Enrolled 50 were prided correctly, but D(34) and G(22) were false, and the graduate (G) got 409 correctly and D(36) and E(77) incorrectly.

|  |
| --- |
| Total Observations in Table: 885    | test\_set$Target  predicted\_TargetSVM | D | E | G | Row Total |  --------------------|-----------|-----------|-----------|-----------|  D | 214 | 32 | 11 | 257 |  | 0.833 | 0.125 | 0.043 | 0.290 |  | 0.754 | 0.201 | 0.025 | |  | 0.242 | 0.036 | 0.012 | |  --------------------|-----------|-----------|-----------|-----------|  E | 34 | 50 | 22 | 106 |  | 0.321 | 0.472 | 0.208 | 0.120 |  | 0.120 | 0.314 | 0.050 | |  | 0.038 | 0.056 | 0.025 | |  --------------------|-----------|-----------|-----------|-----------|  G | 36 | 77 | 409 | 522 |  | 0.069 | 0.148 | 0.784 | 0.590 |  | 0.127 | 0.484 | 0.925 | |  | 0.041 | 0.087 | 0.462 | |  --------------------|-----------|-----------|-----------|-----------|  Column Total | 284 | 159 | 442 | 885 |  | 0.321 | 0.180 | 0.499 | |  --------------------|-----------|-----------|-----------|-----------| |

|  |
| --- |
| Overall Statistics    Accuracy : 0.7605  95% CI : (0.7309, 0.7882)  No Information Rate : 0.4994  P-Value [Acc > NIR] : < 2.2e-16    Kappa : 0.5945    Mcnemar's Test P-Value : 1.574e-09  Statistics by Class:  Class: D Class: E Class: G  Sensitivity 0.7535 0.3145 0.9253  Specificity 0.9285 0.9229 0.7449  Pos Pred Value 0.8327 0.4717 0.7835  Neg Pred Value 0.8885 0.8601 0.9091  Precision 0.8327 0.4717 0.7835  Recall 0.7535 0.3145 0.9253  F1 0.7911 0.3774 0.8485  Prevalence 0.3209 0.1797 0.4994  Detection Rate 0.2418 0.0565 0.4621  Detection Prevalence 0.2904 0.1198 0.5898  Balanced Accuracy 0.8410 0.6187 0.8351 |

## 7.3 K Nearest Neighbour

When building the KNN model, the number of k that was specified was the square root of the number of rows in the training dataset, then pick the closest odd number to it. After that, decide to change the numbers of K to different numbers to see the performance of the model, but the best still remains the number of K gotten from the square root K = 59.

The K nearest neighbour model was used, showing an accuracy of 0.7006 and a kappa value of 0.4558, which is moderate. The cross-table below shows the amount of prediction the model got. It predicted 169 would drop out, which was correct, but E(15) and G(3) mean the KNN model predicted them wrongly. The same also goes for Enrolled 13; they were prided correctly, but D(19) and G(1) were false, and the graduate (G) got 438 correctly and D(96) and E(131) wrongly.

|  |
| --- |
| Total Observations in Table: 885    | test\_set$Target  KNN | D | E | G | Row Total |  -------------|-----------|-----------|-----------|-----------|  D | 164 | 13 | 4 | 181 |  | 0.906 | 0.072 | 0.022 | 0.205 |  | 0.577 | 0.082 | 0.009 | |  | 0.185 | 0.015 | 0.005 | |  -------------|-----------|-----------|-----------|-----------|  E | 18 | 11 | 1 | 30 |  | 0.600 | 0.367 | 0.033 | 0.034 |  | 0.063 | 0.069 | 0.002 | |  | 0.020 | 0.012 | 0.001 | |  -------------|-----------|-----------|-----------|-----------|  G | 102 | 135 | 437 | 674 |  | 0.151 | 0.200 | 0.648 | 0.762 |  | 0.359 | 0.849 | 0.989 | |  | 0.115 | 0.153 | 0.494 | |  -------------|-----------|-----------|-----------|-----------|  Column Total | 284 | 159 | 442 | 885 |  | 0.321 | 0.180 | 0.499 | |  -------------|-----------|-----------|-----------|-----------| |

|  |
| --- |
| Overall Statistics    Accuracy : 0.7006  95% CI : (0.6692, 0.7306)  No Information Rate : 0.4994  P-Value [Acc > NIR] : < 2.2e-16    Kappa : 0.4558    Mcnemar's Test P-Value : < 2.2e-16  Statistics by Class:  Class: D Class: E Class: G  Sensitivity 0.5951 0.08176 0.9910  Specificity 0.9700 0.97245 0.4876  Pos Pred Value 0.9037 0.39394 0.6586  Neg Pred Value 0.8352 0.82864 0.9818  Precision 0.9037 0.39394 0.6586  Recall 0.5951 0.08176 0.9910  F1 0.7176 0.13542 0.7913  Prevalence 0.3209 0.17966 0.4994  Detection Rate 0.1910 0.01469 0.4949  Detection Prevalence 0.2113 0.03729 0.7514  Balanced Accuracy 0.7826 0.52711 0.7393 |

## 7.4 Decision Tree

Decision Tree model was use, showing and accuracy of 0.739 and a kappa value of 0.5691 which is moderate .

The cross-table below shows the amount of prediction the model got. It predicted 204 would drop out, which was correct, but E(27) and G(18) mean the Decision Tree model predicted them wrongly. The same also goes Enrolled 65 were predicted correctly, but D(48) and G(39) were false, and the graduate (G) got 484 correctly and D(32) and E(67) wrongly.

|  |
| --- |
| | test\_set$Target  predicted\_TargetDtree | D | E | G | Row Total |  ----------------------|-----------|-----------|-----------|-----------|  D | 204 | 27 | 18 | 249 |  | 0.819 | 0.108 | 0.072 | 0.281 |  | 0.718 | 0.170 | 0.041 | |  | 0.231 | 0.031 | 0.020 | |  ----------------------|-----------|-----------|-----------|-----------|  E | 48 | 65 | 39 | 152 |  | 0.316 | 0.428 | 0.257 | 0.172 |  | 0.169 | 0.409 | 0.088 | |  | 0.054 | 0.073 | 0.044 | |  ----------------------|-----------|-----------|-----------|-----------|  G | 32 | 67 | 385 | 484 |  | 0.066 | 0.138 | 0.795 | 0.547 |  | 0.113 | 0.421 | 0.871 | |  | 0.036 | 0.076 | 0.435 | |  ----------------------|-----------|-----------|-----------|-----------|  Column Total | 284 | 159 | 442 | 885 |  | 0.321 | 0.180 | 0.499 | |  ----------------------|-----------|-----------|-----------|-----------| |

|  |
| --- |
| Overall Statistics    Accuracy : 0.739  95% CI : (0.7087, 0.7676)  No Information Rate : 0.4994  P-Value [Acc > NIR] : < 2.2e-16    Kappa : 0.5691    Mcnemar's Test P-Value : 0.000644  Statistics by Class:  Class: D Class: E Class: G  Sensitivity 0.7183 0.40881 0.8710  Specificity 0.9251 0.88017 0.7765  Pos Pred Value 0.8193 0.42763 0.7955  Neg Pred Value 0.8742 0.87176 0.8579  Precision 0.8193 0.42763 0.7955  Recall 0.7183 0.40881 0.8710  F1 0.7655 0.41801 0.8315  Prevalence 0.3209 0.17966 0.4994  Detection Rate 0.2305 0.07345 0.4350  Detection Prevalence 0.2814 0.17175 0.5469  Balanced Accuracy 0.8217 0.64449 0.8238> |

## 7.5 Naive Bayes

Naive Bayes is an algorithm based on Bayes theorem, calculating the posterior probability of each class given the input features and selecting the class with the highest probability. Bernoulli Naive Bayes would be use to train the model because it is best suitable for binary features.

Naive Bayes model gives and accuracy of 0.680 and a kappa value of 0.4646 which is moderate.

|  |
| --- |
| Total Observations in Table: 885    | test\_set$Target  predicted\_Naive | D | E | G | Row Total |  ----------------|-----------|-----------|-----------|-----------|  D | 190 | 38 | 31 | 259 |  | 0.734 | 0.147 | 0.120 | 0.293 |  | 0.669 | 0.239 | 0.070 | |  | 0.215 | 0.043 | 0.035 | |  ----------------|-----------|-----------|-----------|-----------|  E | 46 | 39 | 38 | 123 |  | 0.374 | 0.317 | 0.309 | 0.139 |  | 0.162 | 0.245 | 0.086 | |  | 0.052 | 0.044 | 0.043 | |  ----------------|-----------|-----------|-----------|-----------|  G | 48 | 82 | 373 | 503 |  | 0.095 | 0.163 | 0.742 | 0.568 |  | 0.169 | 0.516 | 0.844 | |  | 0.054 | 0.093 | 0.421 | |  ----------------|-----------|-----------|-----------|-----------|  Column Total | 284 | 159 | 442 | 885 |  | 0.321 | 0.180 | 0.499 | |  ----------------|-----------|-----------|-----------|-----------| |

|  |
| --- |
| Overall Statistics    Accuracy : 0.6802  95% CI : (0.6484, 0.7109)  No Information Rate : 0.4994  P-Value [Acc > NIR] : < 2.2e-16    Kappa : 0.4646    Mcnemar's Test P-Value : 0.0001303  Statistics by Class:  Class: D Class: E Class: G  Sensitivity 0.6690 0.24528 0.8439  Specificity 0.8852 0.88430 0.7065  Pos Pred Value 0.7336 0.31707 0.7416  Neg Pred Value 0.8498 0.84252 0.8194  Precision 0.7336 0.31707 0.7416  Recall 0.6690 0.24528 0.8439  F1 0.6998 0.27660 0.7894  Prevalence 0.3209 0.17966 0.4994  Detection Rate 0.2147 0.04407 0.4215  Detection Prevalence 0.2927 0.13898 0.5684  Balanced Accuracy 0.7771 0.56479 0.7752 |

A table of numbers and a number of text

Description automatically generated with medium confidence

|  |  |  |
| --- | --- | --- |
| Model | Accuracy | Kappa |
| Random Forest | 0.7774 | 0.6235 |
| Support vector machine | 0.7605 | 0.5945 |
| KNN | 0.7006 | 0.4558 |
| Decision Tree | 0.739 | 0.5691 |
| Naïve Bayes | 0.6802 | 0.4646 |

# 8.0 Conclusion

In this work, we used advanced machine learning techniques, including data imputation and several ML models, in order to identify the factors that influence student dropout rates and identify at-risk students to focus more on the aspiration to reduce the dropout rate. The five models that were developed were tuned to the maximum limit, but there is also room for more improvement. I believe there are more factors that affect the academic success rate. That is why all the factors in this data set were used.

# Reference

Aized Amin Soofi, Arshad Awan, 2017. Classification Techniques in Machine Learning: Applications and Issues. J. Basic Appl. Sci. 13, 459–465. https://doi.org/10.6000/1927-5129.2017.13.76

Biau, G., Scornet, E., 2016. A random forest guided tour. TEST 25, 197–227. https://doi.org/10.1007/s11749-016-0481-7

Dekker, G.W., Pechenizkiy, M., Vleeshouwers, J.M., 2009. Predicting students drop out : a case study, in: Proceedings of the 2nd International Conference on Educational Data Mining, EDM 2009, July 1-3, 2009. Cordoba, Spain. pp. 41–50.

Kiss Botond, Nagy, M., Molontay, R., Csabay, B., 2019. Predicting Dropout Using High School and First-semester Academic Achievement Measures, in: 2019 17th International Conference on Emerging eLearning Technologies and Applications (ICETA). Presented at the 2019 17th International Conference on Emerging eLearning Technologies and Applications (ICETA), pp. 383–389. https://doi.org/10.1109/ICETA48886.2019.9040158

Kocsis, Á., Molnár, G., 2024. Factors influencing academic performance and dropout rates in higher education. Oxford Review of Education.

Nagy, M., Molontay, R., 2018. Predicting Dropout in Higher Education Based on Secondary School Performance, in: 2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES). Presented at the 2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES), pp. 000389–000394. https://doi.org/10.1109/INES.2018.8523888

Realinho, V., Machado, J., Baptista, L., Martins, M.V., 2022. Predicting Student Dropout and Academic Success. Data 7, 146. https://doi.org/10.3390/data7110146

Ukil Abhisek, 2007. Support Vector Machine, in: Intelligent Systems and Signal Processing in Power Engineering. Springer, Berlin, Heidelberg, pp. 161–226. https://doi.org/10.1007/978-3-540-73170-2\_4

Zhang Zhongheng, 2016. Introduction to machine learning: k-nearest neighbors. Annals of Translational Medicine 4. https://doi.org/10.21037/atm.2016.03.37

# APPENDICS

|  |  |
| --- | --- |
| Attached the Meta Data | <https://uelac-my.sharepoint.com/:x:/g/personal/u2442187_uel_ac_uk/EeaEZKH9F7FJgmXevCfRS6YBu3aYj9v55mkRJ0EJuAzHxw?e=cfWTON> |
| Attached the Data file | <https://uelac-my.sharepoint.com/:x:/g/personal/u2442187_uel_ac_uk/EejnFXEmAtFBkZj5QljLuLIBbraNpPe_Bsc8bS3ZVai2Cg?e=8JfMDd> |
| Attached the R script | <https://uelac-my.sharepoint.com/:u:/g/personal/u2442187_uel_ac_uk/Ed6Uxkx6mdtBizL9F63Ru7sBkQ3blhtMYDa5gXsApzK76Q?e=eeGRhD> |