Universidade de Évora - Escola de Ciências e Tecnologia

Mestrado em Engenharia Informática

Dissertação

Student Dropout Risk Detection at University of Évora

Sharmin Sultana Prite

Orientador(es) | Teresa Gonçalves

Luís Rato

Évora 2021

A dissertação foi objeto de apreciação e discussão pública pelo seguinte júri nomeado pelo Diretor da Escola de Ciências e Tecnologia:

Presidente	|	Paulo Miguel Quaresma (Universidade de Évora)
Vogais	|	Miguel José Barão (Universidade de Évora) (Arguente) Teresa Gonçalves (Universidade de Évora) (Orientador)

Évora 2021

To my son Aayan Ahmed
My Best Achievement in this Journey

Prefácio

This thesis is the final work of my Master's study at the University of Évora from the Informatics department. It serves as documentation of my research during the study, made from September 2018 until November 2020. It presents a machine learning approach for identifying the risk profile of students of the University of Évora. It includes the design and extraction of considered functional attributes that characterize the academic path of the student and the search for the best model. It specifically predicts the dropout of a student in the early stage using the student's previous academic records.

Evora, October 18, 2021
Sharmin Sultana Prite

Acknowledgements

I would like to express my deepest gratitude to my advisor Dr Teresa Cristina de Freitas Gonçalves for her deep inspiration, direction, care, patience, and for giving me a great environment for research. I would not have been able to finish my thesis without her support.

Many thanks to my other advisor Dr Luís Miguel Mendonça Rato for sharing his expertise, ideas, sincere and valuable direction and encouragement.

I would also like to thank Kashyap Rayani, who has always been interested in helping me with my dissertation writing and giving me his best advice.

Finally, I would like to thank my husband Md.Sajib Ahmed, who supported me and encouraged me with his best wishes.

Contents

List of Figures

List of Tables

Acronyms

HEIs Higher Education Institutions

ECTS European Credit Transfer and Accumulation System

ML Machine Learning

DT Decision Tree

NB Naïve Bayes

RF Random Forest

SVM Support Vector Machines

CSV Comma-Separated Values

DBMS Database Management System

Abstract

Currently, student dropout is a global problem in higher education affecting the results of education systems. In addition to providing state-of-the-art education, any institution needs to maintain its student flow rate, which means that predicting dropout is critical to measuring the success of an education system.

This work focuses on identifying the risk of dropout at the University of Évora based on students' academic performance. We propose a set of academic information as predictive attributes and present machine learning models that have a precision of 96.8% and f1-measure of 94.8% as performance in identifying students at risk of dropping out.

In this regard, 13 years of academic data were collected from four different academic programs (the academic years 2006/2007 to 2018/2019 and Management, Biology, Informatics Engineering and Nursing programs). After collecting the students' academic records, anonymizing the information and pre-processing the data, an engineering and attribute selection process was conducted, building the data sets. Various machine learning algorithms were applied and their performance was compared; models were built with Decision Trees (DT), Naïve Bayes (NB), Support Vector Machines (SVM) and Random Forest (RF), with the latter algorithm having obtained the best performance in terms of recall.

Keywords: Academic Data Mining, Machine Learning, Classifier, Dropout, Random Forest

Sumário

Detecção de Risco de Abandono de Alunos na Universidade de Évora

Atualmente, o abandono escolar é um problema global no ensino superior que afeta os resultados dos sistemas educativos. Além de fornecer educação de ponta, qualquer instituição precisa manter a taxa de fluxo de alunos, o que significa que a previsão do abandono escolar é essencial para medir o sucesso de um sistema de ensino.

Este trabalho centra-se na identificação do risco de abandono escolar na Universidade de Évora com base no desempenho escolar dos alunos. Propomos um conjunto de informação académica como atributos preditivos e apresentamos modelos de aprendizagem automática que apresentam uma precisão de 96.8% e f1-medir de 94.8% como desempenho na identificação de alunos em risco de desistência.

Nesse sentido, foram recolhidos 13 anos de dados académicos de quatro cursos diferentes (anos letivos de 2006/2007 a 2018/2019 e cursos de Gestão, Biologia, Engenharia Informática e Enfermagem). Após a recolha do percurso académico dos alunos, a anonimização da informação e o pré-processamento dos dados, foi conduzido um processo de engenharia e seleção de atributos, construindo assim os conjuntos de dados. Foram aplicados vários algoritmos de aprendizagem automática e o seu desempenho foi comparado; foram construídos modelo com Árvores de Decisão (DT), Naïve Bayes (NB), Máquinas de Vetores de Suporte (SVM) e Random Forest (RF), tendo este último algoritmo obtido o melhor desempenho no que respeita à cobertura.

Palavras chave: Mineração de Dados Académicos, Aprendizagem Automática, Classificação, Abandono Escolar, Floresta Aleatória

Chapter 1

Introduction

We live in an information age where data is easy to obtain and store is inexpensive. Information is the primary ingredient to generate new knowledge that can be applied to solve various real-life problems. One real-life problem that exists in the education system all over the world is the student's dropout before completing their graduation from the educational institute.

Educational data mining is an emerging interdisciplinary research field that deals with developing methods for exploring data generated in educational contexts (Rai, 2014). Data mining techniques to analyze academic data are expected to benefit the Higher Educational Institutions (HEIs) since educational organizations are necessary parts of our society and play an important role in the growth and development of any nation. On the other hand, dropout students are an obstacle to the growth and development of any country.

About 40% of students seeking bachelor's degrees do not complete their degree within six years (Raisman, 2013). Student dropout happens when a student ends his academic life after a particular time without completing the degree, being a major concern in education and policy-making communities (Demetriou and Schmitz-Sciborski, 2011). Besides the universities' revenue loss, society also suffers because the skilled workforce shortage can undermine the nation's productive capacity. Moreover, studies indicate that dropout students are more likely to be frequent recipients of welfare and unemployment subsidies (Catterall, 1987). Because of these negative consequences, student dropout has long been considered a serious educational problem by educators, researchers, and policymakers.

Allen et al. (2008) state that the reasons for student dropout can be related to economical, social, and psychological issues. A study funded by the Bill and Melinda Gates Foundation (Bill and Melinda, 2020) found that

the main reason students drop out is the conflict between academic and work and family commitments (Bridgeland et al., 2006). Total five major factors contributing to a student having difficulty graduating or dropping out are mentioned: academic environment, personal characteristics, family influences, health, and community environment (Fig 1.1). In order to determine whether or not a student is at risk, all of these five factors must be considered.

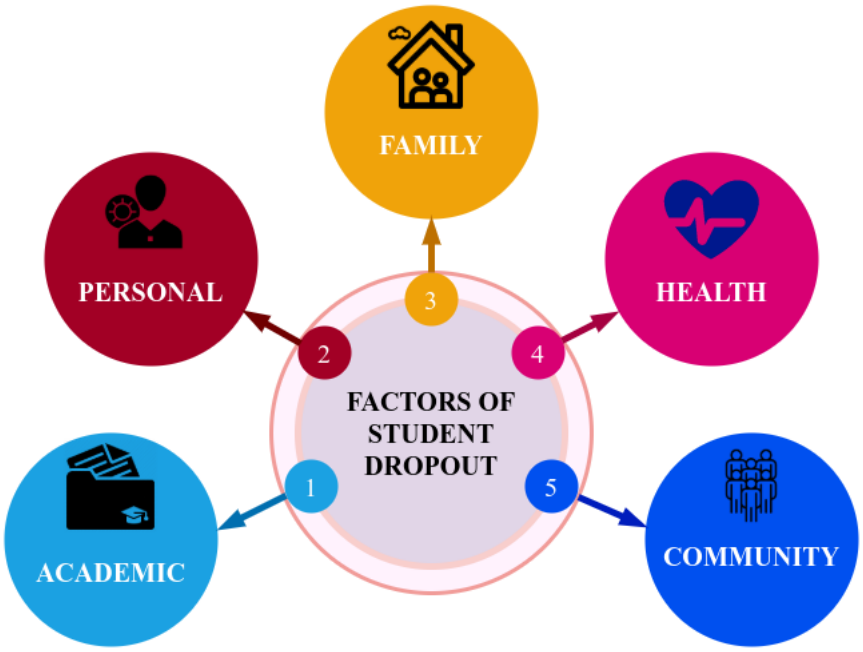


Figure 1.1: Diagram of Student Dropout Factors (Francesca, 2020).

The first factor is the academic issues where involve a student's academic performance, attendance, and unsupportive school culture. These factors affect the dropout rate of a student from an educational institution. The second one is a student's personal life that also depends on a student's mental health, offending records, and his/her other responsibilities. These issues are also responsible for a student dropping out. Family matters such as family income, parental marital status, foster care system, and family criminal records are accountable for a student's dropout. Another factor is student health where includes his/her medical conditions, medical records, and substance use that affect the academic life then a student goes to drop out. The last one is the community factor where involves neighborhood demographics, crime rates, and community resources for a student dropout.

In this work, only students' academic performance records are used to identify the risk student profile in our work since gathering the other factors'

information is time-consuming and expensive. So our proposed machine learning model works on only academic performance. Thus, we want students to do their graduate on time and succeed in life, as well as actively contribute to society and their families.

1.1 Motivation

Nowadays, student dropout in Higher Education Institutions (HEIs) is a crucial concern for educators and managers. Knowing beforehand, the students at risk of dropping out allow higher education players to take measures that can improve the institution's success rate.

Every year a lot of students are admitted to universities, but not all of them complete their degrees. So identifying them at an early stage and providing motivation can be helpful for them not to leave university without completing graduation. This scenario happened with my friends and that's why it was inspiring for me to work with this topic.

1.2 Goals

The primary goal of this work is to design a system that can predict risky profile students. So identifying the risky profile of students is the target of this research. Since we only have information from their academic path and are willing to understand it, this information is enough to detect students at risk, and the research question can be stated as **"Is it possible to identify the profiles of students at risk of dropping out by analyzing their academic records?"**. The main goal is to, with this information, university authorities can take actions to avoid dropout and reduce the university's dropout rate. Reducing the dropout rate will increase the university success rate, and motivated students will succeed in their future careers and lives.

1.3 Proposal

Early prediction of dropout students is a challenging task in Higher Education Institutions (HEIs). Data analysis is one way to scale down the rate of dropout students. In this work, we propose to use student academic records from the University of Évora. The data is collected from the university's database and pre-processed to extract and engineering useful features. Then a machine learning approach is used to build predicting models.

1.4 Main contributions

The main contributions to this work can be stated as:

1. Compilation of state-of-art approaches to detect dropping out students in early stage.
2. A set of engineered discriminant features.
3. A dataset of student academic paths.
4. Machine learning prediction model.

1.5 Dissertation Outline

This dissertation is divided into six chapters:

- The first Chapter is the Introduction and presents the motivation, goals, proposal, and main contributions;
- The second chapter presents the tools and techniques that are used to process and analyze the educational data;
- Similar works on applying different algorithms and approaches for identifying/detecting dropping out students are discussed in the third Chapter;
- Chapter four explains feature engineering including raw data, database build, data pre-processing, instance labelling, dataset construction and characterization.
- The fifth Chapter presents the performance metrics of the experiments, experimental setup, experiments performed and discusses the results obtained.
- Finally, Chapter six presents conclusions and future work.

Chapter 2

Tools and Techniques

Tools and techniques are the primary instruments used to gather, synthesize, and analyze information appropriately and practically. This chapter describes various tools and techniques used in the recent trend to process and analyze educational data. Different types of tools and techniques are introduced in the following sections.

2.1 Database Management System (DBMS)

According to Wikipedia (2020), a database management system is a software that manipulates data with the group of program, organize and analyze the data that interact with the end application. The DBMS is managing the data to interface to the end-user and storing and retrieving the data with considering the appropriate security measure. DBMS support to recover the damaged data and enforcement of constraints to ensure the data follows certain rules.

A database management system (DBMS) is defined by Connolly and Begg (2005) as a "software system that allows users to define, create, maintain, and control database access". A database management system's capabilities might vary greatly; the storing, retrieval, and updating of data are the key functions. According to Edgar F. Codd[1] proposal, a fully-fledged general-purpose DBMS provides the following functions and services (Connolly and Begg, 2005):

- Data storage, recovery, and modification are all possible.

[1] Edgar Frank "Ted" Codd (19 Aug 1923 – 18 Apr 2003) was an English computer scientist who, while working for IBM, invented the relational model for database management.

- The metadata is described in a user-accessible catalogue or data dictionary.
- Transactions and concurrency are supported.
- If the database is damaged, there are options for retrieving it.
- Support for access authorization and data updates.
- Access support from remote places.
- Constraints are used to verify that data in the database follows particular rules.

2.1.1 DBMS Tools

In commercially there are many different data management tools are used. For examples of the most popular DBMS are PostgreSQL (Drake and Worsley, 2002), MySQL (Letkowski, 2015), Microsoft SQL Server (Mistry and Misner, 2014), Oracle Database (Greenwald et al., 2013), and Microsoft Access (Mikheeva, 2006). For this work here used Microsoft SQL Server, version 14.

Microsoft SQL Server is a relational DBMS that was developed by Microsoft[2] Corporation. It is a software product that saves and retrieves data as required by other software applications, which may run on the same computer or on a different computer. In this work, it has been used for the data store, cleaning, filtering, and processing. SQL Server Management Studio (SSMS) is used as a tool to access the MSSQL server since it does not do anything by itself. On the other hand, the MSSQL server import and export wizard tool is used for data storing from excel files to MSSQL servers, and SQL query and stored procedures have been used for data clearing, filtering, and processing. MSSQL server is an overall great tool for generating and managing a strong relational database. It's made easy to all for data storing, cleaning, filtering, and processing.

2.2 Supervised Machine Learning

Machine learning (ML) is the study of computer algorithms that enhance automatically through knowledge (Mitchell, 1997). Supervised learning algorithms create a mathematical model of data aggregation that holds both

[2] www.microsoft.com

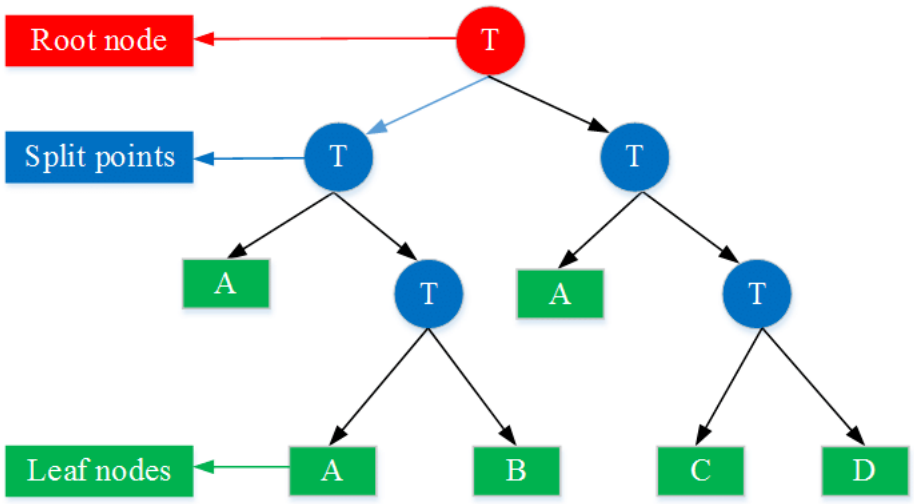
input and desired results (Russell and Norvig, 2002). It is seen as a subset of artificial intelligence.

The discipline of ML employs several approaches to teach computers to accomplish tasks where no entirely satisfactory algorithms are available. Various algorithms have been used and researched for the supervised systems: The following subsection introduces four of them: Decision Tree (DT), Naïve Bayes (NB), Random Forest (RF), and Support Vector Machines (SVM).

2.2.1 Decision Tree (DT)

In general, a decision tree is an inductive machine learning technique for data mining (Quinlan, 1996). Mainly two types decision tree are applied to solve data mining (Wikipedia, 2021) problems: **Classification Tree** and **Regression Tree**.

Briefly, a decision tree (Rokach and Maimon, 2008) is a classification that is expressed as a repetitive division of example space. In the beginning, the decision tree creates the root note that constitutes make the leaf node. Each node has the out node and the incoming node except the root node and divides into two or more subtrees that depend on the incoming node.The term "internal" or "test" refers to a node with outgoing edges. Figure 2.1 shows a general form of a decision tree.



Source: https://iopscience.iop.org/article/10.1088/1757-899X/394/5/052002/pdf

Figure 2.1: Decision Tree (DT) (Yuan et al., 2018).

2.2.2 Naïve Bayes (NB)

A Naïve Bayes (NB) classifier is a generic probabilistic classifier that is based on the Bayes theorem and assumes that each feature is class-conditionally independent (Christopher and alias Balamurugan, 2014). It is a probabilities algorithm that helps to predict immediately. The following equation 2.1 expresses the Bayes theorem mathematically (Ugalde, 2015):

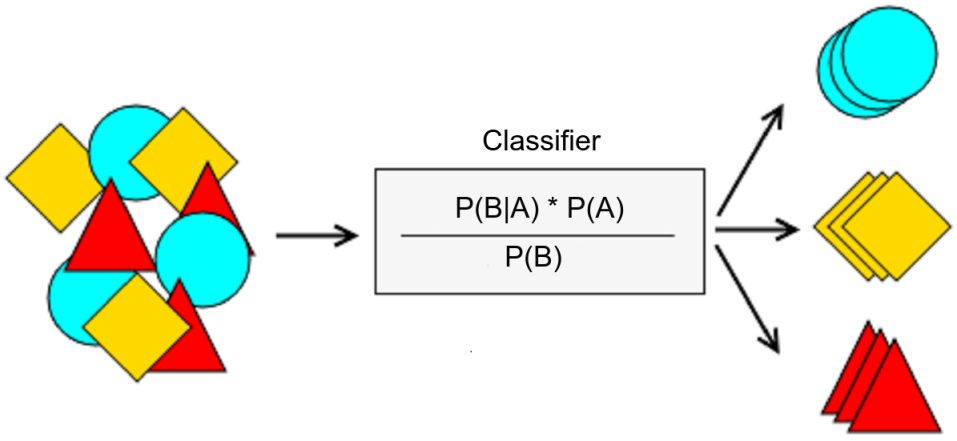
$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \tag{2.1}$$

where A and B are two distinct events,

- The probabilities of A and B without considering P (A) and P (B).
- P(A|B) is the probability of A given that B is true.
- P(B|A) is the probability of B given that A is true.

In Naïve Bayes Learning, each example is described by a set of attributes and takes a class value from a predetermined set of values. The influence of a variable value on a particular class is independent of the values of other variables when a feature is believed to be class-conditionally independent.

And the Naïve Bayes classifier solves real-world problems with document classification and filtering. For that, there require a training set of data to estimate the necessary parameters. Figure 2.2 shows a basic form of a Naïve Bayes algorithm.

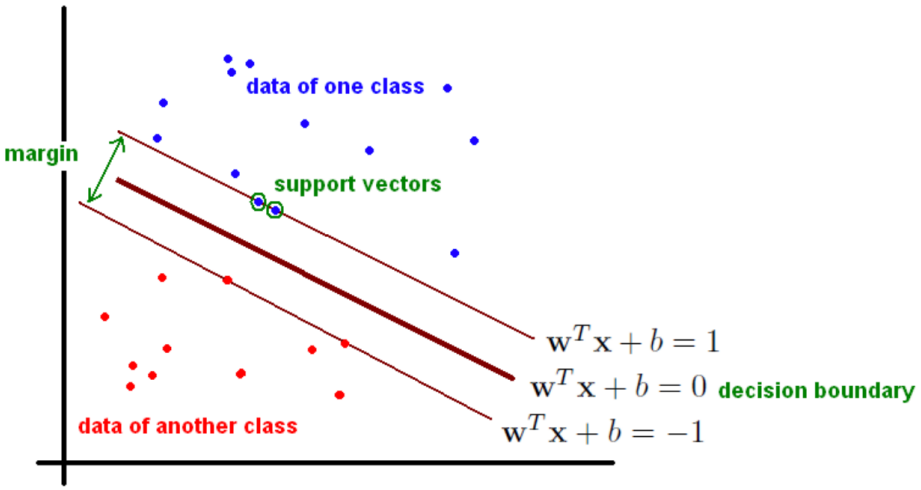


Source: https://bit.ly/3BPw8xc

Figure 2.2: Naïve Bayes (NB).

2.2.3 Support Vector Machines (SVM)

Nowadays, Support Vector Machine (SVM) most of used machine learning algorithms. It classifies the data by constructing a hyperplane (HP) on the high dimensional feature space (Obaidullah et al., 2018), with the hyperplane dividing a dataset into two classes, as shown in Figure 2.3.



Source: http://scjournal.ius.edu.ba/index.php/scjournal/article/view/107/108

Figure 2.3: Support Vector Machines (SVM) (Akcesme and Can, 2016)

The data points closest to the hyperplane, the points in the data set that, if deleted, would change the position of the dividing hyperplane, are called support vectors. As a result, they might be considered essential components of a data set. The margin is the distance between the hyperplane and the nearest data point from either class. The aim is to select a hyperplane with the largest possible margin between the hyperplane and any point in the training set, giving a greater chance of new data being classified correctly. Some features (Wikipédia, 2021) of the SVM are:

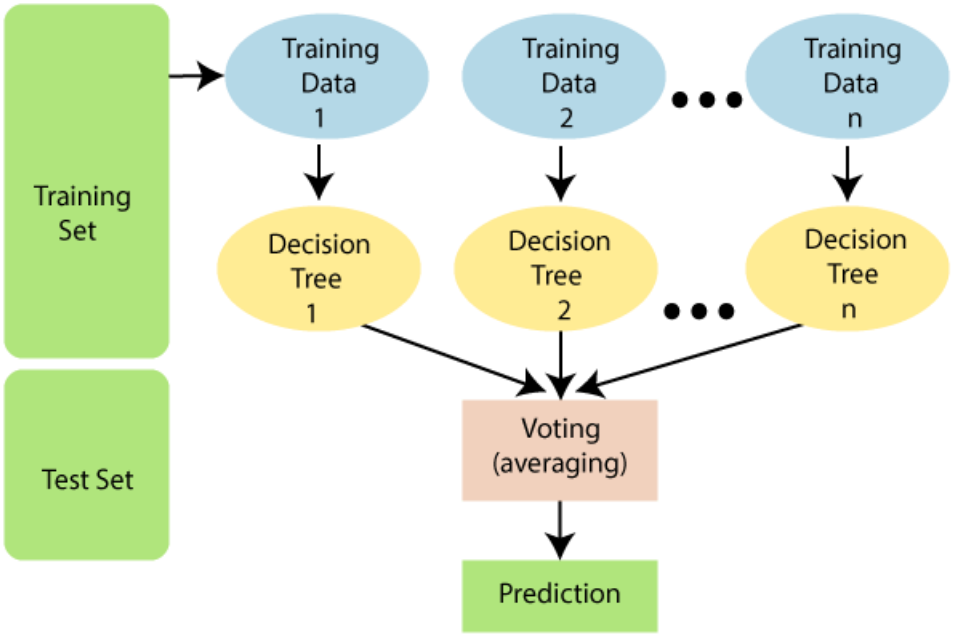
- In the case of an outlier, SVM seeks the best possible form of classification and, if necessary, disregards the outlier;
- It works very well in complicated domains, where there is a clear margin of separation;
- It doesn't work well on very large data sets, as it requires matrix inversion - increasing computational complexity with up to the cube of data volume;
- It doesn't work well on data sets with a lot of noise;

- If the classes are very overlapping, only independent evidence should be used (because it is not very good with data with a lot of noise);

2.2.4 Random Forest (RF)

Random forest is a flexible and user-friendly machine learning method that produces excellent results without hyper-parameter tuning. It is one of the most widely used algorithms, due to its simplicity and diversity (Islam and Amin, 2020).

Random forest is a decision tree-based approach for predicting outcomes and analyzing behavior (Belmokre et al., 2019). It comprises a large number of decision trees representing a unique instance of the classification of data input into the random forest. The random forest method analyzes each instance separately, selecting the one with the majority votes as the selected prediction. Figure 2.4 shows a general form of a random forest algorithm.



Source: https://www.javatpoint.com/machine-learning-random-forest-algorithm

Figure 2.4: Random Forest (RF) (Dey et al., 2020)

Each classification tree uses samples from the initial data set as input. At each node, features are chosen at random and utilized to build the tree. Every tree in the forest should not be pruned until the exercise is completed when the prediction is reached decisively. In this approach, the random forest allows any weakly correlated classifier to become a powerful classifier.

The random forest algorithm was created by Leo Breiman and Adele Cutler in 2001 (Breiman, 2001). It generally exhibits a significant performance improvement compared to single tree classifiers such as C4.5. It produces a lower generalization error rate than Adaboost, but it is more robust to noise. The advantages of Random Forest are (Horning, 2013):

- Overcome the problem of over fitting;
- Outlier data is less sensitive;
- Parameters can be set easily eliminating the need for pruning the trees;
- Automatically produced importance and accuracy variables.

2.2.5 Machine Learning Tool: Weka

Waikato Environment for Knowledge Analysis (Weka) is free software released under the GNU General Public License and is the companion program to the book "Data Mining: Practical Machine Learning Tools and Techniques", which was created at the University of Waikato in New Zealand (Azuaje, 2006). It is a set of data mining-related machine learning techniques. The algorithms can be applied to a dataset directly or called from the user's Java code. Data pre-processing, classification, regression, clustering, association rules, and visualization tools are included in Weka (en.Wikipedia, 2021).

Weka's first non-Java version featured a Tcl[3]/Tk frontend for (mainly third-party) modeling algorithms written in other programming languages, as well as data preparation utilities written in C[4] and a Makefile[5]-based method for executing machine learning experiments. This version was developed with the purpose of analyzing data from agricultural domains (Holmes et al., 1994; Garner et al., 1995). The more current completely Java[6]-based version (Weka 3), which was first developed in 1997, is now utilized in a wide range of applications, including education and research. Weka has a number of advantages, including:

- It is written in Java programming language, thus running on almost any modern computing platform.
- It's a comprehensive repository of data preparation and modelling methods.

[3]https://en.wikipedia.org/wiki/Tcl
[4]https://en.wikipedia.org/wiki/C_(programming_language)
[5]https://en.wikipedia.org/wiki/Makefile
[6]https://en.wikipedia.org/wiki/Java_(programming_language)

- The graphical user interface makes it easy to operate.
- The GNU General Public License[7] allows for free distribution.

Weka uses Java Database Connectivity to connect to SQL databases and can process the results of a database query.

[7] https://en.wikipedia.org/wiki/GNU_General_Public_License

Chapter 3

Related Work

Student dropout has been defined as a student leaving education without obtaining a minimal credential (De Witte et al., 2013). So student dropout prediction is a widely researched area. Similarly, student dropout in Higher Education Institutions (HEIs) is becoming a key concern for educators and a central focus for researchers. Researchers have found that dropout happens due to several reasons related to numerous factors.

This chapter introduces the concept of student dropout and presents related work with studies from different countries' universities and schools, that focus on various different factors for student dropout.

3.1 Student Dropout

This section discusses about the dropout situation and the reasons students dropout in higher educational institutes. A student can dropout because of various reasons such as unsupportive academic culture, attendance record, result, family responsibilities, mental health, household income, parental marital status, medical conditions etc.

To minimize the student dropout rates, the "No Child Left Behind Act" (2001), and the "Europe 2020" goals have been drawn up in the United States and Europe (De Witte et al., 2013), respectively. It was targeted that 90% of students on average finish their high school graduation; later it aimed in the European Union that at least 85% of all 22-year-old complete higher-secondary education.

However, this dropout matter was getting a big part of policymakers' attention, since school dropout is still critical issue. The increasing study about

the early school leaving gives direction about school dropouts, compared with the graduated peers, those are associated with long-term poverty, unemployment, sustained dependence on public assistance, bleak health prospects, single parenthood (in females), political and social apathy, and (juvenile) crime (Headley, 2003).

The high dropout rates in Western countries sharply contrast with government officials and policymakers' social and economic objectives to achieve sustainable economic growth (De Witte et al., 2013). Studies show that about 40% of the university student in completing their higher education by 2020. Similarly, it is also noticed that around one-fourth of the students dropout in the first year before their graduation. In Europe, Portugal has the fourth highest rate for early school leaving (Andrei et al., 2011), and 14% of people who have not to finish their academic year or quit to attend. Now, the most challenging fact is to improve the student drop out rate in the Higher Education Institutions (HEIs).

Predict student dropout in the early stage, is an essential work in the HEIs (Rai, 2014): decreasing the number of dropout students and increasing enrollment is very important for a University. So data analysis is a way to improve this situation by getting out the in-depth information. The knowledge from the data analysis helps in predicting dropout students at an early stage. This information will help management keep an eye on weaker students to improve their status and assist them in different ways where their needs are. These could be an effective way of preventing student dropout.

3.2 Related Work

Significant research has been done in the student dropout field, with several techniques being applied to reduce the number of dropping out students. The following subsection presents the dropout related work that already done, so their proposed methods, technique and outcome will describe in the below sections. These dropout related work descriptions divided into two groups based on the high school level and university level.

3.2.1 High School Level Dropout

Lee and Chung (2019) show that it is possible to improve a dropout early warning system's performance by (a) addressing the class imbalance issue using the synthetic minority oversampling techniques (SMOTE) and the ensemble methods in machine learning, and (b) evaluating the trained classifiers with both receivers operating characteristic (ROC) and precision-recall (PR).

In their work, a large dataset of the 165,715 high school students is used from the National Education Information System (NEIS) of South Korea, which boosted decision tree (BDT) and random forest (with and without SMOTE). The trained classifiers with both ROC and PR curves were used to evaluate their system, with ROC curves being less informative than PRăcurves. And they get from the ROC and PR curve analysis that the optimal performance was achieved through a boosted decision tree.

Similarly, Mduma et al. (2019) proposed a predictive ensemble model based on prototyping to identify dropout students in secondary schools. This deployed model is developed by soft combining a tuned Logistic Regression and Multi-Layer Perception models (Manar and Ploix, 2015). Their developed ensemble predictive model used the Uwezo data[1] collected in 2015 at the country level across hundreds of thousands of households in East Africa; The dataset consists of 61,340 samples of student records and 18 features. They applied feature engineering to obtain the most important features for predicting student dropout, so their system could recognize earlier which students and schools need help.

In another study, Fall and Roberts (2012) analyzed educational data and found that appreciation of social context (parent and teacher support) predicts students self-appreciation, which, in turn, helps predict students academic progress and behavior activity. It identifies that a student's academic and behavioral engagement and achievement in the 10th grade are related to a decreased possibility of dropping out in the 12th grade.

3.2.2 University Level Dropout

Berens et al. (2018) developed an Early Detection System (EDS) using administrative student data from state and private universities to predict student success as a basis for targeted intervention. The EDS used neural networks, decision trees, regression analysis, and AdaBoost to identify student characteristics that differentiate graduates and potential dropouts.

Kim and Kim (2018) examined the possible causes of university dropout. There are four fundamental categories: students, resources, faculty, and university characteristics. A non-linear panel data model was constructed from data from 2013 to 2015. They concluded that the important factors for students' dropout are the faculty's qualitative and quantitative features, financial resources, and university status.

[1] http://www.twaweza.org/go/uwezo-datasets

Chen et al. (2018) presented a survival analysis[2] framework for the early identification of students at the risk of dropping out. They compared the performance of survival analysis approaches to other machine learning approaches including logistic regression, decision trees, and boosting. Their proposed method showed good performance to predict student at-risk in an early stage and was also able to indicate when a student will dropout with high accuracy.

CEME (2017) evaluate the reason for student failure based on the previous data and predict the risk of failure for the next course. They used a dataset consisting of 450 records extracted from five College of Electrical & Mechanical Engineering (CEME)[3]. They applied six ML algorithms for prediction and risk analysis; ID3 gave the best results compared to others based on accuracy: C4.5 with 55.52%, Random Tree with 54.11%, Random Forest with 61.97%, ID3 with 79.23%, CHAID with 49.50% and Decision Stump with 50.95%. The ID3 algorithm reduced the following rule: students who have CGPA $\leq$ 2.2 are in risk of dropout. Their system helps estimate the risk in the early phase which can help teachers design effective planning for the students who are at risk.

Ameri et al. (2016) developed a survival analysis framework that can identify at-risk students using time-dependent Cox (TD-Cox) model, which captures time-varying factors. The framework assists in providing more accurate identification of dropout students with a specific time to dropout. This work was evaluated with real student data at Wayne State University. Their proposed Cox-based framework, predict the student dropout and semester of dropout with high accuracy and precision.

Abu-Oda and El-Halees (2015) used different data mining techniques to examine and predict the dropout students in various programs of AL-AQSA University[4]. To do the analysis, they took 1290 records of the computer science Graduated program from 2005 to 2011. This data records the student's study history: a transcript with the first two years' records, and the high school GPA and average marks. The problem is modelled as a binary classification showing if a particular student finished the study with the first chosen major or not. They trained different classifiers, including Naive Bayes (NB) and Decision Tree (DT), for predicting dropout students. They used a 10-fold cross-validation method and the obtained DT and NB classifiers accuracies were 98.14% and 96.86%, respectively. The work also tried to

[2]Survival analysis is a set of statistical methods for longitudinal data analysis on the occurrence of events.

[3]CEME is a constituent college of the National University of Sciences and Technology, Pakistan

[4]AL-AQSA University is a Palestinian university established in 1955 in the Gaza Strip, Palestine.

discover the hidden relationship between enrolment tenacity and student dropout level using frequent-pattern growth algorithm. They found that a student fails in "Algorithm Analysis", and "Programming Language" courses will be dropping out; Similarly, a student with 80% marks in "Algorithm Analysis" will not drop out.

Bhardwaj and Pal (2012) gathered 300 student records from different degree colleges and institutions in India and preprocessed and transformed the data. They used a Bayesian classification prediction model to analyze the performance of students. The objective of their research was to discover information from student performance in previous year examinations. This work had significant importance to identify students in an earlier stage who needed special attention from teachers to improve their performance and reduce the possibility of dropout.

Chen (2012) proposed and tested a multilevel event history model that identifies the primary institutional attributes related to student dropout risk in a longitudinal process. Her evidence indicates that institutional expenditure on student services is negatively associated with student dropout behaviour.

A study to identify factors that influenced dropouts in undergraduate Majors in the Federal University of Rio de Janeiro was conducted by Manhães et al. (2012). The authors used data mining techniques and observed that the students with active enrollment are students who present a regular behaviour throughout the course, enrol in a high number of subjects, and have grades well above the students who dropout, but lower to those who have completed the graduation. They applied a Naive Bayes model for visualizing the factors that distinguish students who succeed or fail in their courses, and got an overall accuracy of around 80%.

Bayer et al. (2012) described a method to get new features from student dropout data and their behaviour from a constructed social graph. The work represents a novel method that assists a classifier in learning from data aiming to predict student failure to complete the studies. They found that if students' social behaviour is included, it helps to increase the prediction rate accuracy.

Baradwaj and Pal (2012) used data mining techniques to classify the performance of students. They used decision trees and a student database to calculate a performance indicator for the current semester based on previous data. It includes class presence and marks of the class tests, assignments and seminars. Based on this analysis, it became possible to identify probable dropout students. The system also suggests students who need special attention and teachers' advice.

Kumar and Vijayalakshmi (2011) used a decision tree to help tutors get

information of weak students who have a high possibility of dropping out the program. It classifies students as PASS and FAIL level before the exams using their previous history. These are important to improve weak students performance by giving them useful guidelines.

Belloc et al. (2010) analyzed student dropout rate of the Economics and Business Faculty of Sapienza University of Rome. Their analysis used administrative data on 9,725 undergraduates students who enrolled in three-year bachelor programs from 2001 to 2007. They developed a Generalized Linear Mixed Model and identified students' characteristics that significantly impact dropout. Their empirical analysis unveils the statistically significant effect of students' factors, like citizenship and income. Simultaneously, the main findings relate a high dropout probability to a high secondary school final mark and a low individual students' performance. At the same time, they find male students dropout less likely than women and non-Italian students drop-out of the university less than Italians do.

Ayesha et al. (2010) analyzed students' learning behaviour to predict weak students. They applied data mining technologies like K-mean clustering and model-based algorithms to analyze students' learning behaviour, which help the teachers improve the performance of students and reduce the dropout ratio to a significant level.

Kotsiantis (2009) predicted, with high accuracy, the number of new dropout students. The author collected information from an online e-Learning platform for the course "Introduction in Informatics". This work was done using first-year university students' information and became an important tool to reduce dropouts. He identifies that failure in the early stage of the program is a significant factor for dropout. It used decision trees, Bayesian classifiers, logistic models, rule-based learners and random forests to identify dropout student at first-year.

A statistical analysis using a computational risk framework has done by Lassibille and Navarro Gómez (2008). This work found that academic preparedness has a significant impact on students completing their studies. Moreover, it identified that a person who started high school with an older age has a higher possibility of dropping out before graduation. It is also observed from their work that financial support helps to reduce the dropout rate. Family characteristics like family background, parents' education level, and income are essential factors in students' decision to withdraw from school that come out from this research.

Boero et al. (2005) proposed an econometric analysis of withdrawal students. Their research used two Italian universities student administrative data in a probit model of the probability that the student dropout. Their proposed

research suggests that the dropout (withdrawal) rate is high and shows that only around 8% students are likely to finish their study within the institutional time. They concluded that gender and age are two important factors for identifying the dropout factor: males have a higher likelihood of dropping out compared to females; Age has a significant positive effect with young individuals in below-average age are less likely to dropout.

Smith and Naylor (2001) used a binomial probit model to estimate the probability of an individual student to dropout before finishing the degree. It uses information from a group of students who enrolled as full time for a three or four-year degree (1989 – 1990) in an undergraduate program.

Blanchfield (1972) proposed a method for identifying college dropouts at Utica College of Syracuse University; he used multiple discriminant analysis to identify dropouts, reaching an accuracy of around 73%.

3.3 Summary

This chapter presents the student dropout concept, its impact, and previous work in this area. Reasons for a dropout can be related to economical, social and psychological issues. An early detection system is of the utmost importance to identify dropout students and guide them to continue their studies.

Chapter 4

Feature Engineering

Supervised Machine Learning is a sub-area of Machine Learning that aims at learning a function that maps an input into an output from a set of input-output example pairs (Russell and Norvig, 2010). Since it infers a function from labeled training data (Mohri et al., 2012), supervised Machine Learning is heavily data-dependent. Since the learning algorithm generalizes from the training data, the existence of a well-structured, low-noise dataset is of the utmost importance.

In this work, there has been used only students' academic records where they belong in four programs and they are Computer Science, Management, Biology, and Nursing. All the student records are collected from the information system of the University of Evora. A single record of student representing a single enrolled course performance and each record contained nineteen different information, like course enrollment year, name, credits, enroll semester, result details, etc. Mentioned information has been used to generate a dataset that is used in our experiments to identify a risk profile student.

This part of the dissertation presents the steps/process for generating the dataset. The chapter includes the study of raw data, data preprocessing, feature engineering and dataset characterization.

4.1 Raw Data

It is essential to understand the attributes and properties of the raw data for high-end targeted research work. In this study, the academic records of students are gathered from the university information system. Specifically, four programs (Management, Biology, Computer Science, and Nursing) for

undergraduate students are considered for the duration of 13 academic years (from 2006/2007 to 2018/2019)

The student's academic record included information about course enrollments and their results during the student's academic life. The information of a student is considered from the first year when he/she registers at the university until graduation or dropout. Students were anonymized and updates on study programs were considered. Table 4.1 sums-up the existing information for each course enrolled by a student. Using this raw data, the process of storing in a database and the dataset creation steps are discussed next.

4.2 Build Database

To process the raw data needs to transfer all data into a database. So in this work, Microsoft SQL server 2014 was used as a DBMS to clean, process, and store raw data. Figure 4.1 shown a simplified block diagram of the data storage process from a CSV (initially contained raw data) file to a database.

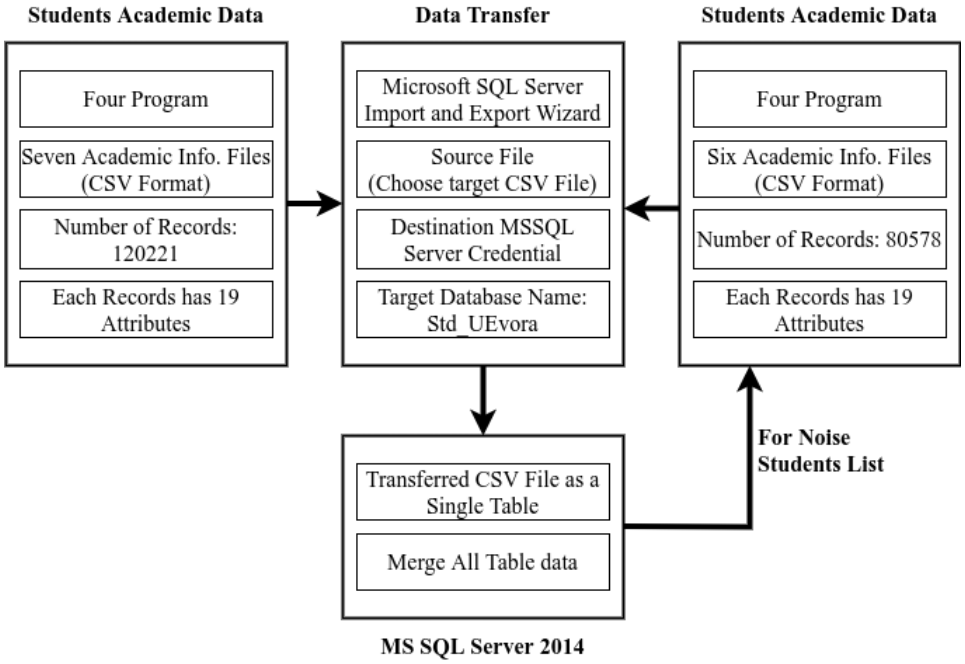


Figure 4.1: Raw data storage process from CSV file to database.

In this transfer process, there were used MSSQL Server Import and Export

Property	Type	Description
School Year	Integer	When a student enrolled in the courses
Degree	String	Name of the student degree
Department	String	Which department student belongs
Course Code	String	Course code represents the course that student take
Course Unit	String	The course belongs to which group
Regime	Character	It's present student's special regime (if have any), but here it's always **"S"**
Course Name	String	Course Name contains the course code
Course Credits	Float	Course how many ECTS contains
Edition	NULL	There is not value contained in this field; all are empty
Speciality	NULL	Same as Edition attribute
Semester	String	Mentioned semester. There are three types of value: Par, Impar, n/d
Time	String	It represents student course timing. Two value: Normal and Extraordinary
Type	String	It's mention course taken to type. Values are: Normal, Mobility, Improvement, Free, Extracurricular
Student Id	Integer	Unique Value, represent a student
Current Status	String	Student current situation (Active, Inactive, Graduate)
Student Type	String	It's present student type (Normal, Mobility)
Mark	Integer/NULL	Present student mark in a course
Result	String	Represent course result status. Statuses are Approved, Disapproved, Missed, Cancelled, Give-Up
Final Status	Character	Final status has two values: **"S"** means student pass the course, and **"N"** means student miss or fail the course.

Table 4.1: Information available for each course enrolment.

Wizard tool to import raw data (CSV file) into a database. Initially, there were seven CSV data files of four programs that contain a total number of 120221 records of student's course enrollment history with 19 attributes. Transferred each CSV file saved as a single table into the database where the data table name is the same as the CSV file. Then, imported raw data tables were merged with additional information in a master table named "LIC_-AllData_List", because it was helpful to easily manipulate all the data from a single table to clear, filter and process them. After merging all table data, these data were used to create a dataset for experiments. The following section will describe the steps of creating a dataset.

4.3 Dataset Creation

As mention earlier, the data used for this study is obtained from the University of Evora through structured academic information based on students' academic yearly performance. The aim of creating a dataset is to enable the machine learning algorithm to identify the potential student profile leading to dropouts. The final dataset is generated by processing the raw data and engineering a set of discriminate features.

4.3.1 Data Preprocessing

As already mentioned, student data was collected over a period of 13 years from four different undergraduate programs: Management, Biology, Nursing and Computer Science. Among these programs, only nursing is a four years program (totalling 240 credits), and the rest are three years of programs (totalling 180 credits).

After collecting the data then marge, all the students records into a single student record that resulted in 3480 students. Over these records, the following steps were applied:

1. Data from the school year 2018/2019 was deleted from the collected data because it only includes enrolled data and not the results obtained by each students information that could give a hint of the student's performance.
2. Data with Semester value of n/d was updated by "Par".

Then, out of 3480 students, a total of 969 students were found who have completed fewer ECTS to finished their degree, but their "Current Status"

showed them as "Graduate". After going through these 969 students, it was found that they changed their program in the middle of their academic life and that's why there is less academic information of them. For those students, academic information from their others programs was collected and merged to build their full period of study at the university.

For these 969 students, a set of conditions were verified that are presenting below:

Condition 1.

`If` $(TotalCompleteCredit \geq (240$ `OR` $180)^a)$
`Then` GRADUATE

[a] 240 for nursing; 180 for other programs.

Merging the student's academic records of their others program and applied this condition, then it was found that 223 students who are completed more or equal (240 or 180) ECTS. So their current status became the same Graduate. So, the remaining 746 students from 969 exist as noise students

Condition 2.

`If` $(TotalCompleteCredit \geq (210$ `OR` $150)^a)$
`Then` GRADUATE

[a] 210 for nursing; 150 for other programs.

Subsequently, applied another similar condition where those students were completed (210 or 150) ECTS, then marked them as Graduate and they were 473 students. Since they had been changed their program, so the rest of the credits they completed in another program. Thus, their current status is Graduate which is the same as before. After applying this condition, the scenarios are,

- Removed 473 students data from 746 noisy student list, so the rest of 273 students data still the noise data.
- A total number of 273 students information are not correctly traceable, since some academic information of these students is missing.
- So, untraceable 273 students records are removed from the student database. Now the total number of students is 3207 from 3480.

After that, it was applied two more conditions to clean and filter the data from the database. The conditions are presenting below,

Condition 3.

```
If (EnrolProgramCount ≥ 2)
Then DELETE RECORD
```

The third condition removes those students who belong to two programs at the same time. It was 46 students out of 3207 students, so we removed them from the database. Thus, now 3161 students data exist in the database from 3207 students.

Condition 4.

```
If ((Register = 2017)
AND (CompleteCredit = 0) AND (EnrolCredit = 0))
Then DELETE RECORD
```

The fourth condition applied to remove those students record who enrolled 2017 academic year but didn't have any other academic results. A total of 227 students found to apply this condition, so removed these students from the database. So final database contains 2934 students academic records.

After applied the conditions, a total number of 2934 students was used to create the dataset for the experiments; for each student, the following 8 enrolment attributes were considered from the database to making an experimental dataset.

1. Academic Year
2. Program : Management, Biology, Computer Science, Nursing
3. Semester
4. Student Id
5. Course
6. Credits
7. Mark
8. Final Status

4.4 Instance Labeling and Dataset Construction

Using the retrieved data, for each student, the annual student performance is calculated, and all the yearly records are jointed together to generate a single record presenting the academic path of a specific student. The final experimental dataset was generated after different logical conditions. The overall process is shown in Figure 4.2.

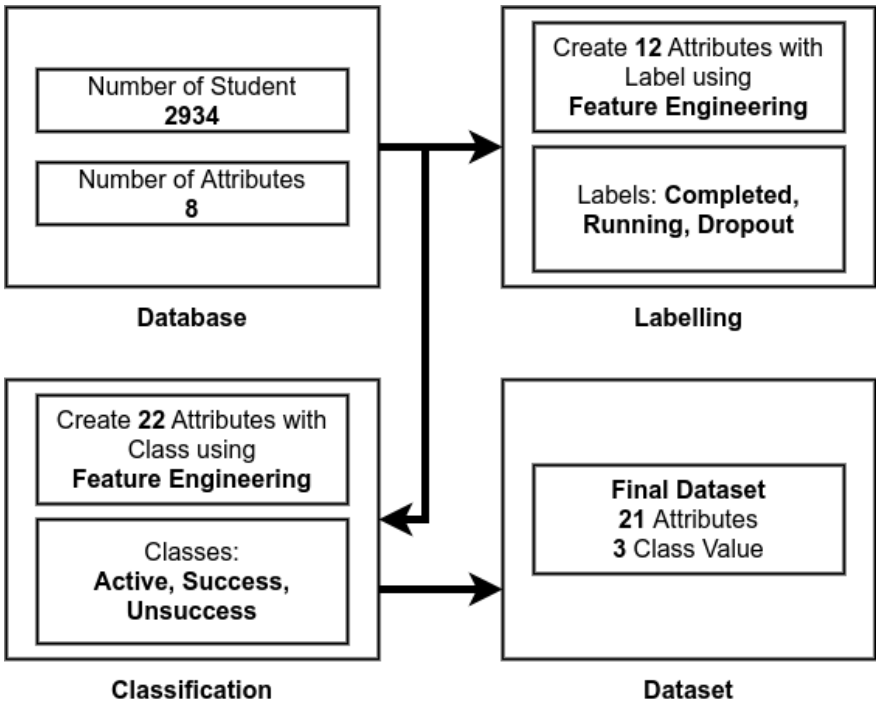


Figure 4.2: Overall process diagram of Dataset Construction.

4.4.1 Labelling

Initially, data from the university information system did not have "Student Current Status" property. So, one class label from "Completed, Running and Dropout" is assigned to each example. Figure 4.3 shows the process used for adding the label.

Parallelly, for each student's consecutive last five years data was considered from the database to make a dataset. The following 21 attributes contained each student information that was used to generate a dataset.

- Student Id
- Program
- Total Attempted Credits
- Total Completed Credits
- Total ECTS
- Student Current Status
- Year_0, 1, 2, 3, 4 (Year, Attempted Credits, Completed Credits)

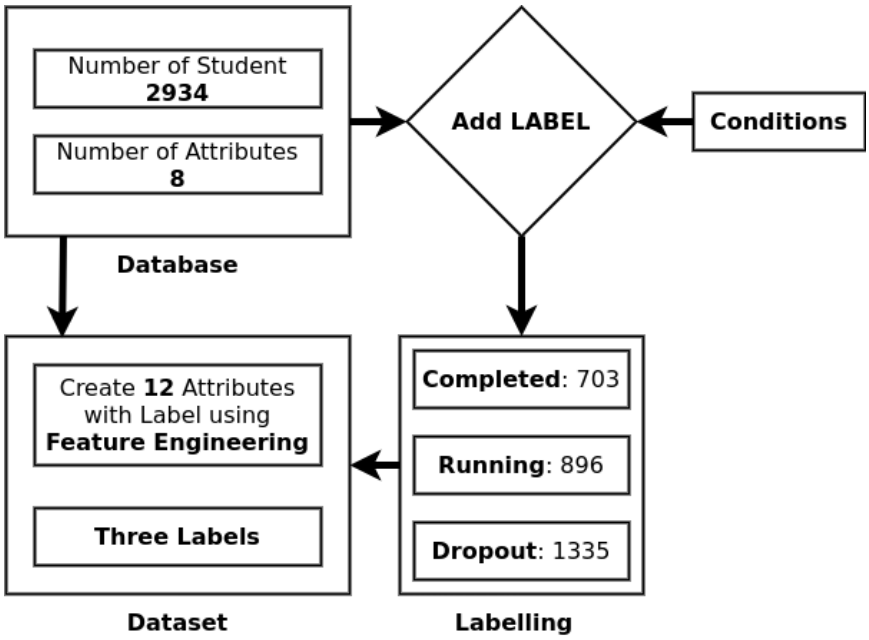


Figure 4.3: Full process of add label and build dataset.

The terms of the terms used to attach labels to the created dataset are presented below,

If $(TotalCompleteCredit \geq (240$ OR $180)$[a])
Then COMPLETED
ElseIf $((EnrolConsecutiveYear = 3$ AND $CompleteAnyECTS = True)$
OR $(EnrolConsecutiveYear = 2$ AND $CompleteAnyECTS = False))$
Then RUNNING
Else DROPOUT

[a] 240 for nursing; 180 for other programs.

After applying the rules where a student completed (240 or 180) ECTS, then they were marked as Completed (Graduate). On the other hand, if a student consecutively enrolls the last three academic years and successfully completed any ECTS or a student consecutively enrolls the last two academic years and did not complete any ECTS. It means this student is a Running student. Lastly, if a student didn't fulfill any previous rules then he/she had been a dropout student.

The outcome of the class label resulted in 703 Completed, 896 Running, and 1335 Dropout students. When this information had compared after getting the student's current status from the institution's information system, it

hadn't add-up. Thus, we have continued to the next rule "Classification" to find a better solution.

4.4.2 Classification

In this iteration, there were introduced three new class, named: "Active", "Success" and "Unsuccess". The overall process of adding class value is shown in Figure 4.4.

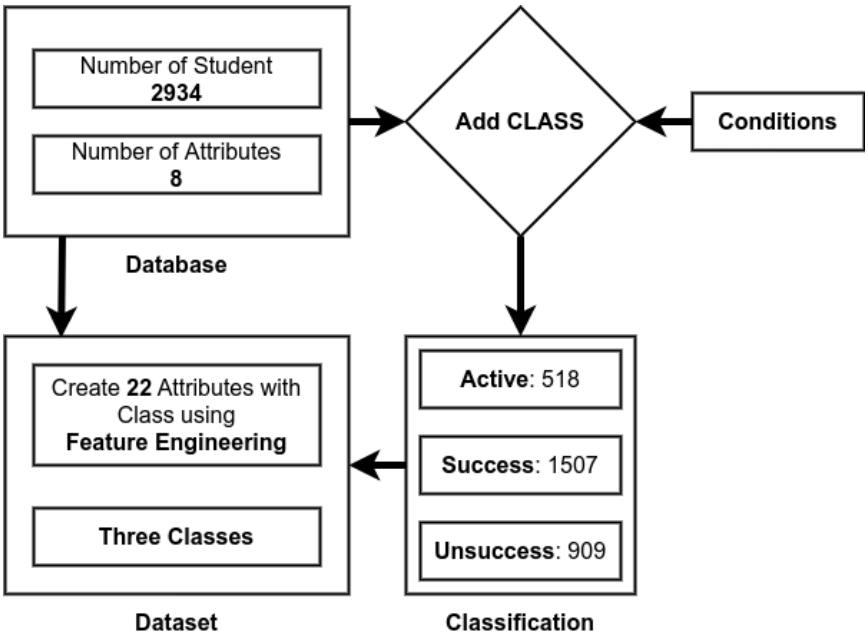


Figure 4.4: Full process of add class and build dataset.

At the same time, making a dataset from the database using feature engineering where considered a student's consecutive five-year academic record. The dataset attributes are presenting below,

- Program ECTS
- Program Name: Management, Biology, CS, Nursing
- Last Year Enrolled ECTS and Average Grade
- Year_1, 2, 3, 4 (Enrolled and Completed ECTS, Average Grade)
- Rest of Years Enrolled and Completed ECTS

The condition rules that were used to add a new class for each row for the created dataset are presenting as follows:

If $Registered = 2017$ AND $CompleteCredit > 0$
Then ACTIVE
ElseIf $Registered < 2017$ AND $CompleteCredit \geq (210$ OR $150)$[a]
Then SUCCESS
Else UNSUCCESS

[a] 210 for nursing; 150 for other programs. This corresponds completing all except the credits of one semester.

From the rules, a student was ACTIVE when he/she registered in 2017 and successfully completed any ECTS. On the other hand, a student was SUCCESS when he/she last registered before 2017 and completed (240 or 180) ECTS. In the end, a student didn't fulfill any previous rules then he/she had been marked as an UNSUCCESS student.

The outcome of the class assigning resulted in 518 Active, 1507 Success, and 909 are Unsuccess students. This information had compared to the institutional information system where rules mark class and institutional status match.

4.5 Dataset Characterization

The final dataset of 13 years composed of 22 attributes with the class was built. Table 4.2 presents them,

Attributes	**Number**	**Type**
program_ects	1	int
program_name: man, bio, cs, nurse	4	bool (all)
year_0: enrol , avg_grade	2	int, float
year_1: enrol, complete, avg_grade	3	int, int, float
year_2: enrol, complete, avg_grade	3	int, int, float
year_3: enrol, complete, avg_grade	3	int, int, float
year_4: enrol, complete, avg_grade	3	int, int, float
year_rest: enrol, complete	2	int, int
Class : Active, Success, and Unsuccess		

Table 4.2: Final Dataset attributes.

To build the final dataset, there were used 119407 academic records of 2934 students. For each student, all the yearly performance were joined together to generate a row for the final dataset which represent the student's academic path.

The annual student performance is given by three attributes: the total num-

ber of enrolled and completed credits and average grade. This information was compiled for the student's five most recent academic years plus the performance calculated over the remaining student academic life. For students that successfully completed the program in less than 5 academic years, the values for attributes of oldest years were filled with zeros.

After finalize the dataset, two labelling used in the experiments. They are,

- Three class: Active, Success and Unsuccess
- Two class: Success and Unsuccess (Active and Success merged)

The distribution of these two labelling are presented in Table 4.3.

	3 Class			2 Class	
Total	Active	Success	Unsucces	Success	Unsuccess
2934	518	1507	909	2025	909

Table 4.3: Class distribution of final dataset for two labelling.

Chapter 5

Classification model

In this chapter, the experiments performed for our work and their results are explained. First, performance metrics are presented that describe the evaluation of machine learning algorithms. In the next section, the experimental setup is revealed which defines the implementation of the algorithm during experiments. Finally, the experimental result section presents the results obtained and discusses.

5.1 Performance Metrics

For this work, the performance of the system was measured by the Accuracy and F1-Measure. Accuracy is the fraction of predictions a model got right. Equation 5.1 calculates accuracy from the confusion matrix.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.1}$$

Table 5.1 presents the confusion matrix values for Equation 5.1,

		Actual	
Predict		Positive	Negative
	Positive	TP: True Positive	FN: False Negative
	Negative	FP: False Positive	TN: True Negative

Table 5.1: Confusion Matrix for Classification.

Equation 5.2 calculate f1-measure, it is harmonic mean between precision and recall. These measures are given by Equation 5.3 and 5.4

$$F1 - Measure = \frac{2 * (Recall * Precision)}{Recall + Precision} \tag{5.2}$$

$$Precision = \frac{TP}{TP + FP} \tag{5.3}$$

$$Recall = \frac{TP}{TP + FN} \tag{5.4}$$

5.2 Experimental Setup

In order to evaluate the performance of the algorithms as well as the most relevant set of attributes, several experiments were carried out. The two labellings used in the experiments: One with classes: Active, Success, and Unsuccess, and another with two classes: Success and Unsuccess.

In this work, there were four machine learning algorithms applied, named: DT, NB, RF and SVM. All experiments have been performed using Weka 3.8.1 toolkit (Hall et al., 2009).

In the experiments, the dataset was split into 70% of examples for training (2052 samples) and 30% for testing (882 samples). Table 5.2 presents the class distribution of the considered two labelling class problem.

		3 Class			2 Class	
	Total	Active	Success	Unsucces	Success	Unsuccess
Train	2052	362	1054	636	1416	636
Test	882	156	453	273	609	273

Table 5.2: Class Distribution.

5.3 Experiments

To test the importance of the enrolled program and grade information, four different attribute subsets were used to build classification models:

- att_1: without *program_name*, without *avg_grade*
- att_2: with *program_name*, without *avg_grade*
- att_3: without *program_name*, with *avg_grade*
- att_4: with *program_name*, with *avg_grade*

Already mentioned that we were considered to use two labelling problems. One is the three-class problem and another is the two-class problem. For each problem, we used four different sets of attributes that previously mentioned. From table 4.2, there were 22 attributes with class value. But in subset att_1 had 13 attributes where program name and average grade were not considered. Similarly, att_2 had 17, att_3 also 18 and att_4 had 22 attributes with class.

The next following section will describe the results of the two labelling problems with the measurement of accuracy, precision, recall and F1-Measure.

5.3.1 Labelling A: Active, Success, Unsuccess

Table 5.3 shows the results obtained over the test set for each set of attributes and machine learning algorithm. The best accuracy is performed by the Random Forest and it reaches the value 90% above. This algorithm presents a similar result for the four sets of attributes. Contrary, Naïve Bayes seems to be more sensitive to the set of attributes: it reaches 74% when not using the program and average grade and 82% when using both attributes.

Attributes	DT (%)	NB (%)	RF (%)	SVM (%)
Att_1	87.3	**74.0**	**90.9**	88.9
Att_2	88.8	80.3	90.4	89.2
Att_3	86.8	75.9	90.0	89.0
Att_4	88.6	82.1	90.3	89.7

Table 5.3: Labelling A: Accuracy

As can be seen in Table 5.4, SVM outperforms all the other algorithms by achieving 98.4% of precision value when using the program name and the average grade (att_4). On the other hand, DT seems to have the worse precision values with not using program name and average grade (att_1).

Attributes	DT (%)	NB (%)	RF (%)	SVM (%)
Att_1	**89.8**	94.1	93.5	96.1
Att_2	91.6	93.8	92.8	96.9
Att_3	92.0	94.6	94.9	98.0
Att_4	92.0	94.3	94.5	**98.4**

Table 5.4: Labelling A: Precision of Unsuccess class.

The recall results over the Unsuccess class presented in Table 5.5. As can be seen, the RF algorithm outperforms all the other algorithms by achieving

95.6% of recall when using no program name and nor average grade (att_1). NB presents the worse performance, but still good recall.

Attributes	DT (%)	NB (%)	RF (%)	SVM (%)
Att_1	93.8	87.9	**95.6**	91.2
Att_2	91.9	88.6	94.5	91.9
Att_3	92.3	89.4	94.9	90.5
Att_4	93.0	90.8	94.9	90.8

Table 5.5: Labelling A: Recall of Unsuccess class.

Table 5.6 presents the f1-measure results over Unsuccess class. For f1-measure, all experiments presented values above 90% ranging from 90.9% for NB without program_name and average_grade (att_1) to 94.9% for RF without program_name but with average_grade (att_1). Here, also RF is outperforming all the other algorithms by achieving 94.9% of f1-measure. As a conclusion and taking into consideration that f1-measure has a big performance (we want to detect as many students as we can in a risk of challenging), one can state that RF without program_name but with average_grade presents the best model. Moreover, program_name and average_grade don't move to overall comparison.

Attributes	DT (%)	NB (%)	RF (%)	SVM (%)
Att_1	91.8	**90.9**	94.6	93.6
Att_2	91.8	91.1	93.6	94.4
Att_3	92.1	91.9	**94.9**	94.1
Att_4	92.5	92.5	94.7	94.5

Table 5.6: Labelling A: F1-Measure of Unsuccess class.

5.3.2 Labelling B: Success, Unsuccess

Table 5.7 shows the accuracy obtained when considering only 2 classes of labelling. The accuracy results are similar for all the algorithms and set of attributes ranging from 92.4% to 96.8%. The highest accuracy is obtained with RF and the smallest with NB, both using average_grade but without program_name (att_3).

Table 5.8 presents the precision results over the unsuccess class. The precision result varies between 84.8% for NB and 96.5% for SVM. Results for

Attributes	DT (%)	NB (%)	RF (%)	SVM (%)
Att_1	94.4	92.9	96.5	95.5
Att_2	94.9	92.7	96.2	96.2
Att_3	96.0	**92.4**	**96.8**	95.9
Att_4	96.2	93.7	96.6	96.5

Table 5.7: Labelling B: Accuracy.

RF are at most 1% smaller (for experiments using program_name but no average_grade).

Attributes	DT (%)	NB (%)	RF (%)	SVM (%)
Att_1	92.4	88.6	95.5	95.7
Att_2	92.5	87.7	95.4	**96.5**
Att_3	93.1	**84.8**	95.9	96.1
Att_4	93.8	88.3	96.2	96.2

Table 5.8: Labelling B: Precision of Unsuccess class.

The recall results over the unsuccess class are presented in Table 5.9. Recall value varies between 88.3% for NB (without program_name and average_grade) to 93.8% for RF (without program_name and with average_grade).

Attributes	DT (%)	NB (%)	RF (%)	SVM (%)
Att_1	89.4	**88.3**	93.0	89.4
Att_2	90.8	89.0	91.9	90.8
Att_3	94.1	91.9	**93.8**	90.5
Att_4	93.8	91.6	92.7	92.3

Table 5.9: Labelling B: Recall of Unsuccess class.

Table 5.10 present the f1-measure results over the unsuccess class of the test set. For f1-measure values range between 85.9% to 94.8% for the experiments mentioned. And RF algorithm is outperforming all the other algorithms by achieving above 93% score.

5.3.3 Overall Comparison

When conjoining the algorithms SVM presents the best precision result while RF the best recall ones, having both similar f1-measure performances. So conjoining the importance of the different sets of attributes, one can say that

Attributes	DT (%)	NB (%)	RF (%)	SVM (%)
Att_1	90.9	**85.9**	94.2	92.4
Att_2	91.7	88.4	93.7	93.6
Att_3	93.6	88.2	**94.8**	93.2
Att_4	93.8	89.9	94.4	94.2

Table 5.10: Labelling B: F1-Measure of Unsuccess class.

NB is non-sensitive for the 3 classes labelling but that's not true for the 2 classes labelling.

Again, conjoining both labelling higher accuracy are achieved for the 2 classes problem with differences of almost 20% (NB with att_1 set of attributes) to 6% (RF with att_1 set of attributes). On the other hand, looking at recall values the difference is smaller: the maximum difference is less than 4% (SVM with att_4 set of attributes).

Chapter 6

Conclusions and Future Work

This work focus on finding out the risky students profile using students previous academic records. This chapter presents the conclusion of the work and also presents the future work.

6.1 Conclusions

"University student dropout" occur at any time in their academic year. Academic performance is the spotlight factor in higher education. Our aim is to predict the current risky students by analyzing the student's academic records using a machine learning algorithm.

In this context, our proposed idea was to feature engineering to enable the ML algorithms to build a classification model to detect students at risk of dropping out. Thus collect the full-time data of the academic year 2006/2007 to 2018/2019 with students enrollment information and the average grade was harvested from the institutional database.

As a part of descriptive feature analysis, 2934 students' academic records were acquired from the University of Évora information system and multiple logical iterations were performed before generating the final attribute-label pair dataset for supervised machine learning.

In our work, we trained multiple classification models using different machine learning algorithms, namely Decision Tree, Random Forest, Naïve Bayes, and Support Vector Machine. The developed models were evaluated over a test set resulting in accuracy of 96.8% and a recall of 93.8% for 2 class problems and 90.9% accuracy and 95.6% recall for the 3 classes labelling. These results were obtained using the Random Forest ML algorithm with

the att_1 (without program_name and avarage_grade) and att_3 (without program_name) set for attributes respectively 3 class and 2 class problems.

6.2 Future Work

Being composed of several modules, it is certain that our approach can still be improved. As future work, we intend not only to continue improving the individual modules but also extend this work to:

- Enrich the dataset by adding more programs. The current version of dataset has only four programs.
- Include students personal (including gender and age), financial and social media information as attributes.
- Develop a hybrid system that combines both algorithms and real-time student activities in social media since social media involvement play in a significant role to dropout. (Mahoney, 2014).

Publication

[1] Sharmin Sultana Prite, Teresa Gonçalves and Luís Rato. "Identifying Risky Dropout Student Profiles using Machine Learning Models". *RECPAD 2020: 26th Portuguese Conference on Pattern Recognition. October 30, 2020 - Remote Event, University of Évora, Portugal.*

Funding

This work was supported by the Erasmus Mundus LEADER (*Links in Europe and Asia for engineering, eDucation, Enterprise and Research Organization*) project.

Bibliography

Abu-Oda, G. S. and El-Halees, A. M. (2015). Data mining in higher education: university student dropout case study. *Data mining in higher education: university student dropout case study*, 5(1).

Akcesme, B. and Can, M. (2016). Support vector machines for predicting protein structural classes via pseudo images derived from amino acid sequences. *Southeast Europe Journal of Soft Computing*, 5(1).

Allen, J., Robbins, S. B., Casillas, A., and Oh, I.-S. (2008). Third-year college retention and transfer: Effects of academic performance, motivation, and social connectedness. *Research in Higher Education*, 49(7):647–664.

Ameri, S., Fard, M. J., Chinnam, R. B., and Reddy, C. K. (2016). Survival analysis based framework for early prediction of student dropouts. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 903–912.

Andrei, T., Teodorescu, D., and Oancea, B. (2011). Characteristics and causes of school dropout in the countries of the european union. *Procedia-Social and Behavioral Sciences*, 28:328–332.

Ayesha, S., Mustafa, T., Sattar, A. R., and Khan, M. I. (2010). Data mining model for higher education system. *European Journal of Scientific Research*, 43(1):24–29.

Azuaje, F. (2006). Witten ih, frank e: Data mining: Practical machine learning tools and techniques 2nd edition.

Baradwaj, B. K. and Pal, S. (2012). Mining educational data to analyze students' performance. *arXiv preprint arXiv:1201.3417*.

Bayer, J., Bydzovská, H., Géryk, J., Obsivac, T., and Popelinsky, L. (2012). Predicting drop-out from social behaviour of students. *International Educational Data Mining Society*.

Belloc, F., Maruotti, A., and Petrella, L. (2010). University drop-out: an italian experience. *Higher education*, 60(2):127–138.

Belmokre, A., Mihoubi, M. K., and Santillán, D. (2019). Analysis of dam behavior by statistical models: application of the random forest approach. *KSCE Journal of Civil Engineering*, 23(11):4800–4811.

Berens, J., Schneider, K., Görtz, S., Oster, S., and Burghoff, J. (2018). Early detection of students at risk–predicting student dropouts using administrative student data and machine learning methods. *CESifo Working Paper*.

Bhardwaj, B. K. and Pal, S. (2012). Data mining: A prediction for performance improvement using classification. *arXiv preprint arXiv:1201.3418*.

Bill and Melinda, G. F. (Accessed November 20, 2020). The bill & melinda gates foundation, previously the william h. gates foundation, is an american private foundation founded by bill and melinda gates. In *https://www.gatesfoundation.org/*.

Blanchfield, W. C. (1972). College dropout identification: An economic analysis. *The Journal of Human Resources*, 7(4):540–544.

Boero, G., Laureti, T., Naylor, R., et al. (2005). An econometric analysis of student withdrawal and progression in post-reform italian universities. Technical report, Centre for North South Economic Research, University of Cagliari and Sassari .

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Bridgeland, J. M., DiIulio Jr, J. J., and Morison, K. B. (2006). The silent epidemic: Perspectives of high school dropouts. *Civic Enterprises*.

Catterall, J. S. (1987). On the social costs of dropping out of school. *The High School Journal*, 71(1):19–30.

CEME, N. (2017). Student performance prediction and risk analysis by using data mining approach. *Journal of Intelligent Computing Volume*, 8(2):49.

Chen, R. (2012). Institutional characteristics and college student dropout risks: A multilevel event history analysis. *Research in Higher education*, 53(5):487–505.

Chen, Y., Johri, A., and Rangwala, H. (2018). Running out of stem: a comparative study across stem majors of college students at-risk of dropping out early. In *Proceedings of the 8th international conference on learning analytics and knowledge*, pages 270–279.

Christopher, A. A. and alias Balamurugan, S. A. (2014). Prediction of warning level in aircraft accidents using data mining techniques. *The Aeronautical Journal*, 118(1206):935–952.

Connolly, T. M. and Begg, C. E. (2005). *Database systems: a practical approach to design, implementation, and management*. Pearson Education.

De Witte, K., Cabus, S., Thyssen, G., Groot, W., and van Den Brink, H. M. (2013). A critical review of the literature on school dropout. *Educational Research Review*, 10:13–28.

Demetriou, C. and Schmitz-Sciborski, A. (2011). Integration, motivation, strengths and optimism: Retention theories past, present and future. In *Proceedings of the 7th National Symposium on student retention*, volume 201.

Dey, S. K., Rahman, M., et al. (2020). Effects of machine learning approach in flow-based anomaly detection on software-defined networking. *Symmetry*, 12(1):7.

Drake, J. D. and Worsley, J. C. (2002). *Practical PostgreSQL*. " O'Reilly Media, Inc.".

en.Wikipedia (Accessed January 15, 2021). Weka (machine learning). In *https://en.wikipedia.org/wiki/Weka_(machine_learning)*.

Fall, A.-M. and Roberts, G. (2012). High school dropouts: Interactions between social context, self-perceptions, school engagement, and student dropout. *Journal of adolescence*, 35(4):787–798.

Francesca, B. (Accessed November 20, 2020). The student net- using machine learning algorithms to address our failing guidance system. In *https://bit.ly/35XrXBO*.

Garner, S. R., Cunningham, S. J., Holmes, G., Nevill-Manning, C. G., and Witten, I. H. (1995). Applying a machine learning workbench: Experience with agricultural databases. In *Proceedings of the Machine Learning in Practice Workshop*.

Greenwald, R., Stackowiak, R., and Stern, J. (2013). *Oracle essentials: Oracle database 12c*. " O'Reilly Media, Inc.".

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18.

Headley, S. (2003). Realising australia's commitment to young people: Scope, benefits, cost, evaluation and implementation. *Youth Studies Australia*, 22(1):54–55.

Holmes, G., Donkin, A., and Witten, I. H. (1994). Weka: A machine learning workbench. In *Proceedings of ANZIIS'94-Australian New Zealnd Intelligent Information Systems Conference*, pages 357–361. IEEE.

Horning, N. (2013). Introduction to decision trees and random forests. *American Museum of Natural History*, 2:1–27.

Islam, S. and Amin, S. H. (2020). Prediction of probable backorder scenarios in the supply chain using distributed random forest and gradient boosting machine learning techniques. *Journal of Big Data*, 7(1):1–22.

Kim, D. and Kim, S. (2018). Sustainable education: analyzing the determinants of university student dropout by nonlinear panel data models. *Sustainability*, 10(4):954.

Kotsiantis, S. (2009). Educational data mining: a case study for predicting dropout-prone students. *International Journal of Knowledge Engineering and Soft Data Paradigms*, 1(2):101–111.

Kumar, S. A. and Vijayalakshmi, M. (2011). Implication of classification techniques in predicting students recital. *Int. J. Data Mining Knowl. Manage. Process (IJDKP)*, 1(5):41–51.

Lassibille, G. and Navarro Gómez, L. (2008). Why do higher education students drop out? evidence from spain. *Education Economics*, 16(1):89–105.

Lee, S. and Chung, J. Y. (2019). The machine learning-based dropout early warning system for improving the performance of dropout prediction. *Applied Sciences*, 9(15):3093.

Letkowski, J. (2015). Doing database design with mysql. *Journal of Technology Research*, 6:1.

Mahoney, J. L. (2014). School extracurricular activity participation and early school dropout: A mixed-method study of the role of peer social networks. *Journal of Educational and Developmental Psychology*, 4(1):143.

Manar, A. and Ploix, S. (2015). Machine learning with python/scikit-learn-application to the estimation of occupancy and human activities. *SIMUREX, 2015*.

Manhães, L. M. B., Cruz, S., Costa, R. J. M., Zavaleta, J., and Zimbrão, G. (2012). Identificação dos fatores que influenciam a evasão em cursos de graduação através de sistemas baseados em mineração de dados: Uma abordagem quantitativa. *Anais do VIII Simpósio Brasileiro de Sistemas de Informação, São Paulo*.

Mduma, N., Kalegele, K., and Machuve, D. (2019). An ensemble predictive model based prototype for student drop-out in secondary schools. *Journal of Information Systems Engineering and Management*, 4(3):em0094.

Mikheeva, V. D. (2006). *Microsoft Access 2003 (+ CD)*. BHV-Petersburg.

Mistry, R. and Misner, S. (2014). *Introducing Microsoft SQL Server 2014*. Microsoft Press.

Mitchell, T. (1997). Introduction to machine learning. *Machine Learning*, 7:2–5.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). Foundations of machine learning. adaptive computation and machine learning. *MIT Press*, 31:32.

Obaidullah, S. M., Ahmed, S., Gonçalves, T., and Rato, L. (2018). Rmid: a novel and efficient image descriptor for mammogram mass classification. In *Conference on Information Technology, Systems Research and Computational Physics*, pages 229–240. Springer.

Quinlan, J. R. (1996). Learning decision tree classifiers. *ACM Computing Surveys (CSUR)*, 28(1):71–72.

Rai, S. (2014). Student dropout risk assessment in undergraduate course at residential university. *arXiv preprint arXiv:1405.3727*.

Raisman, N. (2013). The cost of college attrition at four-year colleges & universities. policy perspectives. *Educational Policy Institute*.

Rokach, L. and Maimon, O. Z. (2008). *Data mining with decision trees: theory and applications*, volume 69. World scientific.

Russell, S. and Norvig, P. (2002). Artificial intelligence: a modern approach. -.

Russell, S. J. and Norvig, P. (2010). Artificial intelligence-a modern approach, third international edition.

Smith, J. P. and Naylor, R. A. (2001). Dropping out of university: a statistical analysis of the probability of withdrawal for uk university students. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 164(2):389–405.

Ugalde, D. S. (2015). Android app for automatic web page classification: Analysis of text and visual features. Master's thesis, -.

Wikipedia (Accessed January 1, 2021). Decision tree learning. In *https://en.wikipedia.org/wiki/Decision_ tree_ learning*.

Wikipedia (Accessed November 29, 2020). Database. In *https://en.wikipedia.org/wiki/Database*.

Wikipédia (Accessed January 15, 2021). support vector machine. In *https://pt.wikipedia.org/wiki/M%C3%A1quina_de_vetores_de_suporte*.

Yuan, L., Chen, H., and Gong, J. (2018). Classifications based decision tree and random forests for fanjing mountains tea. In *IOP Conference Series: Materials Science and Engineering*, volume 394, page 052002. IOP Publishing.