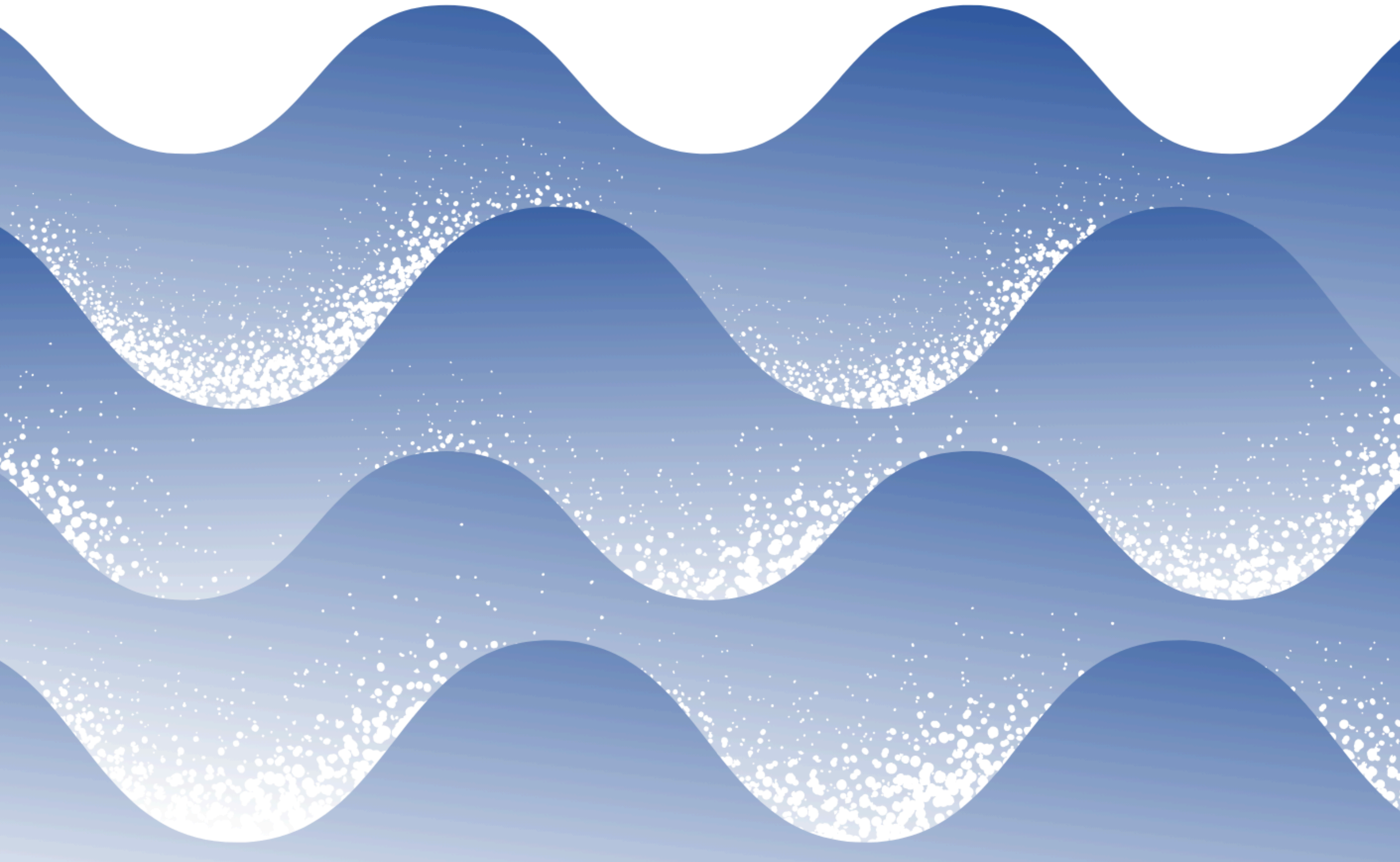
# HOW TO PREDICT THE DROPOUT RATE OF THE STUDENTS?



LUCAS SCHREIJ (1026157)
JIA QIANG LI (1035208)
TAWAN HARINK (1009798)
JASMIN ABDULLAHI (1001453)
VIOLET JIANG (0989421)
MELLE HERLAAR (1029218)

CMI-MINBOD-2425
Souravlias, D.
Dobrkovic, A
Wypych, A.
First version
19-01-2025

**Table of Contents**

# Chapter 1: Introduction

## 1.1 Background

Dropout rates have been long used as the main indicator in educational systems worldwide to give an impression of a program's quality and experience it offers. High dropout rates often signify low student satisfaction and poor institutional backing. Whereas low dropout rates are often considered to display a strong student support system, a positive environment and a great alignment between student expectations and program delivery (Hadjar et al., 2022).

Although students might believe dropping out only influences their academic life, it's influence reaches much further than just the academic setting. Various studies have shown that students who drop out are more likely to face unemployment and have limited opportunities for personal growth in comparison with students that finish their degree. Therefore, being able to identify the key factors involved in a student's decision to dropout and potentially predict a class's dropout rate is of very high importance (Patzina & Wydra-Somaggio, 2020).

When it has become more clear why students drop out, institutions are able to identify at-risk students at a more early stage and intervene effectively. By identifying factors that influence the dropout rate, analyzing distinct patterns using data machine learning techniques and determining the earliest indicators of a student potentially dropping out, this study aims to predict student dropout rates.

## 1.2 Research Question

This study focuses on answering the following main research question:

*How to predict the dropout rate of the students?*

Student dropouts can cause problems in the educational systems, they have a direct effect on people's future opportunities and affect the overall economy as well. The question can recognize and reduce dropout risks which improves the educational outcomes. Therefore, we decided to predict the student dropout rate as it suited the dataset the most.

To answer this question, two sub-questions were formulated, allowing the main question to be answered in a step by step manner:

1. What factors influence the dropout rate?
2. How early can we detect warning signs?

The first sub-question outlines the important factors that affect students' dropout rate such as academic performance. For the second sub-question, this question helps us address observable characteristics or patterns that can be used to identify the students who are about to drop out or close to. These early signs are very important, how sooner the risks of a student are identified, the greater the chances of preventing the dropout. Answering all the subquestion helps the study build a better approach to predicting student dropouts.

## 1.3 Methodology

For the research methodology, a structured approach was followed, which is divided into four milestones. First it starts with data storytelling: During this initial phase, the focus is to understand the data thoroughly, ensuring the dataset is complete and aligned with the research topic. During this a masterfile was created of all the merged worksheets. Only the most important columns were included, such as grades, attendance etc. Merging all the important files made it easier to compare data and gain a comprehensive overview. A bar graph was created to compare grades between first-chance and second-chance scores. For attendance, data was compared from ANL1 and ANL2. ANL1 and ANL2 can be seen as periods in a school year. A school year is made up of ANL1, ANL2, ANL3 and ANL4, so each of these can be seen as a period in a school year.

During the next phase the idea is to identify a dependent variable of high relevance and determine if there is a correlation between the dependant and the independent variable(s). Besides that, the linearity is proven by means of graphs and hypothesis testing is carried out. Correlation testing was done between grades and dropout rate, attendance percentage and dropout rate and homework grades and drop out rate. Dropout rate being the dependent variable. For each of them a hypothesis test was also made. In result,  a logistic regression graph was made to illustrate the correlation between the dependent and independent variables.

During the literature review, credible papers addressing the problem of "Predicting student dropout rates" were researched for their relevance to the main research question. After carefully selecting these sources, a summary was created, and the literature was categorized. Data mining algorithms and tools used in the studies were analyzed to identify those most compatible with the dataset. To provide a clear and thorough understanding of the options, the advantages and disadvantages of each algorithm were outlined. Finally, a conceptual model was developed and presented in response to the main research question.

The final phase involved experimentation, where data mining techniques were applied to explore the dataset. Three different algorithms were selected and compared: Logistic Regression, Random forest, and Decision trees. The results were analyzed by using the following statistical metrics: accuracy, precision, recall, F1-score, presented clearly, and explained in detail.

## Chapter 2: Literature review

This chapter provides an insight into the methods used to research how to predict student dropout. In order to gather more insight about the problem, it is crucial to research previous solutions. A structured and thorough literature review is performed by employing the research method (Verhoeven, 2022). The first step is to define a main research question alongside sub questions that will be answered through the literature review. Following that, the scope is defined by searching literature using search queries and selection process. In the third step, documents containing relevant academic information are collected, and after that the most important information is extracted. Finally, the collected information is shown in a relevance table.

### 2.1 Literature search and selection process

Utilizing well-formulated research questions helps narrow down the scope of the study to make a predictive model for predicting student dropout. As a result, the research can remain relevant and manageable (Verhoeven, 2022). In order to research the student dropout prediction a main research question was formed. The main research question is “How to predict the dropout rate of the students?”. This main research question will help identify the key factors influencing the student drop out and developing effective predictive models. In order to answer the main research question, two sub questions have been formulated:

1. What factors influence the dropout rate?
2. How early can we detect warning signs?

After formulating these research questions, a search for prior papers that have examined the same subject or a related one can be conducted. The usage of search queries was employed. The search queries helped gather relevant information to answer the research question. The table below shows which search queries and filters were utilized to receive accurate and relevant results.

| | Site | Query | Filter | Article |
|---|---|---|---|---|
| 1 | https://hogeschoolrotterdam.on.worldcat.org/ | “Predicting” AND “Dropout” | **Search within libraries:** Libraries worldwide<br>**Publication year:** 2019-2024 | Full article: Predicting student dropout: A machine learning approach |
| 2 | Google Scholar | “Predicting” AND “ Student” AND “Dropout” | **Publication year:** 2019-2024 | Predicting dropout from higher education: Evidence from |

| | | | | |
|---|---|---|---|---|
| | | | | Italy - ScienceDirect |
| 3 | Google | “Prediction model” AND “ subscription” | - | Predicting Subscription Churn Using PySpark ML \| by Noemi Ramiro \| Analytics Vidhya \| Medium |
| 4 | Google | “Predicting student” AND “dropout” | **Publication year:** 2019-2024 | Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study |
| 5 | https://hogeschoolrotterdam.on.worldcat.org/discovery | “Predicting student dropout machine learning” | **Publication year:** 2020-2024 | Predicting Student Drop-Out in Higher Institution Using Data Mining Techniques |
| 6 | Google | “Predicting student dropout machine learning” | **Publication year:** 2019-2024 | A Survey of Machine Learning Approaches and Techniques for Student Dropout Prediction |

*Table 1: Search log table*

## 2.2 Information retrieved from search

### 2.2.1 Predicting student dropout: A machine learning approach

This study used a dataset only based on individual student performance, preventing the usage of possible privacy related information and discrimination based on student’s non-study-related information. The combined features such as count of passed exams, count of failed exams and average grades were the most relevant single factors to predict academic success. The machine learning algorithms employed to make a predictive model were logistic regression and classification trees. Logistic regression preprocessing and model interpretation are intricate processes. On the other hand, although classification trees are prone to over- and underfitting, the model is deemed easier to compute and interpret (Kemper et al., 2019).

### 2.2.2 Predicting dropout from higher education: Evidence from Italy

The goal of this paper is to evaluate the machine learning models as early warning systems to identify the students who are at-risk of dropping out. The paper included socio-economic issues into their dataset. The machine learning models Random Forest (RF), Least Absolute Shrinkage and Selection Operator (LASSO), Gradient Boosting Machines (GBM), and Neural Networks (NN) were employed. RF and GBM achieved high accuracy and robustness. The machine learning models proved effective for predicting the dropout rate of students (Delogu et al., 2024).

### 2.2.3 Supervised machine learning algorithms for predicting student dropout and academic success; a comparative study

This comprehensive study compared both supervised and unsupervised ML algorithms to predict student dropout rates as well as their academic success. The paper puts a heavy emphasis on the exceptional performance of boosting algorithms, such as CatBoost and LightGBM. According to the paper, boosting algorithms surpass traditional methods like logistic regression and decision trees. However, it’s important to note that boosting algorithms lack interpretability, leading the researchers to employ the SHAP method. The study concluded that there is no substantial difference in the predictive performance of LightGBM and CatBoost and has confirmed that these boosting algorithm models are very proficient in distinguishing different classes within the dataset (Villar & De Andrade, 2024).

### 2.2.4 Predicting student Drop-Out in Higher Institution Using Data Mining Techniques

The paper describes the usage of various ML models within a data mining framework to predict student dropout rate of undergraduate students after 3 years of enrolment in a specific discipline. Their dataset consisted of 64 students collected from two sources; the admission center where minimal demographic information (f.e. gender, birth date) was retrieved. And the transcript records of each student which included the courses taken and their respective grades. Five different types of algorithms: k-NN, Decision Tree, Neural Network, Logistic Regression and Random Forest were investigated for this specific data set. Among these algorithms, the paper concluded that Logistic Regression outperformed the other four with better accuracy and a more comprehensive classification approach (Yaacob et al. 2020).

### 2.2.5 Predicting Subscription Churn Using PySpark ML

The article focuses on predicting customers who are mostly likely to stop subscribing to a platform. This is similar to predicting student dropout, as both involve identifying individuals at risk of disengaging based on behavioral data. In order to predict the subscription churn, logistic regression, random forest classifier and gradient-boosted tree classifier were applied.

The models were evaluated based on their performance based on the metrics accuracy and F-1 score, which showed that random forest outperformed the other machine learning algorithms (Ramiro, 2021).

### 2.2.6 A Survey of Machine Learning Approaches and Techniques for Student Dropout Prediction

This paper shows how machine learning can help predict the student dropout. First it's noticed that In developing countries the research on applying these methods is restricted for addressing this problem. The use of data machine learning has grown significantly, despite that many researchers also overlook the problem of imbalanced data, this leads to less reliable data. Many researches focus on early prediction, only a few include ranking or forecasting methods to address the problem. Additionally, the paper recommends identifying at-risk schools for early intervention by using school-level data. Several algorithms are used to predict student dropouts such as: Decision Trees, Random Forests, Support Vector Machines (SVM), Neural Networks, and Logistic Regression.

The table below shows what pages include the most important information to answer the main research question. Furthermore, it includes possible machine learning models that offer as a solution to the main research question, alongside the most effective and relevant ML model(s).

| | **Author** | **Title** | **Most important pages** | **Possible solutions** | **Most relevant solutions** |
|---|---|---|---|---|---|
| 1. | Kemper et al., 2019 | Predicting student dropout: A machine learning approach | p33-p41 & p44 | Logistic regression<br>Classification trees | Logistic regression<br>Classification trees |
| 2. | Delogu et al., 2024 | Predicting dropout from higher education: Evidence from Italy | p5-9 | Random Forest<br>Least Absolute Shrinkage and Selection Operator<br>Gradient Boosting Machines | Random Forest |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
|  |  |  |  | Neural Networks (NN) |  |
| 3. | Ramiro, 2021 | Predicting Subscription Churn Using PySpark ML | p1-6 | Logistic regression, Random forest, Gradient-boosted machines | Logistic regression,<br>Random forest |
| 4. | Villar & De Andrade, 2024 | Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study | p4-12<br>p21-23 | Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting Extreme, Gradient Boosting (XGBoost), CatBoost Light Gradient Boosting Machine | Decision Tree<br>Random Forest |
| 5. | Yaacob et al. 2020 | Predicting Student Drop-Out in Higher Institution Using Data Mining Techniques | p8-11 | Logistic Regression<br>Random Forest<br>Decision Tree<br>Artificial Neural Network (ANN)<br>k-Nearest Neighbor (k-NN) | Logistic Regression<br>Random Forest |

| | | | | | |
|---|---|---|---|---|---|
| 6. | Mduma et al. 2019 | A Survey of Machine Learning Approaches and Techniques for Student Dropout Prediction | p4-6 | Logistic regression<br>Decision tree<br>Random forest<br>Neutral Networks<br>Support Vector Machines | Logistic regression<br>Decision tree<br>Random forest |

*Table 2: Literature categorization table*

## 2.3 Gaps in the research

The study of Ramiro (2021) did not employ any advanced model feature to help interpret the predictions or any hyperparameter tuning. Moreover, the study of Mduma et al. (2019) focuses solely on students who are enrolled at the University of Bangladesh. This limits the generalizability of the findings. Although the study shows the results of the predicted student dropout rate, it did not provide any interventions that the school can take in order to help students who are predicted to drop out. Additionally, the study does not discuss any ethical challenges associated with predictive modeling, such as the risk of reinforcing biases in the data or the potential misuse of predictions to unfairly penalize students, such as minority groups. Furthermore, the study of Kemper et al. (2019) solely relied on two traditional machine learning approaches: logistic regression and decision trees, limiting the potential for exploring more advanced techniques that could potentially provide more accurate predictions.

Yaacob et al. (2020) also did not mention employing personal interventions to reduce student dropout. Moreover, the research does not investigate how dropout risks evolve over time, which could be critical for timely interventions. Additionally, the utilized dataset consists of 64 students, which could limit the reliability and robustness of the results. The study of Villar & De Andrade (2024) has not provided information on how to interpret the results of the paper, which is crucial for readers who do not have machine learning knowledge. Finally, the findings of Delogu et al. (2024) are limited to Italy, as they only researched Italian universities. A comparative analysis with other countries, such as the Netherlands, would help contextualize the predictors and validate the model's robustness.

The aforementioned gaps, explained in the first two paragraphs, were addressed by taking certain steps in this paper. This dataset did not contain privacy-related information and non-academic factors like socioeconomic status and mental health status, which helps objectively evaluate the impact of academic performance predictors on the dropout rate. Furthermore, the dataset includes 333 students, which is larger than the dataset of Yaacob et al. (2020), providing a more comprehensive sample for analysis. This larger sample size increases the reliability and generalizability of the findings in relation to academic performance predictors and dropout rates. Additionally, to combat the lack of generalizability of the papers to a study in the Netherlands, this paper will construct a model that is applicable to all the facilities of University of applied science Rotterdam, which will facilitate the creation of a model that can be applied to all universities in the Netherlands.

In order to observe how the dropout risk evolves, this study will look at how the risk of dropping out evolves over four quarters. Furthermore, this paper will facilitate replicating the study by incorporating an extensive explanation on how to interpret the findings. Apart from solely employing logistic regression and decision trees,  random forest will be employed to improve the accuracy of the dropout risk prediction model. This approach will provide a more robust analysis and enhance the overall reliability of the study's findings. Lastly, the paper will mention interventions that can be taken to provide assistance to students who are at risk of dropping out.

## Chapter 3: Solution design

This chapter dives into the steps and methodologies employed to design and develop the student dropout predictor. It covers the data set, preparation process and the machine learning algorithm selection process. This section offers an understanding of the design solutions used.

### 3.1 Dataset and preprocessing

The dataset used for this study was acquired from a data science minor course. The dataset itself was collected from the Informatica course 2019/2020 at Hogeschool Rotterdam, and is very detailed. It consists of multiple excel files that contain relevant information collected during the course such as previous education levels, grades and attendance throughout the year. The dataset has two excel files where students are mentioned with their student id. The first excel file has a total of 2486 rows and the second excel file has a total of 4640 rows, however these contain a lot of duplicates of students. When the two excel files are filtered on unique students and students that are studying computer science full time (INFV) or parttime (INFD), there are 333 rows of students left. The dataset allows for the estimation of student performance each semester  by considering grades, attendance and homework made. The student dropout rate of first year computer science students can be determined. To achieve this, the dataset will be preprocessed, each feature will be analyzed, and all useful information will be consolidated into a “clean” master file that serves as the foundation for the analysis. The master file included the following data:

| Feature name | Description |
|---|---|
| Student ID | School given ID. |
| Full name | Full name of the student. |
| Previous education | Name of previous school. |
| Graduated | J if a student graduated, N if not. |
| Full time / part time | VT if a student is full time, DT if part time. |
| Dropout date | Date student dropped out, empty if student hasn’t dropped out. |
| ANL1-ANL4 attendance rates | Attendance on a weekly basis for every period. |
| ANL1-ANL4 homework grades | Homework grades on a weekly basis for every period. |
| ANL 1-ANL4 exam grades | Exam grades for first and second chance, all answers included. |

*Table 3: All data columns from the master file*

### 3.1.1 Handling missing values

The dataset posed several challenges. Many features contained null values or duplicates, while other features required modifications.

To handle the null values, they were replaced with 1 instead of being removed entirely. This decision was made to ensure compatibility with the chosen algorithm, as many machine learning models cannot process missing values directly. Replacing nulls with 1 provided a neutral, non-disruptive placeholder that allowed the algorithm to operate without errors, while still preserving the overall structure and integrity of the dataset. Removing rows or columns with null values entirely would have resulted in significant data loss, which could negatively impact the model's ability to learn effectively.

The dataset also included a significant number of duplicate entries for students, which posed a challenge for accurate analysis. Multiple rows existed for the same student. To address this issue, a filtering process was implemented to ensure that each student appeared only once in the dataset. The duplicate entries were identified based on unique identifiers such as the Student ID or Full Name, and the most relevant record for each student was retained.

Additionally, students who belonged to the second year were excluded from the final dataset to focus on the relevant cohort. This is done because second-year students could have more experience. By limiting the dataset to first-year students, we aimed to maintain a more homogeneous group to better the prediction model.

### 3.1.2 Feature engineering

The dataset contained a significant amount of data deemed unnecessary or irrelevant for the analysis. For instance, each student had detailed exam answers and was assigned to a specific group or “class” (e.g., A, B, C). This information was used in the dataset to calculate the percentage of success or failure for exams and homework assignments, which were stored in a separate Excel sheet. Additionally, data on homework completion recorded whether assignments were done solo or in a team, and tracked weekly homework completion over five weeks, represented as numerical values (0, 1, or 2). This data was excluded from the cleaned dataset, as it was not deemed relevant for the analysis.

Another challenge arose with the "Previous School" feature. In its original form, this feature listed the names of the schools that students had previously attended. However, these names provided little value for analysis or predictive modeling. To enhance the dataset’s predictive power, the school names were mapped to corresponding school levels (e.g., MBO, HAVO, WO). This transformation standardized the data and made it more useful for machine learning models by categorizing it into a consistent format.

Additionally, the "Dropout" feature required preprocessing to be usable. This feature originally listed the dropout date for each student, leaving the field empty for students who did not drop out. In its raw format, this data was not suitable for analysis. To address this, the dropout information was transformed into a boolean format: 1 for students who dropped out and 0 for those who did not. This transformation simplified the data and enabled it to be used effectively in predictive modeling.

The preprocessed features are combined into one master excel file, this file contains all the useful and cleaned data. The independent variables used in the final model, where the dropout status serves as the dependent variable, are as follows:

| Feature name | Description |
|---|---|
| ANL1 grade | The grade of a student during period ANL1. |
| ANL2 grade | The grade of a student during period ANL2. |
| ANL3 grade | The grade of a student during period ANL3. |
| ANL4 grade | The grade of a student during period ANL4. |
| Previous education | The previous education level (MBO, HBO, WO). |

*Table 4: Final model features*

## 3.2 Correlation testing and Hypothesis testing

### 3.2.1 Linear relationship

Before starting correlation testing, linearity needs to be proven within the data. Predicting the student dropout is a classification problem, this means that Pearson's correlation coefficient can't be used to prove linearity. By visualizing the data into graphs, a linear relationship can be observed. The graphs contain the grades grouped into the following 4 categories.

- 0 < 3 = failed miserably
- 3.0 < 5.5 = failed
- 5.5 < 7.5 = passed
- 7.5 ≤ 10.0 = passed greatly

On the Y-axis, the percentage of students who dropped out is displayed, while the X-axis represents the four grade categories. A clear linear relationship can be observed: as grades decrease, the likelihood of a student dropping out increases.

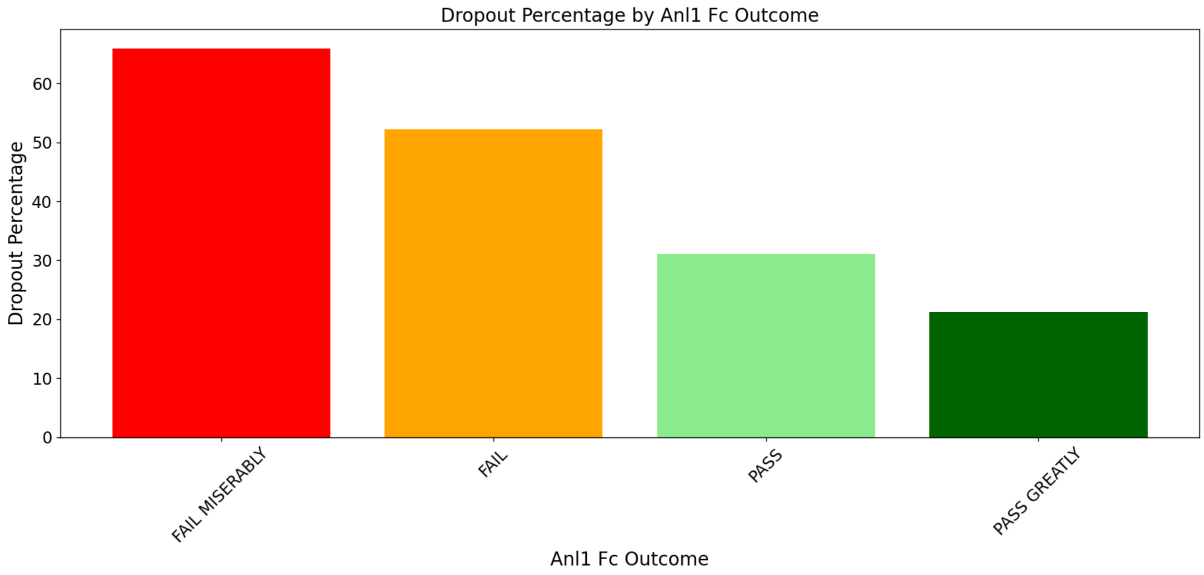


*Figure 1. Analysis 1 first chance barchart.*

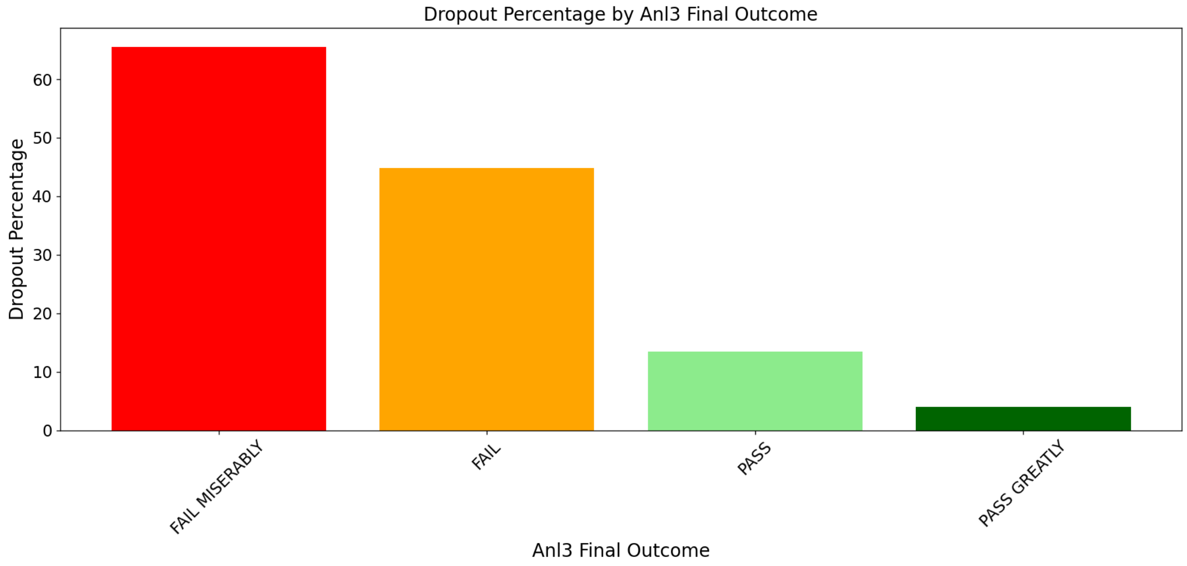


*Figure 2. Analysis 3 final grade barchart.*

In this next graph a feature that does not have a linear relationship has been visualized.

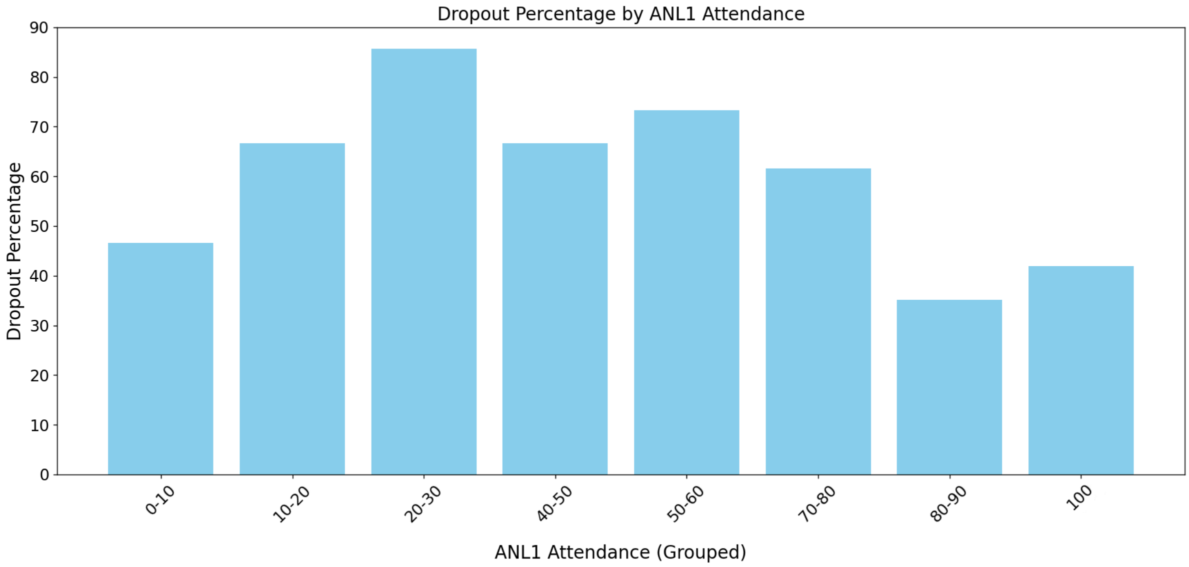


*Figure 3. Analysis 1 attendance rate barchart.*

The attendance percentage has been visualized in steps of 10%, with students who achieved a 100% attendance rate grouped together. The Y-axis shows the percentage of students who dropped out within each attendance category. The visualization indicates that the relationship between attendance percentage and dropout rates is non-linear. While this suggests that simple linear models may not fully leverage this feature, it could still be useful in the final model if machine learning algorithms capable of capturing non-linear relationships are employed.

In the project log odds were used to evaluate the linearity between students' Analysis 1 Final Grade and the likelihood of dropping out (binary variable: yes or no). By binning the grades into intervals, the average grade and dropout probability were calculated for each bin. The dropout probabilities were then transformed into log odds, which allows it to visualize the relationship between the average grade and the log odds of dropping out. This approach makes it possible to verify that the relationship between the grades and the log odds of dropping out was approximately linear, supporting the assumption of linearity in the analysis. The results were plotted into this graph.

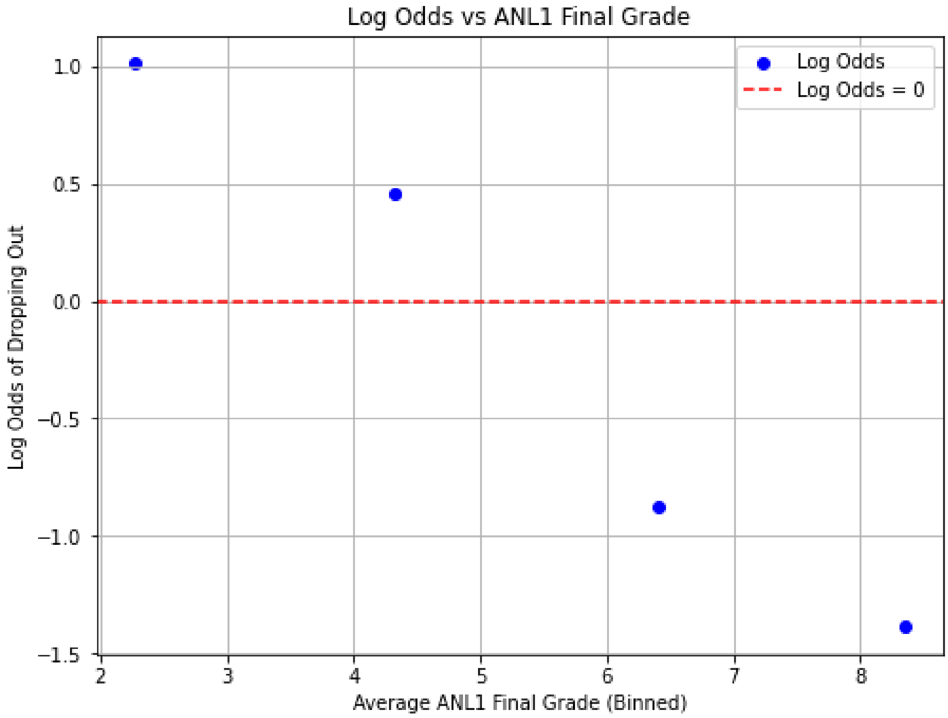


*Figure 4. Log odds average analysis 1 final grades*

### 3.2.2 Correlation testing results

To assess feature correlations, Spearman correlations were calculated and visualized in a heatmap.

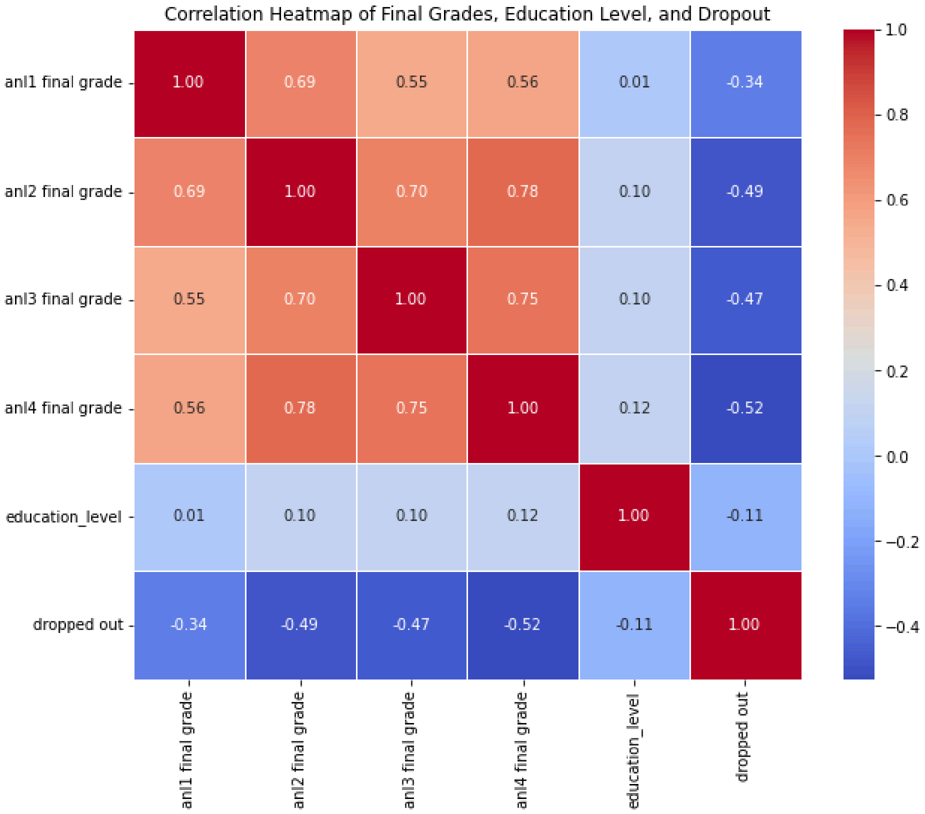


*Figure 5. heatmap of feature correlation.*

The analysis focuses on the correlation of features with the "dropped out" variable, as it serves as the dependent variable in the model. The following correlations were observed:

- Analysis 1 Final Grade: Weak to moderate correlation (−0.34)
- Analysis 2 Final Grade: Moderate correlation (−0.49)
- Analysis 3 Final Grade: Moderate correlation (−0.47)
- Analysis 4 Final Grade: Moderate to strong correlation (−0.52)
- Education Level: Very weak to no correlation (−0.11)

Features with stronger correlations will be prioritized for inclusion in the final model.

#### 3.2.3 Hypothesis testing results

To further validate the relationship between features and the dependent variable (student dropout), hypothesis testing was conducted. Point biserial correlation, a statistical measure for assessing relationships between a binary variable (e.g., dropout: yes or no) and a continuous variable (e.g., grades or attendance), was employed. The hypotheses for the grades were as follows:
Null Hypothesis ($H_0$): There is no significant correlation between the grades and student dropout. (r=0)
Alternative Hypothesis ($H_1$): There is a significant correlation between the grades and student dropout. (r≠0)
For "Analysis 1," "Analysis 2," "Analysis 3," and "Analysis 4" final grades, p-values were below 0.05, and the confidence intervals did not include 0. Thus, it can be concluded with 95% confidence that grades have a statistically significant correlation with student dropout (LibGuides: Statistics Resources: Point Biserial, n.d.).

### 3.3 Machine learning algorithm

Through the literature review it was observed that the most commonly used techniques were logistic regression, random forest and decision trees. By using these algorithms, researchers were able to incorporate diverse variables such as academic achievement, and psychographic factors in their testing, similar to the approach used in this research. The advantages and disadvantages of employing these algorithms on the abovementioned dataset will be discussed in the following paragraphs.

### 3.3.1 Logistic regression

Logistic regression is a statistical method used to model the probability of a binary outcome based on one or more predictor variables. Hence, by using this algorithm the outcome can be classified as dropped out or not dropped out which helps answering the main question. Categorical and continuous variables will be used in the testing, the categorical variables will be encoded in numerical format as logistic regression is an algorithm that can accommodate these variables. Moreover, it is a straightforward algorithm that shows results that are easy to interpret, which is convenient for beginners (Alva, 2021). In the correlation testing it was observed that there was a weak or decent correlation between the independent variables and the dependent variable. Fortunately, logistic regression can perform well even when there is not a strong linear relationship between the variables. Although the model can perform well, it assumes a linear relationship between the independent variables and the dependent factor. This could result in a poor model fit and incorrect predictions. Moreover, it complicates the observation of complex relationships. Another major limitation of logistic regression is that the model can overfit due to a higher number of features in regards to the number of observations (Thanda, 2024). Hence, other algorithms that do not rely on linearity were explored.

### 3.3.2 Random forest

One of the explored algorithms, which did not rely on linearity, is Random Forest. To utilize Random Forest, three main parameters must first be specified before training. These are the node size, the number of trees, and the number of features sampled (Probst, 2019). The random forest algorithm constructs the previously set amount of trees during training and combines their results. Random Forest is an appealing algorithm because it naturally handles both regression and classification, is relatively fast to train and predict and its reduced likelihood of overfitting to the training data (Cutler, 2012). At first, the use of Random Forest could lead to the use of an extreme amount of trees to get the best predictive accuracy. However, the use of an extreme amount of trees leads to a higher computational time and extreme memory usage. Random Forest, when compared to Linear Regression (LR) has several key advantages. Unlike LR, Random Forest is able to identify and capture non-linear patterns effectively. Additionally it also has the capability of handling potentially incomplete datasets more effectively.

For the previously mentioned dataset, Random Forest was implemented using the scikit-learn library. The algorithm incorporated various features such as (retake) grades, attendance rates ect. On top of that, it also underwent hyperparameter tuning *(4.2 Hyperparameter tuning).*

### 3.2.3 Decision trees

Besides Linear Regression and Random Forest, a third algorithm was used. This algorithm is Decision Trees. A decision tree consists of nodes which represent decision rules, for example “student passed course x”. Based on the outcome of this rule, branches are created, eventually arriving at a predicted outcome. Due to a decision tree's versatility of being able to be applied for both classification and regression tasks and its capability of capturing non-linear relationships between features and target variables, making it a very appealing algorithm (Rokach, 2008). However, decision trees are known to easily overfit training data, especially in cases with many nodes. Because of this weakness and the algorithm’s nature, its precision and accuracy are lower compared to methods like Random Forest, as can be seen in the results (Chapter *4 Results / testing and evaluation)*.

## 3.4 Tools

Python is the main programming language used during the development of the student dropout predictor tool. Python is a high-level programming language known for its simplicity and readability (*Welcome to Python.org*, 2025). It's widely used in data science and machine learning due to its vast ecosystem of libraries and  ease of use for data manipulation and analysis (Raschka et al., 2020). This makes python a highly applicable programming language for the student dropout predictor tool. Furthermore, the high accessibility of Python makes it a useful tool for beginning data science students.

One of the python libraries used is Scikit-learn. Scikit-learn is a Python library for machine learning, offering simple and efficient tools. It excels in ease of use, comprehensive documentation and integration with other Python libraries like NumPy and pandas (*Scikit-learn: Machine Learning in Python — Scikit-learn 1.6.1 Documentation*, n.d.). The simple tools make Scikit-learn a great tool for developing the student dropout predictor. Moreover, the comprehensive documentation makes Scikit-learn an accessible library for beginning data science students.

The second python library used is Pandas. Pandas is a Python library for data manipulation and analysis, providing tools for working with structured data like tables and time series. It simplifies data cleaning and transformation (*Pandas - Python Data Analysis Library*, n.d.). The ability to clean and transform data easily makes Pandas a useful tool for cleaning the dataset and developing the student dropout predictor tool.

## 3.5 Evaluation metrics

The performance of the classification algorithms was evaluated using the following metrics: accuracy, precision, recall, F1-score, and the area under the curve (AUC) for receiver operating characteristic (ROC) curves. These metrics are derived from the confusion matrix, which summarizes the relationships between actual and predicted classifications for student dropout prediction.

| | Actual drop out | Actual not drop out |
|---|---|---|
| Predicted drop out | True Positive | False Positive |
| Predicted not drop out | False Negative | True Negative |

*Table 5. Confusion Matrix for student drop out rate.*

The confusion matrix provides the foundation for defining key performance metrics:

- Accuracy: Proportion of all correct predictions, both positive and negative.
  $$Accuracy = \frac{correct\ classifications}{total\ classifications} = \frac{TP + TN}{TP + TN + FP + FN}$$
- Precision: Proportion of predicted dropouts that were correct.
  $$Precision = \frac{correctly\ classified\ actual\ positives}{everything\ classified\ as\ positive} = \frac{TP}{TP + FP}$$
- Recall (Sensitivity): Proportion of actual dropouts correctly predicted
  $$Recall = \frac{correctly\ classified\ actual\ positives}{all\ actual\ positives} = \frac{TP}{TP + FN}$$
- F1-Score: Harmonic mean of precision and recall, balancing both metrics.
  $$F1\ score = 2 \bullet \frac{Precision \bullet Recall}{Precision + Recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

(Classification: Accuracy, Recall, Precision, And Related Metrics, n.d.)

The Receiver Operating Characteristic (ROC) curve is a graph that shows the performance of the model across all thresholds (0 to 1). ROC plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various classification thresholds. The ideal model has a

TPR of 1.0 and a FPR of 1.0 at some threshold. The Area Under the Curve (AUC) quantifies the model's ability to distinguish between classes, so the probability that the model given a random negative or positive example will rank the positive higher than the negative. A high AUC indicates that the model performs well across all thresholds, with 1.0 being ideal. AUC is useful to compare the performance of the models, because the dataset is fairly balanced. The model with a larger area under the curve is usually the better model.
(Classification: ROC And AUC, n.d.)

### 3.5.1 Application to the Problem

The selected key performance metrics are: precision and recall for the following reasons. They are easy to understand and the primary goal of the model is to identify potential dropouts rather than correctly classifying non-dropouts. Precision and recall specifically focus on the model's performance for the dropout class, whereas accuracy evaluates performance across both classes equally.

Although the dataset is relatively balanced (53% did not drop out, and 47% did drop out), precision and recall remain critical for understanding how well the model identifies at-risk students. For example:

Precision minimizes unnecessary help by ensuring that students predicted as dropouts are likely to be truly at risk, reducing false positives (students who have not dropped out, but were predicted as dropping out).
Recall ensures that the model captures the majority of at-risk drop out students, minimizing false negatives (students who have dropped out, but were predicted as not dropping out), which could result in missed opportunities for help.

## Chapter 4: Results / testing and evaluation

This section presents the results of the ML models predicting student dropout. The implementation and experiments of the proposed approach were implemented in Python using scikit-learn. The machine learning algorithms, including logistic regression, decision tree, and random forest, were used to predict student dropout. Cross-validation was applied to evaluate the performance of the models.

### 4.1 Cross validation strategy

To evaluate the performance of the machine learning models, we adopted a 5-fold cross-validation strategy.  The choice of 5-fold cross-validation was primarily motivated by the size of the dataset, which consists of 333 rows. Since the dataset is relatively small, using a higher number of folds, such as 10, could result in each fold containing only a few data points, which might lead to less reliable performance estimates.The 5-fold cross-validation offers a balanced approach by ensuring that each fold contains enough data for both training and validation. This allows the model to be tested on different subsets of data multiple times, ensuring a comprehensive evaluation without overloading any single fold with too few samples. The result is a more reliable estimate of the model's performance, as it leverages the available data effectively while reducing the risk of bias or overfitting.

### 4.2 Hyperparameter tuning and threshold

In our study, we employed hyperparameter tuning with GridSearchCV to identify the optimal parameters for our machine learning models, including Logistic Regression (LR), Random Forest (RF), and Decision Tree (DT). Hyperparameter tuning is a crucial step in model optimization, as it helps in enhancing model performance by selecting the best combination of parameters.

GridSearchCV automates the process of testing various combinations of hyperparameters and selecting the one that gives the best performance based on cross-validation. For each of the models, we defined a set of hyperparameters to improve the performance. Table 6 presents the parameters tested, along with the results from the grid search, highlighting the best parameters for each algorithm.

| Algorithm | Parameter | Value |
|---|---|---|
| Logistic regression | C | 0.1 |
| | Penalty | L1 |
| | solver | Liblinear |
| Random forest | n_estimators | 50 |
| | max_depth | 10 |
| | min_sample_ split | 2 |
| | min_sample_l eaf | 2 |
| | max_features | sqrt |
| Decision tree | criterion | entrophy |
| | max_depth | 3 |
| | max_features | log2 |
| | min_sample_l eaf | 2 |
| | min_sample_ split | 2 |

*Table 6. Table of used/ best parameters for each algorithm*

The default threshold of 0.5 was used for each algorithm. The threshold in is the decision boundary that determines the cutoff point for classifying instances as belonging to one class or the other. When using the default threshold of 0.5, a prediction with a probability higher than or equal to 0.5 is classified as the positive class, while a prediction with a probability below 0.5 is classified as the negative class.

The choice of using the default threshold of 0.5 was made because it provides a straightforward and interpretable decision boundary, which is commonly used in practical applications. Additionally, for this study, we valued precision and recall equally, as both metrics are crucial for ensuring the model's ability to correctly identify students at risk of dropping out (recall) while minimizing false positives (precision).

## 4.3 Comparison of prediction performance

The prediction scores, which are precision, recall, and F1 score, were calculated for each machine learning algorithm, with dropout being designated as the positive class. The scores averages were then computed for every model based on cross-validation, and Table 7 presents the results. In Table 7 and figure 6 respectively, it is shown that the Logistic Regression (LR) model demonstrated the strongest performance among the evaluated classifiers, achieving the highest F1 score (75%) and precision (70%), along with a recall of 83%. These results highlight its robust ability to correctly classify both dropout and non-dropout cases. Furthermore, LR achieved the highest Area Under the Curve (AUC) value of 0.79, showcasing its superior discriminative capability.

The Random Forest (RF) model also performed well, with an F1 score of 71%, precision of 69%, and recall of 75%.

The Decision Tree (DT) model, while simpler and computationally less intensive, showed relatively lower performance with an F1 score of 66%, precision of 64%, and recall of 68%.

Overall, Logistic Regression emerged as the most effective classifier in this study, consistently outperforming the Random Forest and Decision Tree models in terms of both precision and recall. The results demonstrate that while ensemble methods like Random Forest offer competitive performance, simpler models like Logistic Regression can excel when the dataset is well-suited to linear classification.

| **Classifier** | **AUC** | **F1 (%)** | **Precision (%)** | **Recall (%)** |
|---|---|---|---|---|
| Logistic regression | 0.79 | 75 | 70 | 83 |
| Random Forest | 0.78 | 71 | 69 | 75 |
| Decision tree | 0.68 | 66 | 64 | 68 |

*Table 7. Table with the metrics for each algorithm using k-fold 5*

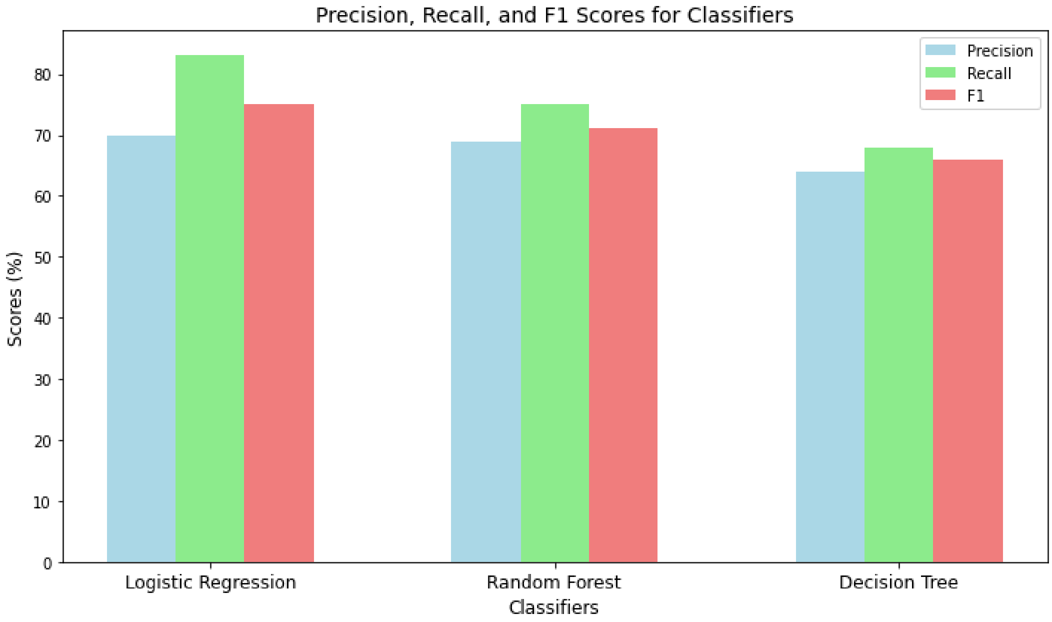


*Figure 6. Barchart of the metrics for each algorithm*

The results of the evaluation indicate that the Logistic Regression model effectively handles the task of predicting dropout. With a recall of 83%, the model demonstrates the ability to correctly identify 83% of dropout cases. Similarly, a precision value of 70% signifies that among the instances predicted as dropouts, 70% are accurate. This balance between precision and recall highlights the model's capability to classify dropout and non-dropout instances with moderate/high reliability.

The ROC curve is a graphical representation of a binary classification model's performance. It illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate (1 - specificity) as the decision threshold varies. The curve helps in understanding how well the model distinguishes between the positive and negative classes across different threshold values. In the provided ROC curve shown in figure 7, the performance of three machine learning models was evaluated: Logistic Regression (LR), Random Forest Classifier (RF), and Decision Tree Classifier (DT).

A curve positioned near the top-left corner of the graph signifies stronger model performance. Upon analyzing the ROC curve, it is evident that Logistic Regression (LR) outperforms the other models, as its curve is closest to this ideal position. This indicates that LR achieves the highest AUC value, which suggests superior classification performance. A higher AUC score reflects better classification ability, as points representing the model’s predictions lie above the diagonal line, indicating that the model performs better than random guessing.

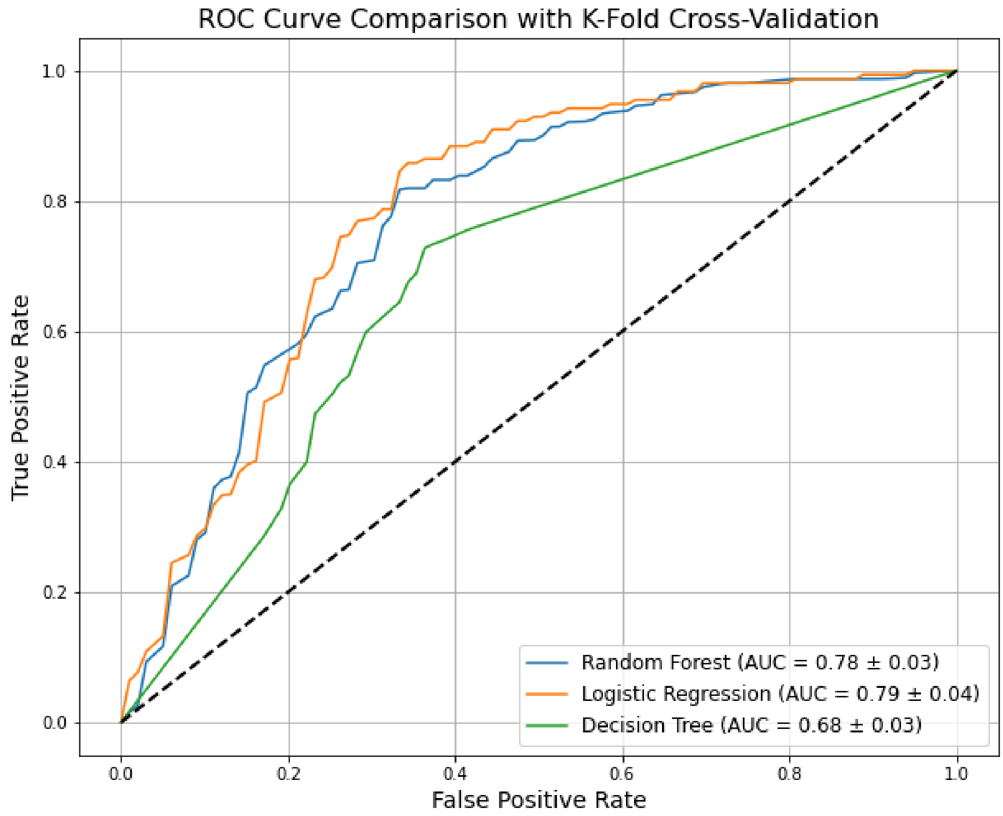


*Figure 7. ROC curves for each algorithm*

The performance of Logistic Regression (LR), Random Forest (RF), and Decision Tree (DT) algorithms was evaluated using their respective confusion matrices. These matrices summarize the classification results in terms of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The results for each model are shown in table 8,9 and 10 and are as follows:

Logistic regression confusion matrix:

| Logistic regression | Predicted No (0) | Predicted Yes (1) |
|---|---|---|
| Actually No (0) | 122 | 56 |
| Actually Yes (1) | 26 | 129 |

*Table 8. Confusion matrix for Logistic Regression*

Random Forest confusion matrix:

| Random Forest | Predicted No (0) | Predicted Yes (1) |
|---|---|---|
| Actually No (0) | 122 | 56 |
| Actually Yes (1) | 26 | 129 |

*Table 9. Confusion matrix for Random Forest*

Decision tree confusion matrix:

| Decision tree | Predicted No (0) | Predicted Yes (1) |
|---|---|---|
| Actually No (0) | 119 | 59 |
| Actually Yes (1) | 50 | 105 |

*Table 10. Confusion matrix for Decision tree*

## 4.4 Feature importance

Multiple machine learning algorithms were used to predict student dropout. Among the algorithms tested, Logistic Regression and Random Forest were the best-performing models in terms of predictive accuracy, recall and precision.To gain insights into the factors influencing the predictions, we analyzed the feature importance.

Figure 8 presents the feature importance derived from the Random Forest model. The analysis identified ANL3 grades as the most influential feature in predicting student outcomes, followed by ANL4 grades. While these features emerged as the most significant predictors, it is important to note that their apparent importance might not fully reflect their true impact. Both ANL3 and ANL4 contained a considerable number of missing values, which were imputed with a default value of 1. This imputation strategy may have artificially increased the importance of these features by introducing a bias in the data.

Despite this limitation, the analysis underscores that grades, in general, are critical factors in predicting whether a student will drop out. While the specific importance of ANL3 and ANL4 grades may be influenced by data imputation, the overall trend highlights the central role of academic performance in determining a student's likelihood of success or dropout.

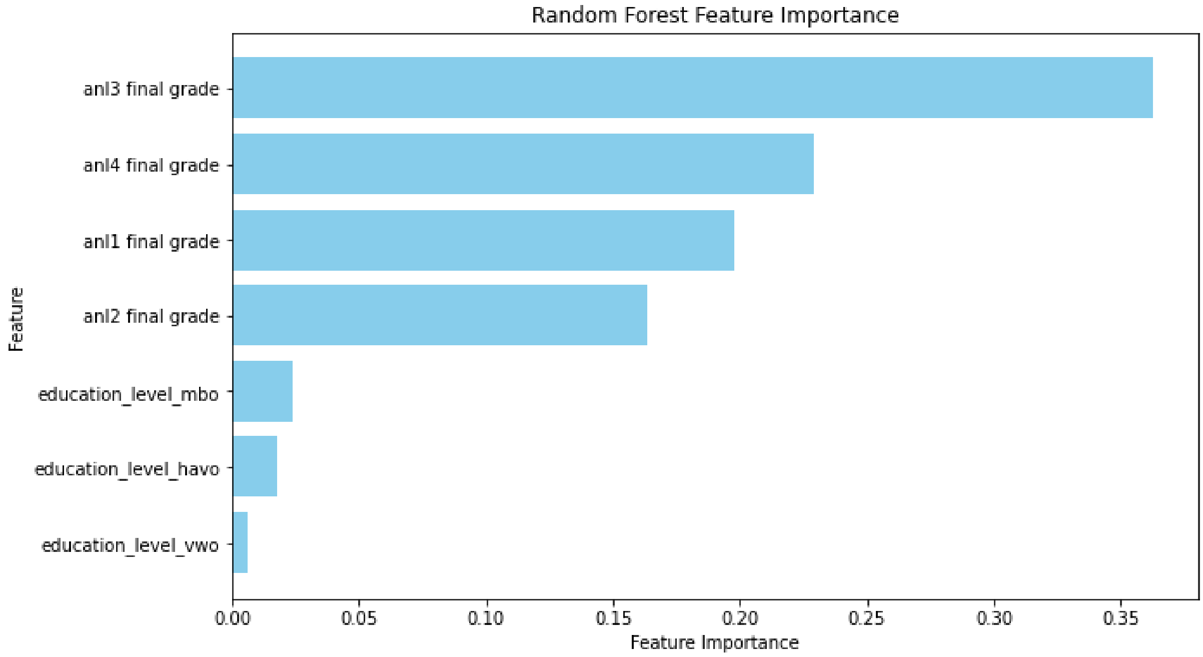


*Figure 8. Feature importance for Random Forest*

### 4.5 Early warning signs to detect dropouts

Identifying early warning signs of student dropouts is important for improving educational outcomes and implementing timely interventions. For this purpose, we analyzed the grades from the first exam students undertook, which is Analysis 1, to assess its potential as a predictor of dropout risk. The rationale for using the ANL1 grade lies in that it is the first exam students take, providing a practical opportunity to detect at-risk students at an initial stage.

Figure 9 shows a graph of ANL1 grades against the number of dropouts per grade. A significant pattern emerges: students who score less than 4.7 are considerably more likely to drop out than students who score above this number. The distribution of students who drop out versus those who do not appears to approximate a normal distribution, though it is not perfectly symmetrical.

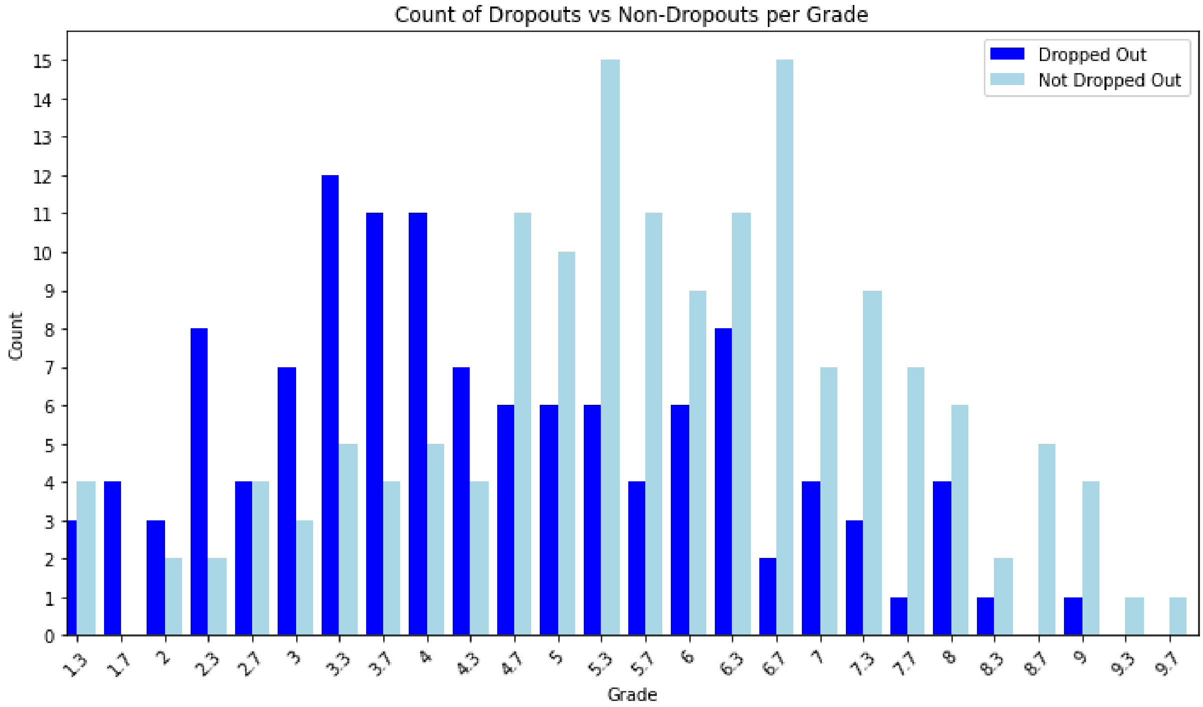


*Figure 9. Barchart of the amount of dropouts per grade for ANL1*

The analysis of ANL1 grades highlights its potential as an early warning indicator for student dropouts. The observed pattern, where students scoring below 4.7 are significantly more likely to leave their studies, provides a valuable threshold for identifying at-risk individuals. These findings show how early grades can be used to detect potential dropouts. By providing the right help early on, universities can lower dropout rates and give more students the chance to succeed.

# Chapter 5: Conclusion

Through the use of data mining tools and an analysis of academic performance indicators, this study looked at predicting student dropout rates. Three machine learning algorithms; logistic regression, random forest, and decision tree, were chosen after a thorough literature review. To guarantee accuracy and usability, the dataset, which contained attendance and grade metrics, was carefully preprocessed. The considerable correlation between grades and dropout rates was proven using correlation and hypothesis testing, and this provided the basis for developing prediction models.

Logistic regression was the most effective among the three algorithms as it provided a good balance of accuracy, precision, and recall. With the highest F1 score of 75% and an AUC of 0.79, it worked well with the dataset and produced results that were easy to understand. Making it a great choice for the problem.

By delivering a reliable way of identifying students who are at danger of dropping out early, the solution gives schools the opportunity to support these students in a timely and focused way. By focusing on academic factors like grades, the model helps schools reduce dropout rates and improve student success. However, this study faced several limitations that must be considered when interpreting the findings. The dataset was small, which may have affected the results. Missing values in key features like ANL3 and ANL4 grades were replaced with a default value of 1, which could have introduced bias and made those features appear more important than they actually are. Also, the dataset contained a limited number of features, focusing primarily on academic grades. While grades are highly informative, they represent only one aspect of a student's experience. Other factors, such as socio-economic background and mental health, were not included but could significantly influence dropout predictions. Additionally, some features, like attendance, had weaker connections to dropout rates and this made them less useful for the model.

Future research should aim to address these limitations by using larger datasets with more diverse features. Including a broader range of student data could enhance the predictive power of the models, leading to more accurate and reliable dropout predictions.

Despite these challenges, the study shows that machine learning is a useful tool for solving problems like student dropout rates. The model is practical and easy to use, giving schools a helpful way to step in early and support students. Although the academic data in the provided dataset serves as the basis for this research, future studies could include more factors, like mental health or family background. This will make the model more accurate and useful in different school systems.

## Reference list

Alva, J. V. (2021, February). *Beginning Mathematica and Wolfram for data science: applications in data analysis, machine learning, and neural networks*. O'Reilly Online Learning. https://learning.oreilly.com/library/view/beginning-mathematica-and/9781484265949/?sso_link=yes&sso_link_from=hogeschool-rotterdam

*Classification: Accuracy, recall, precision, and related metrics*. (n.d.). Google For Developers. https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall

*Classification: ROC and AUC*. (n.d.). Google For Developers. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. In *Springer eBooks* (pp. 157–175). https://doi.org/10.1007/978-1-4419-9326-7_5

Delogu, M., Lagravinese, R., Paolini, D., & Resce, G. (2023). Predicting dropout from higher education: Evidence from Italy. *Economic Modelling*, *130*, 106583. https://doi.org/10.1016/j.econmod.2023.106583

Hadjar, A., Haas, C., & Gewinner, I. (2022). Refining the Spady–Tinto approach: the roles of individual characteristics and institutional support in students' higher education dropout intentions in Luxembourg. *European Journal of Higher Education*, *13*(4), 409–428. https://doi.org/10.1080/21568235.2022.2056494

Kemper, L., Vorhoff, G., & Wigger, B. U. (2020). Predicting student dropout: A machine learning approach. *European Journal of Higher Education*, *10*(1), 28–47. https://doi.org/10.1080/21568235.2020.1718520

LibGuides: Statistics Resources: Point Biserial. (n.d.). https://resources.nu.edu/statsresources/Pointbiserial

Mduma, N., Kalegele, K., & Machuve, D. (2019). A survey of Machine learning Approaches and Techniques for student dropout Prediction. *Data Science Journal*, *18*. https://doi.org/10.5334/dsj-2019-014

Pandas - Python Data Analysis Library. (n.d.). https://pandas.pydata.org/about/

Pandey, B., & Mishra, S. (2023). Comparative Study of Decision Tree and Random Forest Algorithms for Classification Tasks. In *Proceedings of the International Conference on Computational Intelligence and Data Science* (pp. 1-12). Springer. Retrieved from https://link.springer.com/chapter/10.1007/978-981-99-3315-0_12

Patzina, A., & Wydra-Somaggio, G. (2020). Early Careers of Dropouts from Vocational Training: Signals, Human Capital Formation, and Training Firms. *European Sociological Review*, *36*(5), 741–759. https://doi.org/10.1093/esr/jcaa011

Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(3), e1301. https://doi.org/10.1002/widm.1301

Ramiro, N. (2021, December 26). Predicting subscription churn using PySpark ML - Analytics Vidhya - Medium. *Medium*. https://medium.com/analytics-vidhya/predicting-subscription-churn-using-pyspark-ml-b6e265c8d72f

Raschka, S., Patterson, J., & Nolet, C. (2020). Machine learning in Python: main developments and technology trends in data science, machine learning, and artificial intelligence. Information, 11(4), 193. https://doi.org/10.3390/info11040193

Rokach, L., & Maimon, O. (2008). *Data Mining with Decision Trees: Theory and Applications* (2nd ed.). Springer. Retrieved from https://link.springer.com/chapter/10.1007/978-1-4471-7493-6_9

scikit-learn: machine learning in Python — scikit-learn 1.6.1 documentation. (n.d.). https://scikit-learn.org/stable/index.html

Thanda, A. (2024, December 19). What is Logistic Regression? A Beginner's Guide [2025]. *CareerFoundry*. https://careerfoundry.com/en/blog/data-analytics/what-is-logistic-regression/

Villar, A., & De Andrade, C. R. V. (2024). Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study. *Discover Artificial Intelligence*, *4*(1). https://doi.org/10.1007/s44163-023-00079-z

*Welcome to Python.org*. (2025, January 14). Python.org. https://www.python.org/

Yaacob, W. F. W., Sobri, N. M., Nasir, S. a. M., Yaacob, W. F. W., Norshahidi, N. D., & Husin, W. Z. W. (2020). Predicting Student Drop-Out in higher Institution using data mining techniques. *Journal of Physics Conference Series*, *1496*, 012005. https://doi.org/10.1088/1742-6596/1496/1/012005