Satakunnan ammattikorkeakoulu
Satakunta University of Applied Sciences

GAUTAM LOK MANI

Student retention analysis

DEGREE PROGRAMME IN DATA ENGINEERING
2021

ABSTRACT

Gautam, Lok: Student retention analysis
Data engineering degree programme
April 2024
Number of pages: 24

For colleges and universities, student attrition comes with substantial financial and reputational cost. This study aimed to analyse the different factors contributing for the student dropout and to develop a predictive model to identify students at risk of dropping out. Dataset encompassing 4424 students' records were analysed, including demographic factors, economic factors, Academic performance, social and special needs, and macro-economic factors.

Along with feature selection techniques, various machine learning algorithms were employed, among which Random Forest model achieved the highest accuracy (76%). Academic performance, Demographic factors and economic factors emerged as the key predictive factors for student success. These findings help to identify high-risk students and provide support and develop polices to foster a conductive learning environment.

This thesis utilizes exploratory data analysis (EDA) and machine learning techniques to forecast and explain the factors contributing for student dropout.

Keywords: attrition, dataset, machine learning algorithms, Exploratory data analysis (EDA)

CONTENTS

1 INTRODUCTION

The ongoing internationalization of higher education is fuelled by the influence of Western education systems and offers compelling benefits to institutions worldwide(Altbach & Knight, 2007). Finland, following the global trend, has experienced a significant increase in international students over the past two decades. The number of international students has tripled since 2001, with 31,913 international students enrolled in Finnish higher education institutions in 2019, accounting for 10% of the total student population(Lu & Everson Härkälä, 2024).

This thesis examines the open-source student data to identify the most significant factors influencing student retention. It aims to develop a nuanced understanding of the motivations behind students' decisions to stay at institutions and the challenges that may lead them to leave. This study aims to find concrete and actionable recommendations aimed at enhancing support systems and improving the overall experience students.

A combination of Exploratory data analysis (EDA), and machine learning algorithms will be employed. By analysing trends and exploring independent factors, this research seeks to illuminate the potential factors and the strategies institution can adapt for student success.

2 LITERATURE REVIEW

2.1 Overview of student retention

The concept of student retention, which emerged roughly three quarters of a century ago and refers to the multifaceted nature of a student's involvement in their studies. It not only shows the level of attentiveness students exhibit during their learning process but also how integrated and connected with their classes, among the peers, and throughout their college experience(Caruth, 2018). As student retention can be defined as an institution's ability to retain a student from admission through graduation, ensuring student success and institutional growth, understanding, and improving student retention is paramount(Haverila et al., 2020). Student retention involves the patterns of student enrolment, persistence, graduation, or drop out at a particular higher education institution. It describes how students progress through college over a defined time period(Leone & Tian, 2009).

The issue of student retention has become increasingly prominent throughout the history of higher education nationwide. Over time, its importance prompted administrators to delve into extensive research and explore what they can do to mitigate the number of students transferring from, or dropping out of, from their respective institutions (Leone & Tian, 2009). Institutions globally are being pushed to lower the rates of students 'dropping out' and devise fresh and innovative approaches that encourage students to continue (Thomas, n.d.). There are variety of reasons, why students withdraw from their studies. Past research, however, often find it's not just one thing; multiple personal and institutional factors typically intertwine to cause withdrawal(Haverila et al., 2020).

Generally, higher graduation rates reflect positively on the academic, administrative, and financial standing of institutions (Aljohani, 2016). However, enhancing student completion and retention rates can be a challenging task. Higher education establishments invest substantial sums annually to bring students to colleges or universities, while simultaneously witnessing a significant number depart within a single year (Leone & Tian, 2009). A potential solution toward this goal lies in adopting strategies and techniques that are informed by the findings of theoretical models and empirical studies (Aljohani, 2016). Individual student characteristics, including direction, determination, and dedication, play a crucial role in academic success. Students possessing these traits tend to take on heavier course loads, ultimately leading to graduation (Caruth, 2018).

2.2 Factors associated with student attrition.

According to the study conducted by Nieuwoudt & Kelly, Personal challenges such as family responsibilities, financial burdens, difficulties with time management, work obligations, and mental health struggles were frequently cited, while institutional factors like assessments, academic writing, and referencing requirements were also identified as significant contributors. Furthermore, many students reported experiencing feelings of fear, being overwhelmed, stressed, afraid of failing, anxious about assessments, or uncertain about expectations, which fuelled their thoughts of leaving. Negative interactions or experiences with supervisors, lecturers, or tutors also played a role in some students' contemplation of withdrawing from their studies (Nieuwoudt & Pedler, 2023).

Severiens and Wolff discovered that students who feel a sense of belonging, who are well connected with their peers and instructors and who who actively participate in extracurricular activities are more likely to graduate (Severiens & Wolff, 2008). This emphasizes the significant impact of student's social life beyond the academic realm on academic integration and success.

2.3 Methods utilized by different studies.

Exploring the intricate aspects of retaining international students requires a broad array of research methods. In this review, we examine the common techniques used to illuminate the factors that influence the decision of international students to stay or leave an institution.

Descriptive analysis was found to be one of the popular analysis techniques among many researchers. Nieuwoudt & Pedlerused employed descriptive analysis to explore the study's population characteristics. Similarly, it is used to analyse enrolment trends, demographic characteristics, and completion timelines within the international student population(Aljohani, 2016; Haverila et al., 2020). These analyses reveal fundamental patterns and insights into the dynamics of this specific student body.

Correlation Analysis had been in practice to explore potential links between different variables. For instance, (Rienties et al., 2012)have investigated correlations between Students' Adaptation to College Questionnaire (SACQ) components. The Pearson's Product Moment Correlation coefficient was employed to examine the strength of association between n the dependent variable, retention rate, and a set of independent variables, namely: student-to-faculty ratio, enrolment rate, institutional aid rate, default rate, and acceptance rate (Chiyaka et al., n.d.). Identifying such correlations aids in pinpointing factors that might positively or negatively impact retention.

Regression techniques are used to develop models capable of predicting student persistence. By considering a complex interplay of factors such as academic preparedness, social integration, and financial resources, researchers can estimate the probability of an international student continuing their studies (Caruth, 2018; Leone & Tian, 2009).

3 METHODOLOGY

3.1 Datasets

This dataset offers a deep dive into the students enrolled in diverse undergraduate programs at a higher education institution. It encompasses demographic details, socioeconomic factors, and academic performance data, providing valuable insights into potential predictors of both student attrition and academic achievement. Composed of multiple distinct databases, this dataset captures relevant information available upon enrolments, including application mode, marital status, course selection, and more. Additionally, it facilitates the estimation of overall student performance at the conclusion of each semester by evaluating curricular units credited, enrolled, evaluated, and approved, along with their corresponding grades. Furthermore, the dataset incorporates regional unemployment rate, inflation rate, and GDP data, enabling further exploration of how economic factors may influence student dropout rates and academic success outcomes (Devastator, 2023).

```
In [4]: data.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4424 entries, 0 to 4423
Data columns (total 35 columns):
 #   Column                                          Non-Null Count  Dtype
---  ------                                          --------------  -----
 0   Marital status                                  4424 non-null   int64
 1   Application mode                                4424 non-null   int64
 2   Application order                               4424 non-null   int64
 3   Course                                          4424 non-null   int64
 4   Daytime/evening attendance                      4424 non-null   int64
 5   Previous qualification                          4424 non-null   int64
 6   Nacionality                                     4424 non-null   int64
 7   Mother's qualification                          4424 non-null   int64
 8   Father's qualification                          4424 non-null   int64
 9   Mother's occupation                             4424 non-null   int64
 10  Father's occupation                             4424 non-null   int64
 11  Displaced                                       4424 non-null   int64
 12  Educational special needs                       4424 non-null   int64
 13  Debtor                                          4424 non-null   int64
 14  Tuition fees up to date                         4424 non-null   int64
 15  Gender                                          4424 non-null   int64
 16  Scholarship holder                              4424 non-null   int64
 17  Age at enrollment                               4424 non-null   int64
 18  International                                   4424 non-null   int64
 19  Curricular units 1st sem (credited)             4424 non-null   int64
 20  Curricular units 1st sem (enrolled)             4424 non-null   int64
 21  Curricular units 1st sem (evaluations)          4424 non-null   int64
 22  Curricular units 1st sem (approved)             4424 non-null   int64
 23  Curricular units 1st sem (grade)                4424 non-null   float64
 24  Curricular units 1st sem (without evaluations)  4424 non-null   int64
 25  Curricular units 2nd sem (credited)             4424 non-null   int64
 26  Curricular units 2nd sem (enrolled)             4424 non-null   int64
 27  Curricular units 2nd sem (evaluations)          4424 non-null   int64
 28  Curricular units 2nd sem (approved)             4424 non-null   int64
 29  Curricular units 2nd sem (grade)                4424 non-null   float64
 30  Curricular units 2nd sem (without evaluations)  4424 non-null   int64
 31  Unemployment rate                               4424 non-null   float64
 32  Inflation rate                                  4424 non-null   float64
 33  GDP                                             4424 non-null   float64
 34  Target                                          4424 non-null   object
dtypes: float64(5), int64(29), object(1)
memory usage: 1.2+ MB
```

Fig1: Data information

```
data['Target'].unique()

array(['Dropout', 'Graduate', 'Enrolled'], dtype=object)
```

The dataset has 4,424 student records. It contains 35 variables to analyse student outcomes. You'll find 34 independent variables and a single dependent variable called 'Target'. 'Target' tells you if the student dropped out, is still enrolled, or has graduated. All variables, except 'Target', are either integers or floats.

3.2 Data cleaning

This dataset seems to have undergone through preprocessing phase. My initial inspection confirmed that there were no missing values present in the dataset. Although, there was a variable misspelled which was corrected. After that I carefully verified that all the data types of variables are correct.

```
##Nationality is misspelled
data.rename(columns = {'Nacionality':'Nationality'}, inplace = True)
```

```
## Checking for missing values
data.isnull().sum().sum()
```

```
## check for correct data types
data.dtypes
```

```
Marital status                                    int64
Application mode                                  int64
Application order                                 int64
Course                                            int64
Daytime/evening attendance                        int64
Previous qualification                            int64
Nationality                                       int64
Mother's qualification                            int64
Father's qualification                            int64
Mother's occupation                               int64
Father's occupation                               int64
Displaced                                         int64
Educational special needs                         int64
Debtor                                            int64
Tuition fees up to date                           int64
Gender                                            int64
Scholarship holder                                int64
Age at enrollment                                 int64
International                                     int64
Curricular units 1st sem (credited)               int64
Curricular units 1st sem (enrolled)               int64
Curricular units 1st sem (evaluations)            int64
Curricular units 1st sem (approved)               int64
Curricular units 1st sem (grade)                float64
Curricular units 1st sem (without evaluations)    int64
Curricular units 2nd sem (credited)               int64
Curricular units 2nd sem (enrolled)               int64
Curricular units 2nd sem (evaluations)            int64
Curricular units 2nd sem (approved)               int64
Curricular units 2nd sem (grade)                float64
Curricular units 2nd sem (without evaluations)    int64
Unemployment rate                               float64
Inflation rate                                  float64
GDP                                             float64
Target                                           object
dtype: object
```

Fig2: Data types

Finally, I employed numerical encoding to the target variable to map each label to a corresponding numerical representation. Dropout is mapped to 0, Enrolled to 1 and Graduate to 2. This encoding is crucial because most machine learning model requires numerical inputs.

```
## converting Target variable
data['Target'].unique()
```

```
map_value = {'Dropout':0,
     'Enrolled':1,
     'Graduate':2}
data['Target'] = data['Target'].map(map_value)
print(data['Target'].unique())
```

3.3 Exploratory Data Analysis

Exploratory data analysis (EDA) is a method that employs descriptive statistics and visual representations to gain deeper insights into data (Camizuli & Carranza, 2018). In this section, I have incorporated EDA with descriptive statistics.

```
data_num = ['Application order','Age at enrollment', 'Curricular units 1st sem (credited)','Curricular units 1st sem (enrolled)',
data[data_num].describe()
```

	Application order	Age at enrollment	Curricular units 1st sem (credited)	Curricular units 1st sem (enrolled)	Curricular units 1st sem (evaluations)	Curricular units 1st sem (approved)
count	4424.000000	4424.000000	4424.000000	4424.000000	4424.000000	4424.000000
mean	1.727848	23.265145	0.709991	6.270570	8.299051	4.706600
std	1.313793	7.587816	2.360507	2.480178	4.179106	3.094238
min	0.000000	17.000000	0.000000	0.000000	0.000000	0.000000
25%	1.000000	19.000000	0.000000	5.000000	6.000000	3.000000
50%	1.000000	20.000000	0.000000	6.000000	8.000000	5.000000
75%	2.000000	25.000000	0.000000	7.000000	10.000000	6.000000
max	9.000000	70.000000	20.000000	26.000000	45.000000	26.000000

Fig3: Descriptive Statistics of numerical variable

Firstly, application order data suggests a possible rolling admissions process, with most students having lower application order numbers. The average age at enrolment was 23 years, with a notable spread, indicating a mix of fresh high school graduates and potentially returning learners or those entering higher education later in life. An analysis of first-semester curricular units shows disparities between enrolment, evaluation, and final approval. Focusing specifically on units approved, the average student successfully completed approximately five units in their first semester.

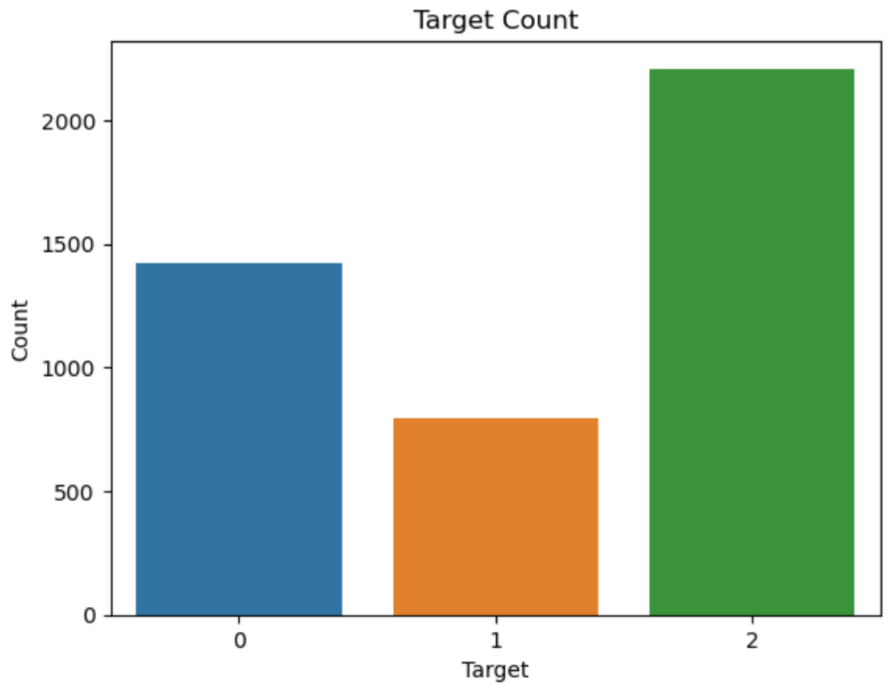

Fig4: Target variable counts

The above figure shows the count of different labels. The maximum number of students in this dataset contains students who graduated followed by students who dropped out and at last the students who enrolled.

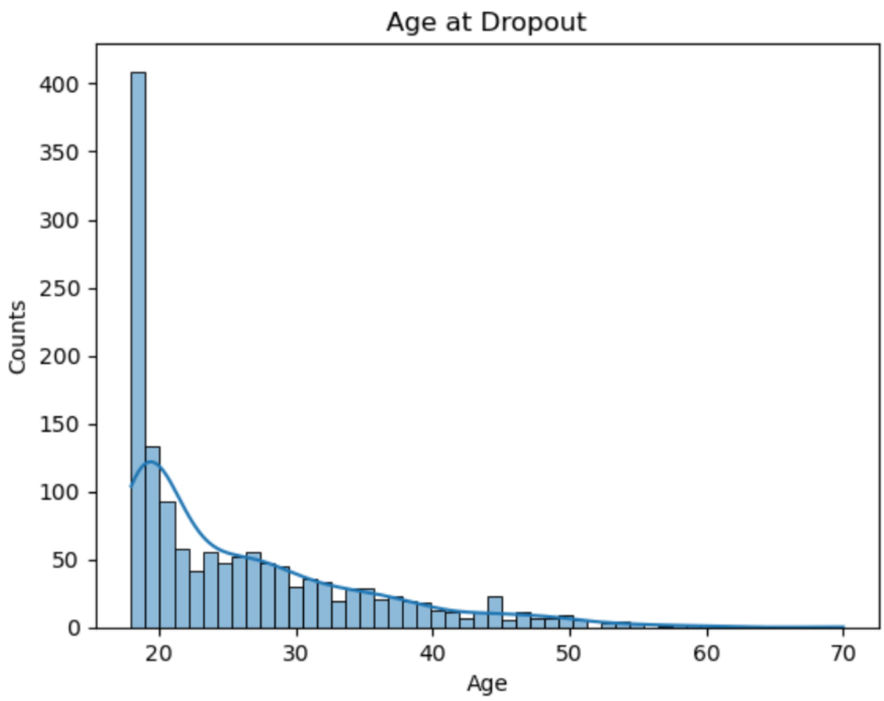

Fig5: Age of students Dropping out

According to this histogram, we can see that the students who dropped out the most were in the age group 17-20. Dropout rates is decreasing slowly as student age increases.

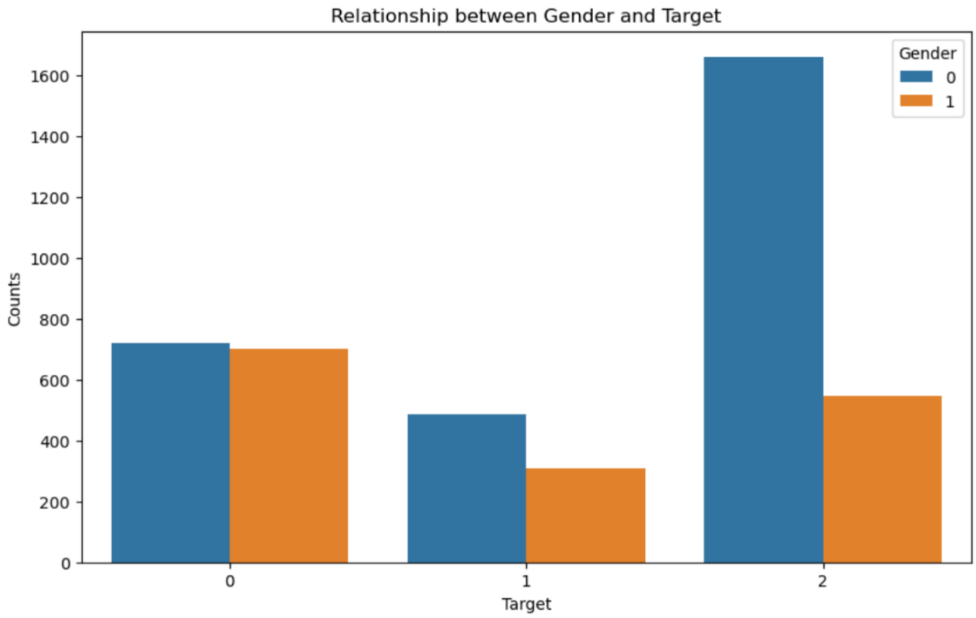

Fig6: Relationship between gender and target variables

The plot demonstrates a significant disparity between male and female students in graduation rates, female students who graduated are significantly higher than the male students. Additionally, male student’s enrolment appears slightly higher then then female. Interestingly, the dropout counts are almost similar for both the gender.

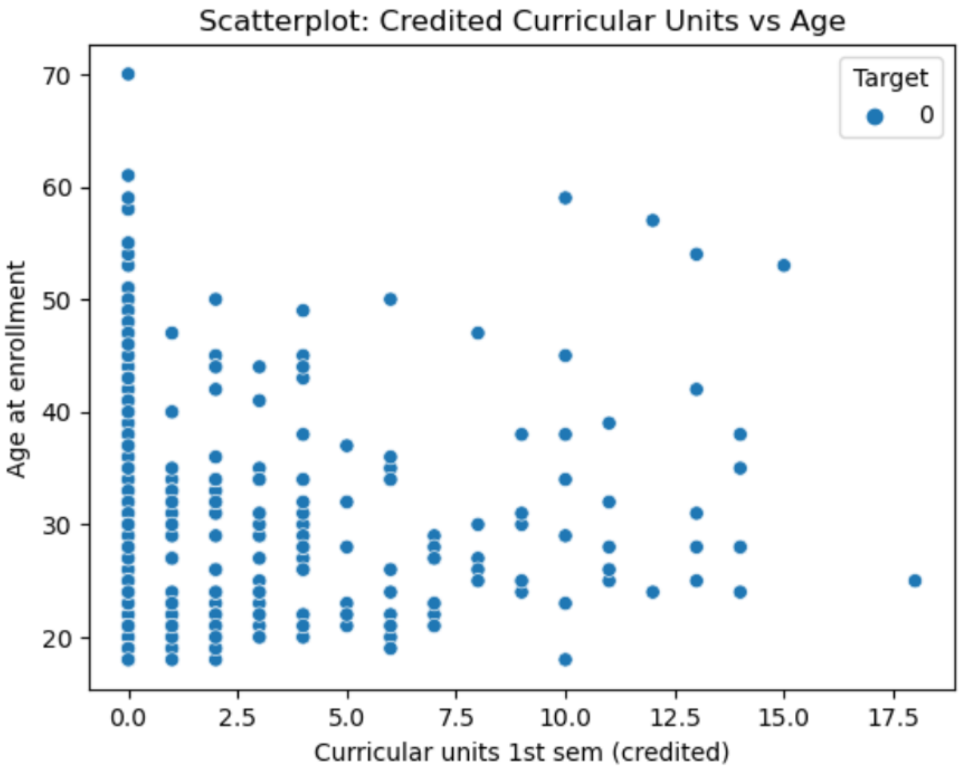

Fig7: Scatterplot between credited 1st semester units and Age

The scatterplot comparing credited curricular units (first semester) versus enrolment age for dropout instances reveals a higher likelihood of dropout among students with 0 to 5 credits earned in their first semester.

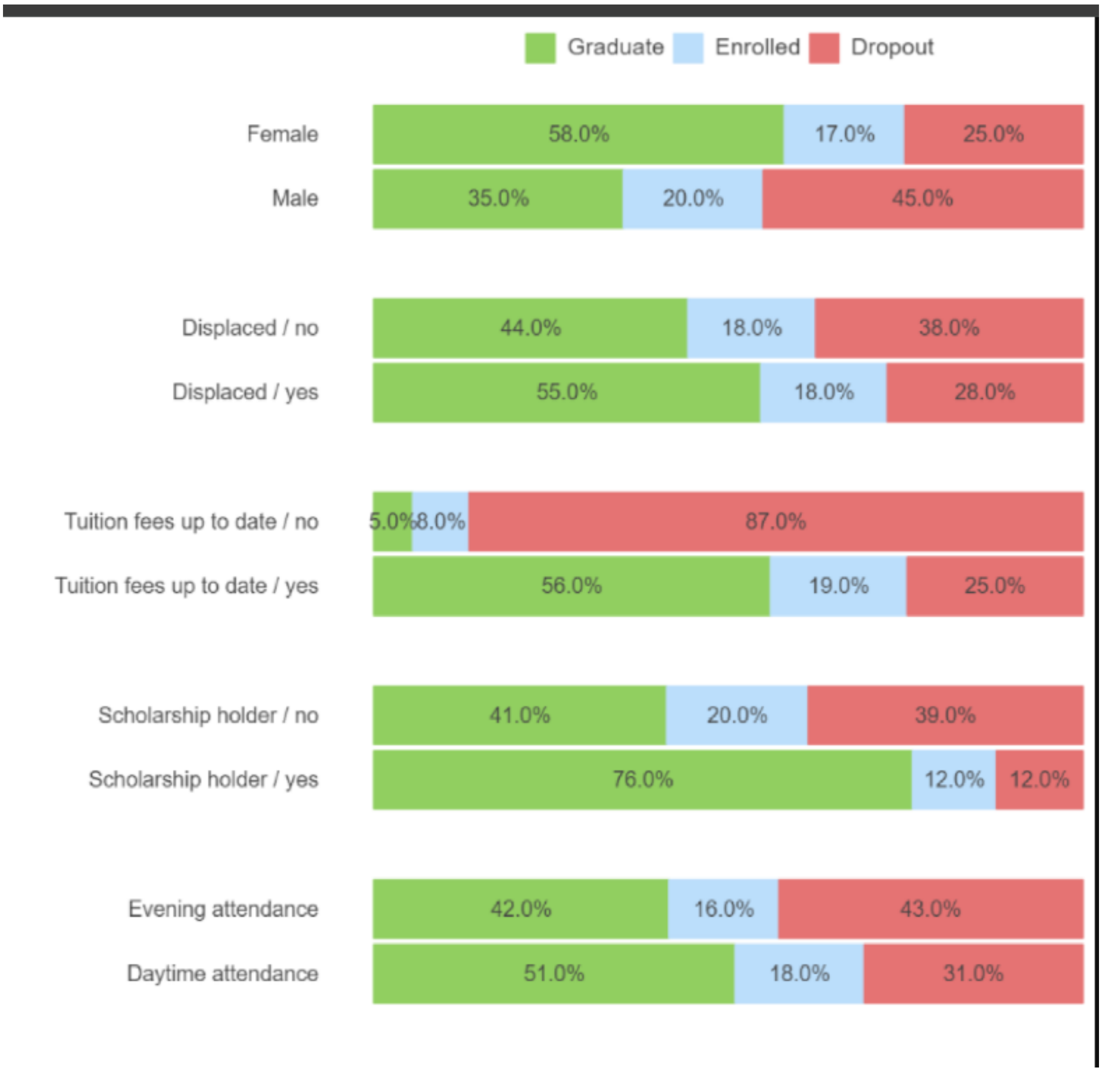

Fig8: Relationship between different categorical variables and the target variable. Source: https://www.mdpi.com/2306-5729/7/11/146

The above figure depicts that, students whose tuition fee is not up to date are more prone to dropout. Furthermore, it reveals that scholarship recipients have a higher likelihood of graduation compared to those without scholarships. Additionally, students who attended daytime studies experienced greater graduation success than evening class students.

3.4 Feature Selection

Feature selection, a longstanding area of interest in data analysis, has garnered considerable attention and research (DASH & LIU, 1997). Feature

Selection is the method that involves streamlining the input variables for a model by retaining only relevant data and eliminating noise data (Menon Kartik, 2024).

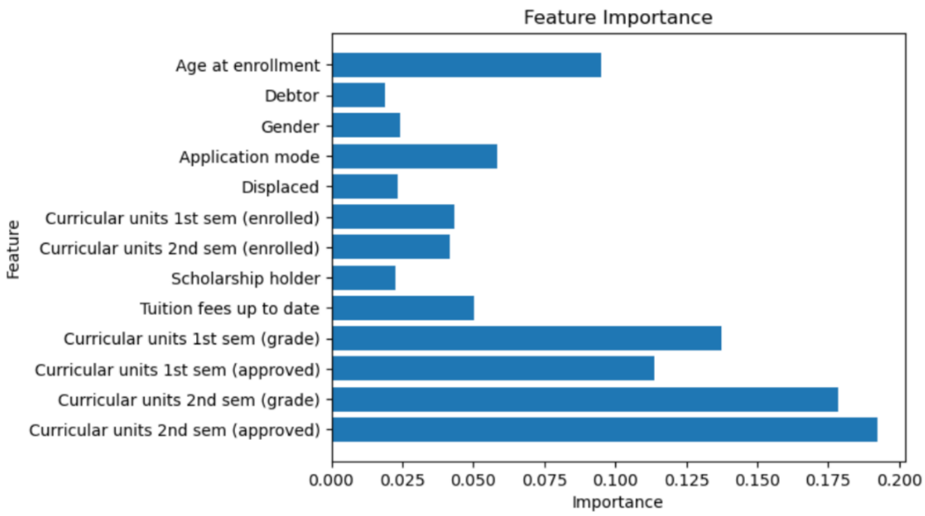

Fig9: Feature selection using Random Forest

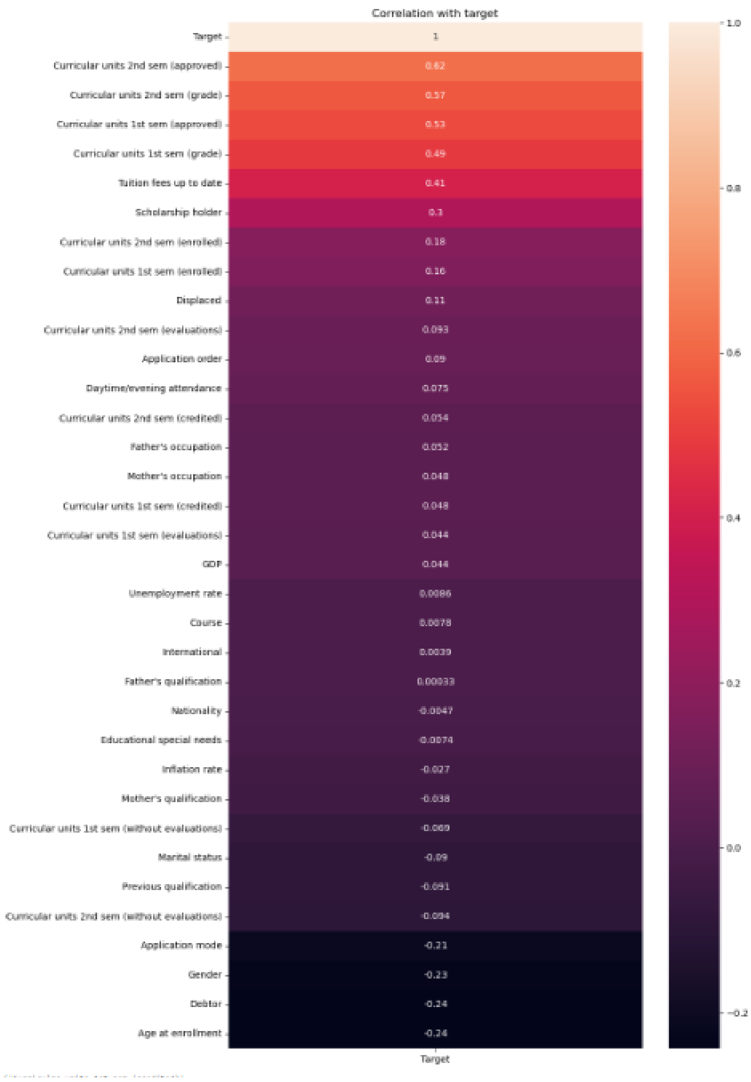

Fig10: Correlation with Target

To avoid noise, selecting features that strongly influence the target variable is important. Combination of correlation analysis with a Random Forest classifier was applied to select the most suitable features. Correlation analysis identifies linear relationships, while Random Forest effectively handles non-linear relationships and feature interactions. This combined approach allowed me to confidently select the top 14 most informative features for the machine learning task.

```
corr_features = data.corr()[['Target']].sort_values(by='Target', ascending=False)
top_15_features = corr_features.index.tolist()[:10]
last_4_features = corr_features.index.tolist()[-4:]
top_features = top_15_features + last_4_features

data2 = data.copy()
new_data = data2[top_features]
```

Both correlation method and Random Forest classifier picked out the same important features. Above code extracts the top 10 positive features and last 4 features, combines them, and stores into new variable “new_data”.

3.5 Model Building

```
## Splitting the data into Train and test sets
X = new_data.drop('Target', axis = 1)
y = new_data['Target']
```

```
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.20,random_state=42)
```

To prepare for model training and evaluation, dataset was split into training and testing sets. Firstly, independent features were assigned to variable 'X' and the dependent variable to 'y'. Utilizing the 'train_test_split' method from the scikit-learn library, 20% of the data was allocated for testing and the re-maining 80% for training. This split make sure that the model is trained on a representative portion of the data and evaluated on an unseen portion to as-sess its true performance.

3.5.1 Model selection

Various classification algorithms such as Logistic Regression, Decision Trees, Random Forests, K-Nearest Neighbors, and Support Vector Machines (SVM) were used along with grid search approach to fine tune hyperparame-ters and to select the best model. For each classifier, a pipeline was created, incorporating feature standardization using StandardScaler followed by the classifier itself. Grid search with 3-fold cross-validation was used to explore various hyperparameter combinations, aiming to maximize the accuracy score.

```
]: def train_classifier(clf, param_grid, X_train, y_train, X_test, y_test):
       pipe = Pipeline([('scl', StandardScaler()), ('clf', clf)])
       grid_search = GridSearchCV(pipe, param_grid, scoring='accuracy', cv=3)
       grid_search.fit(X_train, y_train)
       test_accuracy = grid_search.score(X_test, y_test)
       best_params = grid_search.best_params_
       return test_accuracy, best_params
```

```
]: classifiers = [
       (LogisticRegression(), {'clf__penalty': ['l1', 'l2'], 'clf__C': [ 0.1, 1, 10], 'clf__solver': ['liblinear']}),
       (DecisionTreeClassifier(), {'clf__criterion': ['gini', 'entropy'], 'clf__min_samples_leaf': [1, 2, 3, 4, 5], 'clf__max_d
       (RandomForestClassifier(), {'clf__min_samples_leaf': [1, 2, 3, 4, 5], 'clf__max_depth': [1, 2, 3, 4, 5], 'clf__min_sampl
       (KNeighborsClassifier(), {'clf__n_neighbors': [1, 2, 3, 4, 5], 'clf__weights': ['uniform', 'distance'], 'clf__metric': [
       (SVC(), {'clf__kernel': ['linear', 'rbf'], 'clf__C': [1,2,3,4,5, 6]})

   ]

   trained_classifiers = []

   for clf, param_grid in classifiers:
       test_accuracy, best_params = train_classifier(clf, param_grid, X_train, y_train, X_test, y_test)
       best_clf = clf.__class__()  # Create a new instance of the classifier with best parameters
       trained_classifiers.append((type(clf).__name__, best_params))
       print(f'{type(clf).__name__} Test Accuracy: {test_accuracy}')
       print(f'{type(clf).__name__} Best Params: {best_params}')
       print()
```

```
LogisticRegression Test Accuracy: 0.7480225988700565
LogisticRegression Best Params: {'clf__C': 100, 'clf__penalty': 'l1', 'clf__solver': 'liblinear'}

DecisionTreeClassifier Test Accuracy: 0.7378531073446327
DecisionTreeClassifier Best Params: {'clf__criterion': 'gini', 'clf__max_depth': 5, 'clf__min_samples_leaf': 5, 'clf__min_sampl
es_split': 15}

RandomForestClassifier Test Accuracy: 0.7627118644067796
RandomForestClassifier Best Params: {'clf__max_depth': 12, 'clf__max_features': 'log2', 'clf__n_estimators': 100}

KNeighborsClassifier Test Accuracy: 0.7333333333333333
KNeighborsClassifier Best Params: {'clf__metric': 'manhattan', 'clf__n_neighbors': 5, 'clf__weights': 'uniform'}

SVC Test Accuracy: 0.752542372881356
SVC Best Params: {'clf__C': 100, 'clf__gamma': 0.01, 'clf__kernel': 'rbf'}
```

```
]:
```

Among these classifiers, Random Forest classifier performed the best on test data with an accuracy score of 76 percent, followed by Support Vector machines (SVC) and Logistic Regression with the accuracy score of 75 percent and 74 percent respectively. Decision tree classifier performed the least among all these algorithms.

4 GUIDELINES FOR EDUCATIONAL INSTITUTION

From this research, various significant factors have been identified, which can be used by educational institution to further investigate, aiming to enhance student retention and overall organizational effectiveness.

Presented below are actionable guidelines derived from the findings.

1. Financial Support and tuition fees: Since, unpaid tuition fees strongly corelates with increased dropout risk, educational institutions can implement a system that identifies the students who have tuition fee due. Institution can also give financial counselling workshops, advising for budgeting and fee management.
2. Scholarship Expansion: Scholarship holders have higher graduation rates, so organization should promote more scholarship programs (merit-based, need-based, field-specific).
3. More attention to younger students: Older students (25+) demonstrate better retention than younger students. Institution can offer tailored advising, focusing on study skills and time management.
4. Academic Success: There was a high correction between students who dropout and curricular units' status (approved, grade, evaluation). Therefore, tracking early academic performance to identify students struggling with courses and increasing tutoring availability in core subjects could be a good approach.
5. Other Guidelines: Institutions can strengthen students' desire to stay by offering guidance on social integration through activities like community service, social events, and dances (Haverila et al., 2020). According to research conducted by Haverila et al., indicates that counselling, extracurricular activities, housing services, study workshops, and other support initiatives are crucial for fostering a positive experience for both international students and domestic students, ultimately encouraging retention (Haverila et al., 2020).

5 CONCLUSION

With the goal of identifying influential factors for student success this thesis explores the student dataset through Exploratory Data Analysis (EDA) and with the development and evaluation of classification models. A significant association was observed between the target variable and several factors, including academic performance metrics like 'Curricular units 2nd sem (approved)', 'Curricular units 2nd sem (grade)', 'Curricular units 1st sem (approved)', and 'Curricular units 1st sem (grade)'. Additionally, financial indicators such as 'Tuition fees up to date' and 'Scholarship holder' showed a strong correlation. Demographic factors like 'Displaced', 'Application mode', 'Gender', 'Debtor', and 'Age at enrollment' also exhibited a notable relationship with the target variable. These variables were used to predict the student dropout.

Likewise, in the modelling phase we conducted thorough hyperparameter tuning across a varied array of classification algorithms, such as Logistic Regression, Decision Trees, Random Forests, K-Nearest Neighbors, and Support Vector Machines. The highest test accuracy achieved was 76 percent, which shows the potential to early identification of students at risk of dropout.

Achieving higher accuracy may require the additional ingestion of data and more sophisticated modelling techniques and algorithms. This research sets the stage for further improvement and highlights the role of machine learning and data analytics in improving student success.

6 REFERENCES

Aljohani, O. (2016). *Higher Education Studies*. *6*(2). https://doi.org/10.5539/hes.v6n2p1

Altbach, P. G., & Knight, J. (2007). The Internationalization of Higher Education: Motivations and Realities. *Journal of Studies in International Education*, *11*(3–4), 290–305. https://doi.org/10.1177/1028315307303542

Camizuli, E., & Carranza, E. J. (2018). Exploratory Data Analysis (EDA). *The Encyclopedia of Archaeological Sciences*, 1–7. https://doi.org/10.1002/9781119188230.SASEAS0271

Caruth, G. D. (2018). Student Engagement, Retention, and Motivation: Assessing Academic Success in Today's College Students. *Participatory Educational Research (PER)*, *5*(1), 17–30. https://doi.org/10.11203/per.18.4.5.1

Chiyaka, E. T., Sithole, A., Manyanga, F., Mccarthy, P., & Bucklein, B. K. (n.d.). *Institutional Characteristics and Student Retention: What Integrated Postsecondary Education Data Reveals About Online Learning*.

DASH, M., & LIU, H. (1997). Feature selection for classification. *Intelligent Data Analysis*, *1*(1–4), 131–156. https://doi.org/10.1016/S1088-467X(97)00008-5

Haverila, M. J., Haverila, K., & Mclaughlin, C. (2020). Variables Affecting the Retention Intentions of Students in Higher Education Institutions: A Comparison Between International and Domestic Students. *Peer-Reviewed Article © Journal of International Students*, *10*(2), 2166–3750. https://doi.org/10.32674/jis.v10i2.1849

Leone, M., & Tian, R. G. (2009). Push Vs Pull: Factors Influence Student Retention. *American Journal of Economics and Business Administration*, *1*(2), 122–132.

Lu, W., & Everson Härkälä, T. (2024). International student experience of employment integration in Finland. *Research in Comparative and International Education*, 17454999241238172. https://doi.org/10.1177/17454999241238172

Menon Kartik. (2024). *Feature Selection In Machine Learning [2024 Edition] - Simplilearn*. https://www.simplilearn.com/tutorials/machine-learning-tutorial/feature-selection-in-machine-learning

Nieuwoudt, J. E., & Pedler, M. L. (2023). Student Retention in Higher Education: Why Students Choose to Remain at University. *Journal of College Student Retention: Research, Theory and Practice*, *25*(2), 326–349. https://doi.org/10.1177/1521025120985228

Rienties, B., Beausaert, S., Grohnert, T., Niemantsverdriet, S., & Kommers, P. (2012). Understanding academic performance of international students: The role of ethnicity, academic and social integration. *Higher Education*, *63*(6), 685–700. https://doi.org/10.1007/S10734-011-9468-1/TABLES/4

Severiens, S., & Wolff, R. (2008). A comparison of ethnic minority and majority students: Social and academic integration, and quality of learning. *Studies in Higher Education*, *33*(3), 253–266. https://doi.org/10.1080/03075070802049194

Thomas, L. (n.d.). *Improving Student Retention in Higher Education*.

Devastator, T. (2023). Predict students' dropout and academic success. *Predict students' dropout and academic success*, 3. Retrieved from https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention/data