Royaume du Maroc
Université Mohammed V de Rabat
Faculté des Sciences de l'Education

المملكة المغربية جامعة
محمد الخامس بالرباط
كلية علوم التربية

Master : Applied Linguistics & English Language Teaching

Predicting Students Performance by means of Machine Learning Algorithms : K-NN & Decision Tree CART

A thesis submitted in partial fulfilment of the requirements for a Master's Degree in Applied Linguistics and English Language Teaching

Submitted by:

Fatima Zahrae Aouane

Supervised by:

Salmane Tariq El Allami

Academic Year 2023/2024

Acknowledgements

I would like to express my profound appreciation to my supervisor Salmane Tariq El Allami for his indispensable contributions, his guidance, encouragement and support have been invaluable in shaping the intellectual depth and quality of this academic endeavor.

I am incredibly grateful to all professors who made this master program a heaven of knowledge and productivity. Professor Ichebah Mina, Pr. Mehdaoui Zoulikha, Pr. Amzil Amine, Pr. Ech-charfi Ahmed, Pr. Arssi Abdelaziz, Pr. Asri Khalid, Pr. Boutouam Abdallah and Pr Salmane Triq El Allami.

I would also like to thank my parents and brothers for their continuous support and encouragement throughout my academic journey. I extend my deepest thanks and gratitude to my brother Mohammed Aouane, his endless love, trust and support have been driving forces throughout my life.

Lastly, I would like to thank my dear friends and classmates, for the emotional support, laughs and encouragement during this wonderful two years of my master degree.

Abstract

Education is the cornerstone of all societies, providing individuals with the necessary knowledge and the needed skills to navigate the world and build the future. With the rapid advancement of digitalisation and the increased volume of data within educational contexts, there is a growing curiosity in comprehending how these advancements can be utilized to forecast students' performance. Increasingly, educational institutions start to recognize the potential of Big Data as well as Data Mining Techniques for more effective and more efficient learning and teaching approaches. Measuring, analysing and understanding learning processes and outcomes are all encompassed in the field of Learning Analytics. LA presents opportunities to anticipate learners' outcomes, prompt interventions and adaptations of the curricular, and suggest innovative strategies to increase the number of successful students. This research uses machine learning to predict students' success or failure based on their historical and demographic data. Two well-known machine learning algorithms were employed on educational datasets. These techniques are KNN 'K-Nearest Neighbors' and Decision Tree. The results indicate that these machine learning algorithms were successful in predicting students' performance, which validate the potential of using them to obtain accurate understanding of the learning and teaching processes and methodologies.

Key words*; predicting students' performance, data mining techniques, machine learning, success, failure, teaching.*

List of Tables:

Table of Contents

General introduction

Nowadays, to contribute to the progress of a nation, the focus if often directed towards investing in the main profitable domains, namely the economy, politics and industry. Undoubtedly, these sectors hold significant impacts on a nation’s growth, however, it is crucial not to underestimate the role of the educational system. It is the cornerstone and the primary driver of a nation’s success.

In Morocco, education plays a significant role in shaping responsible, and intellectual citizens. Thus, supporting students throughout their academic career remains one of the crucial measures to adopt for the success of their development. Students are the main actors in the whole system, therefore, they necessitate assessment and guidance constantly. As long as we cannot change a confirmed past, getting in-depth insights on the factors influencing students’ success or failure is imperative for correcting and intervening before irreversible losses occur. That is to say, there is an immense need for implementing predictive techniques to control and intervene at the right moment. We often lack decision-making tools in this sector, wherein we need it most.

In light of these circumstances, this research has undertake the responsibility of using machine learning algorithms to predict students’ performance, filling a critical gap in teaching and learning. Unlike conventional approaches, the purpose behind this research is to forecast students’ success or failure in advance. By doing so, it significantly contributes to applied linguistics, providing effective feedback before students encounter any challenges. This ability to predict has further implications for teaching methodologies, decisions, and best practices. The findings enhance our understanding of students' performance and offer actionable insights as well, reshaping how educators support students and the strategies they utilize to intervene. This research represents the intersection of applied linguistics and AI, and opens the door for further investigations.

Students’ failure is a real issue in education. Certain students from the same class, studying the same courses, with the same teachers, and setting for the same exams succeed, while others fail. The ultimate goal of all schools is for their students to achieve good results; as it reflects the quality of education. During the academic year, the pedagogical teams hold meetings to discuss and assess the progress of their students, and proceed with deliberations afterwards. Although students differ in terms of their learning styles, there are still a large part of external and internal variables responsible for students’ success and failure. Over the years, and due to the fact that the obtained results from students did not yield any satisfaction, it becomes of crucial importance to analyse students’ educational paths and backgrounds in order to determine the variables contributing to students’ success or failure.

Students’ academic performance can be affected by several factors. According to *GIIS Abu Dhabi.org,* social or environmental factors have tremendous impact on students; uncomfortable learning atmosphere, family background such as conflicts, parents’ income, negative home surroundings can all affect students’ motivation and performance. Learning infrastructure can also limit students’ access to resources, an unhealthy lifestyle, drinking alcohol, being distracted, losing interest, and even small details such as how long it takes a student to go to school each day can also impact their academic performance. These and other variables, which can either be external to the students such as economic and social aspects, or internal ones such as motivation, family relationships, romantic relationships etc, constitute the dataset for this research paper.

Defining the problem:

The true problem of students’ failure is not blaming them for not studying hard nor blaming teachers for not dedicating more efforts and time, although people often blame these two, the real problem is our inability of changing the fact that a student failed. We definitely cannot alter the past, but we can predict the future, and hence, prepare for it. Conversely to the traditional approaches, the main objective here is to forecast students’ success or failure before they happen rather than assessing their past achievements. Generally, our actions is solely a reaction to something that already happened, however, and by means of data-driven techniques, we possess the ability of predicting upcoming results.

Purpose of the study:

Driven by the need to address the aforementioned concerns, this research aims to use machine learning algorithms to analyse the variables contributing to students’ success or failure for a more general aim which is predicting students’ future academic struggles. Employing machine learning algorithms, this study seeks to purify and deepen our understanding of the complex factors influencing students’ outcomes.

Hypotheses:

General hypothesis:

Machine learning techniques can accurately forecast students' academic performance based on the historical and contextual data provided.

Specific hypotheses:

1. Machine learning algorithms can accurately identify variables that differentiate students at risk of failing from students who perform well.
2. Machine learning techniques; K-NN, and Decision Tree CART can be used to predict with a high degree of accuracy students who are more likely to succeed and those who are less likely to, depending on students' data.

Research questions:

In more operational terms, this study attempts to answer the following questions:

1. How well do machine learning algorithms detect hidden patterns in students' performance datasets?
2. Can these detected patterns accurately predict a new unseen datapoint and classifies it as success or at risk of failure?

Definition of terms:

Algorithms: step-by-step instructions that mine large datasets to uncover hidden patterns and make predictions.

Machine Learning: type of AI focused on building computers that learn from data.

Data Mining: the process of extracting useful and relevant insights from large datasets.

EDM: involves the use of DM techniques on data collected via various educational systems. (Berland et al., 2014)

Data set: an organized collection of information used for analysis and discovery.

K-NN: it predicts the class of a new est data point finding its K nearest neighbours' class.

Decision Tree: the most robust classification technique in DM. It is a flowchart similar to a tree.

NB: these concepts will be explained as we proceed.

Chapter 1

Chapter 1: Review of the Literature

1. Introduction to Data Mining and Machine Learning in Education:

This part introduces various definitions and experiments dealing differently with Data mining techniques and machine learning algorithms but driven by pretty much the same needs and objectives. In brief, this part aims to investigate what experts have said about predicting students' performance by means of machine learning and data mining techniques.

The attempt to conduct an extensive literature review found about data mining techniques is definitely a real challenge. While similar research exists, it is important to note that the volume is obviously lower in Morocco compared to other contexts, which in turn, creates a critical gap.

To ensure the comprehensibility of the content presented in this literature review, we will tackle the general concepts first, then, as we build the necessary knowledge, we move to the more specific.

2. Overview of Educational Data Mining:

Having a short history, Educational Data Mining succeeded to reach the globe thanks to the growing use of technology. In 2005, EDM first emerged with a series of workshops. By 2008, these workshops became an annual conference, which continued in 2009 with the launch of a dedicated journal, and was followed by formalizing the International Educational Data Mining society. By definition, EDM is ' An emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in.'' according to *www.educationaldatamining.org* website.

EDM bridges the gap between data mining techniques and educational practices (Berland et al., 2014). Through analysing data collected from different educational systems, it aims to advance the quality of education (Baker & Yacef, 2009). Recent EDM researches focus on inquiring students' learning and behaviour, examining educational techniques and materials, forecasting their performance and dropout, as well as providing valuable recommendations and insights to students (Romero & Ventura, 2020). Through inquiring their learning and behaviour, EDM assists educators to comprehend the ways in which students learn and interact with the educational materials presented to them, doing so, it helps enlarge our comprehension about teaching methods. Moreover, thanks to the prediction of students' performance, educators can accurately identify students at risk of failing, hence, intervene and support to assure their academic success (Lu et al., 2008; Wakelam et al., 2020).

3. Implementation of EDM to forecast students' performance:

Despite the fact that data mining is gaining traction all over the globe, the EDM conducted studies are still minimal in Morocco, especially from the part of applied linguistics. Recently, a research was conducted on predicting students' final performance using artificial neural networks in the region of Guelmim Oued Noun. Ahajjam, et al (2022) published a paper in which they use ML (machine learning) and data mining techniques to forecast the students' Baccalaureate average through the grades of their common core courses. Researchers M. Aitdaoud et al (2019) conducted a study seeking to improve the quality of education in Moroccan high schools through leveraging data mining techniques. Their research centred around gathering and analysing educational data from a private school in the region of Casablanca over a period of five years. The study revealed the application of data mining techniques to gain insights from the data extracted from students' performance and forecasting their academic results, hence, highlighting the importance and value of leveraging data for educational purposes.

4. DM Evolution and Techniques:

DM or data mining is an essential component of data analytics. It refers to digging into or mining large datasets in several ways to get insights and identify patterns and relationships. This is further supported by *Bootcamp.rutgers.edu* stating that ' Data mining goes beyond the search process, as it uses data to evaluate future probabilities and develop actionable analyses.'' Historically, the concept of DM existed even before computers did. The foundation for DM was set into motion via statistical methods, mainly Bayes' Theorem – a mathematical formula for determining conditional probability- that was established in 1763, and the development of regression analysis in the 1805. Afterwards, the advancements of technology, especially the Turing machine in the 1930s, neural networks in 1943, databases in the 1970s, and knowledge discovery in databases in the 1989 paved the path for our current comprehension of the nature of data mining nowadays. The 1990s and 2000s witnessed growth in computing power and data storage, offering more powerful and more prolific data mining applications in various situations. DM gains its popularization through the book ''Moneyball'' in 2003, demonstrating how a baseball team with limited resources, could establish a competitive team thanks to data analytics. Back then, the book challenged the traditional, intuition-based ways of making decisions. This, along with the growth of big data solutions in numerous fields of industry, has enlarged the significance and use of data mining.

There are two types of of data mining techniques that are commonly used in most of the previous studies; the descriptive and the predictive models. ''The descriptive analysis is used to mine data and specify the current data on past events.' Stated by *javatpoint.com* website.

As its name suggests, it describes the data, it allows us to learn from the past through answering ‘‘What has happened in the past?’ . Descriptive mining can provide correlation, frequency, and cross-tabulation -which is a method used to quantitatively analyse the relationship between multiple variables- illustrated by *humansofdata.atlan.com* website. These techniques, hence, are essential in determining the data regularities and reveal patterns. While, according to *javatpoint.com*, the predictive analysis ‘ gives the answers to all queries related to recent or previous data that move across using historical data as the main principle for decision.’’ It predicts the possible future events and trends.

5. Common DM methods in EDM:

In EDM research, the most common Data Mining (DM) methods utilized are classification and clustering, claim (Mohamed & Tasir, 2013). Classification is often utilized to assign data points to pre-determined categories and groups based on certain characteristics. Based on its features, classification permits an understanding of the category to which a new data point might belong to. Clustering on the other hand, sorts data into similar groups without any pre-defined categories. This permits discovering hidden patterns and potential new categories. *Javatpoint.com* defined the objective of classification as ‘‘to find which class a new object belongs to form the set of predefined classes.’’ and the objective of clustering as ‘‘to group a set of objects to find whether there is any relationship between them.’’

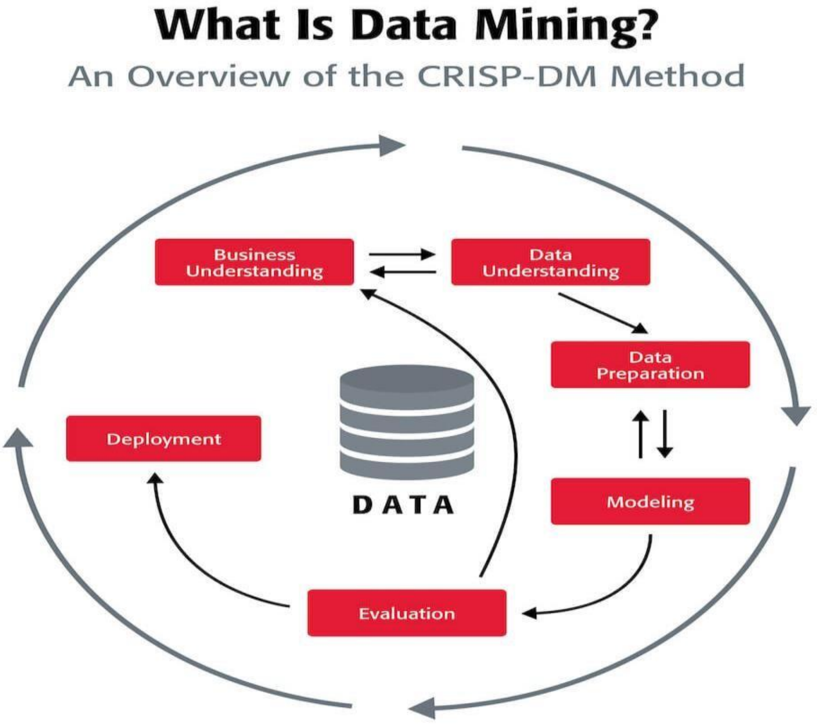
Understanding the relationship between the two categories; descriptive and predictive techniques and classification and clustering is indeed crucial. The descriptive and predictive techniques layout the basis and foundation for classification and clustering, as the techniques used in them help identify significant features and comprehend the data distribution before performing classification or clustering algorithms. However, these last algorithms can be utilized as descriptive or predictive tools; since they are able of describing existing data patterns and forecasting the likely category a specific data point might falls into.

6. The CRISP-DM Process:

Data mining typically follows a standardized six-step method known as CRISP-DM ‘ Cross-Industry Standard Process for Data Mining’’. As stated by *bootcamp.rutgers.edu* website, this process ‘‘encourages working in stages and repeating steps if necessary. In fact, repeating steps is often essential to account for changing data or to introduce different variables.’’ These processes are represented in the diagram below.

As demonstrated in the diagram, the first process or phase of the CRISP method is Business understanding. It involves defining the goal, objective, and problem we are trying to solve as well as the needed data to solve it. With careful consideration to the data; as choosing the wrong data can lead to misleading errors, irrelevant results, and/or results that do not answer our questions.

The second step is Data understanding. Once we set the objective, we need to collect the proper and necessary data. There are various sources of data, mainly records, surveys, geolocation data etc. Then, it is the step for data preparation, which is the most time-consuming process. It has three main stages known as ETL; Extract, in which we gather data from various sources. Transform, which is basically cleaning, fixing errors and dealing with missing values. And Load, this phase is for transferring the data we prepared into a usable database. The fourth process is Modeling. It requires choosing the appropriate statistical technique- such as the ones we discussed above: classification and clustering- to generate answers to the questions we asked before using the prepared data. It is worth mentioning that, as stated in the website *bootcamp.rutgers.edu/blog/, ‘‘*it’s also not uncommon to use different models on the same data to address specific objectives’’. After building and testing the models, it is time for evaluating their success in answering the questions asked during the first phase. In this step, experts must intervene to judge the efficacy of the models’ output, and adjust them accordingly. The last process in the CRISP method is Deployment. Once the data mining model accurately answers the questions, it is time to put the acquired knowledge and insights into practice. This can involve presentations, reports, or making decisions and taking direct actions.



Scheme: the processes of data mining

7. Data Mining and Machine Learning:

Data mining and Machine Learning are often used interchangeably regardless of the differences they have. To prevent confusing the two, we present some of the main distinctions between them. Data mining aims to reveal existing patterns and trends within data, it basically extracts the knowledge in order to gain insights. It also requires human's intervention to define the problem, determine the right techniques and interpret the results. An example of data mining could be analysing the history of purchase of certain customers to uncover trends in their buying habits. Machine learning from the other hand, ''focuses on developing algorithms and models that learn from data and make predictions or decisions.'' Shivam Arora's definition from *simplilearn.com* website. During its process, machine learning relies on algorithms that automatically learn from experience (data). Hence, it creates models able of predicting and making decisions based on new data. An example of machine learning could be training a software to differentiate between images of cats and dogs through learning from massive datasets of labeled pictures.

By definition, Machine Learning is ''an application of Artificial Intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.'' *Simplilearn.com.* To illustrate more, let us use this example to comprehend how machine learning works. Imagine we provide a system with the input data that carries images of different kinds of fruits, and we want the system to distinguish between the various kinds of fruits, and group them accordingly. The system, hence, analyses the given input data and attempt to find patterns -shape, colour, size- based on which the system will try to predict the various kinds of fruit and segregate them. Finally, it keeps track of all the decisions it made throughout the process for similar operations in the future. It is worth mentioning that there are three types of Machine Learning, the supervised, unsupervised, and reinforcement learning. As its name suggests, the supervised learning requires supervision while it is being trained to work on its own, it necessitates labeled training data. Unsupervised learning, from the other hand, requires training data but it is non-labeled. The third type is reinforcement learning in which the machine learns on its own via given feedback.

For further illustration of the differences between DM and ML, the following table is provided.

	Data Mining	**Machine Learning**
Focus	Discovery of hidden patterns or knowledge from data	Development of algorithms that learn from data

Goal	Extract insights and information from existing datasets	Build models to make predictions or perform tasks
Usage	Identifying patterns, trends, and anomalies	Predictive modeling, classification, clustering, etc.
Input	Historical data or large datasets	Labeled or unlabeled data for training and testing
Output	Knowledge in the form of patterns or rules	Predictions, classifications, recommendations, etc.
Methods	Descriptive statistics, clustering, association rules	Decision trees, regression, neural networks, SVM, etc.
Scope	Broader in terms of analysing various types of data	Focused on developing models for specific applications
Domain	Widely used in business, marketing, healthcare, etc.	Widely used in AI, robotics, pattern recognition, etc.

Table 1: Differences between Data Mining and Machine Learning

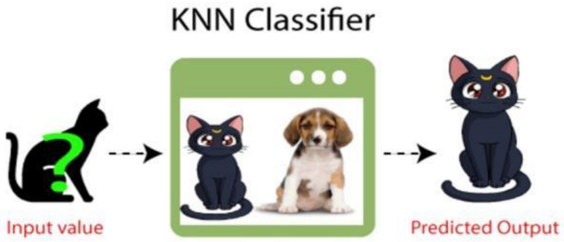
Before delving into the machine learning algorithms, it is essential to keep in mind that, as Shivam Arora clarifies, ‘ data mining is a process that incorporates two elements: the database and machine learning. The former provides data management techniques, while the latter supplies data analysis techniques.’

1. Machine Learning Algorithms: K-NN and Decision Tree:

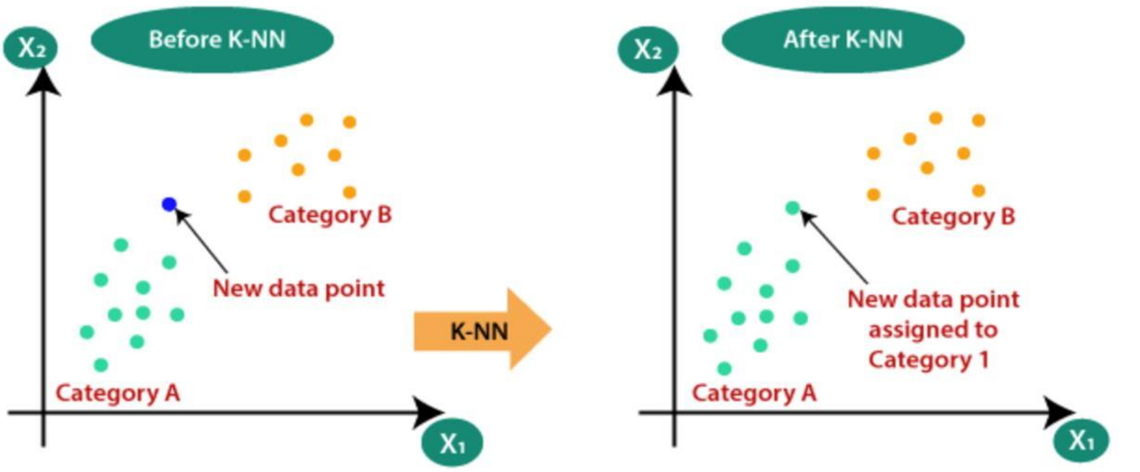
Choosing an algorithm over the other requires understanding what would work best for the data we have. Classification for instance, is a kind of methods which falls under supervised learning. It used when the output we are seeking is categorical; such as ‘yes’ or ‘no’, ‘true’ or ‘false’, ‘success’ or ‘failure’. etc. The algorithms that fall under classification are: K-nearest neighbour (KNN), Decision Tree, Random Forest, Naive Bayes, and Logistic Regression. Other methods are Regression and Clustering. Regression is used when the predicted data is numerical in nature, for example, if a shopkeeper wants to predict the price of a product based on its demand, the algorithm to use is Linear regression. Clustering, on the other hand, is a kind of unsupervised learning that is used when the

data needs to be organized to find patterns, for instance, this method is used in search engines to study the search history and figure out one’s preferences to provide the best search results.

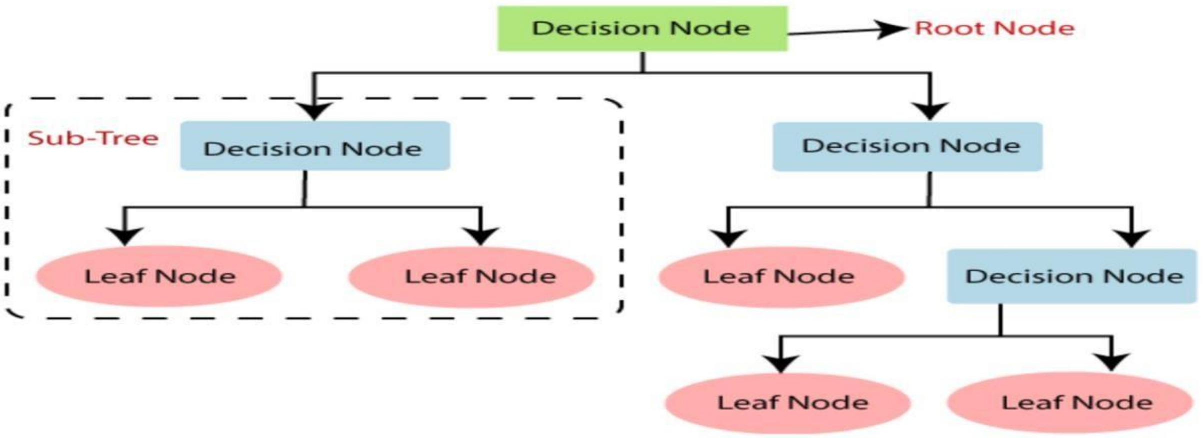
The algorithms to be reviewed in this section are KNN and Decision Tree. K-Nearest Neighbors is a simple but powerful machine learning algorithm used mostly for classification problems. It is often referred to as a lazy learner because, unlike other algorithms that perform complex calculations during training, KNN does not learn from the training set immediately, it rather stores all the available data and then classifies the new data point based on its similarity. That is to say, when a new data appears, it can be classified into the most appropriate category by means of K-NN algorithm. To further illustrate how K-NN works, consider the following example taken from *javatpoint.com* website: suppose we have a picture of an animal that looks like a cat and a dog. In order to classify it as definitively a cat or a dog, we can use K-NN since it relies on similarity measure. K-NN model will discover the similar features of the new data point to the images of the cats and the dogs, and classifies it as either cat or dog.



K-NN is mostly used in situations where we need to classify new data points. For instance, suppose there are two different categories (category A and category B), and we have a new data point x1, this new data point will lie in one of these categories A or B. K-NN helps in solving such problems. It enables us to easily identify the category of the new data point. Consider the following diagram:

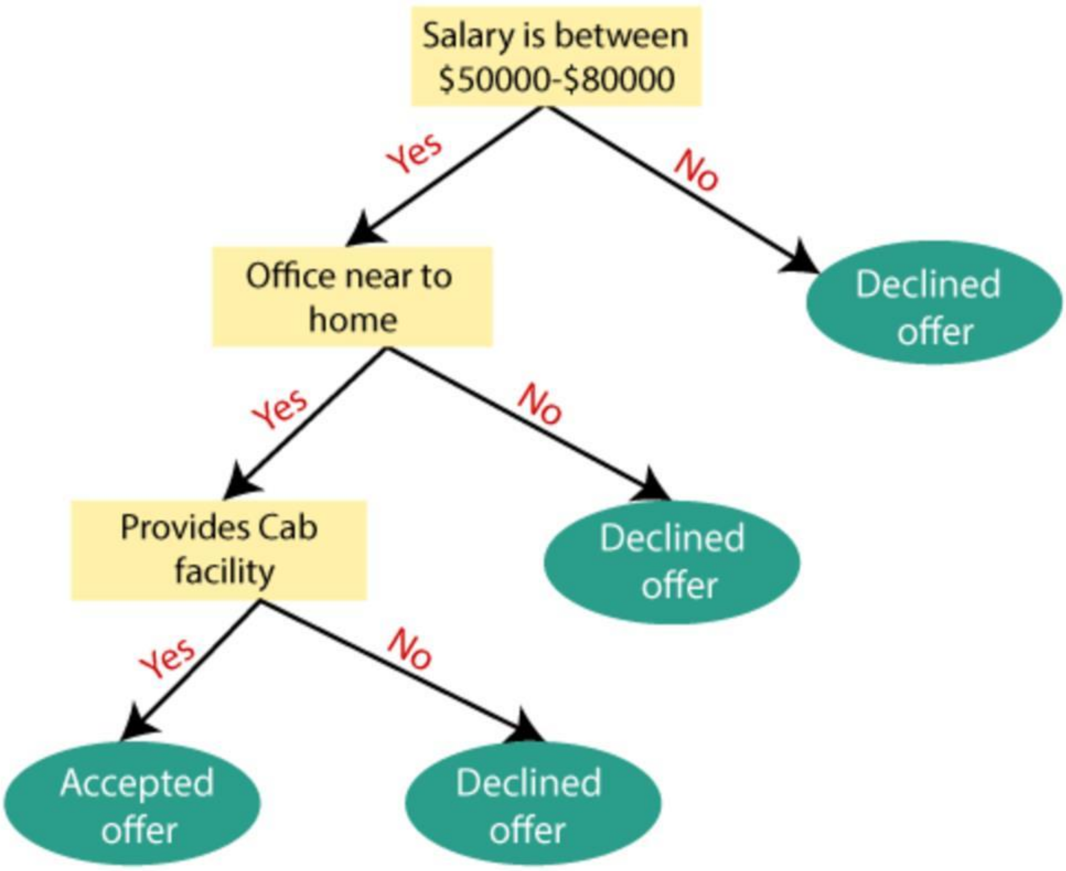


Another essential algorithm for classification is Decision Tree, it falls under the category of supervised learning. As stated by *javatpoint.com* website Decision Tree ‘ is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.’’ It is called Decision tree because it is similar to a tree; as it starts with the root node, further expands on other branches and constructs a tree-like structure. There are two nodes in a decision tree, the decision node and the leaf node. Consider decision nodes as asking questions for the aim of splitting the data set; they can have various branches., while leaf nodes are simply the output of those decisions and do not contain any other branches. The following is a diagram of a Decision Tree:



As for the terminology, the diagram above contains several nodes which we shall see their meanings. In brief, Root Node is basically the starting point of the decision tree. It encompasses the entire dataset. Leaf Node on the other hand refers to the final outcome or classification gained after segregation, and it the tree cannot be segregated further. Finally, Splitting is the process of dividing the root node or decision node into more sub-nodes according to the given criteria and condition.

To comprehend how Decision Tree algorithm works, let us consider this example: you are a job candidate with an offer and want to decide whether to accept the offer or decline it. You can simply use the decision tree, which will start with a question about; let us say salary as a root node, then you might accept or decline the offer based on the salary whether it is high or low. If you accept it then it is a leaf node. However, if you don't like the salary, you might want to ask other questions before jumping to the final decision. For example you might consider the distance from the office and cab facility. The root node in this case splits further into a decision node, which in turn splits into another decision node and one leaf node, and the process continues until you reach your final decision. Consider the diagram below for further comprehension:



8. K-NN and Decision Tree in EDM:

Most of the conducted studies in EDM use at least one of these algorithms: K-NN and Decision tree, if not both. Simply because they gave the best performance with high accuracy. Bigdoli et al (2003) claimed that ‘ K-Nearest Neighbor method had taken less time to identify the students performance as a slow learner, average learner, good learner and excellent learner.’’ Moreover, ‘ K-NN has been proved to give good accuracy in estimating the detailed pattern for learner’s progression in tertiary education’’ said Gray et al. (2014). Thanks to its clarity, simplicity, and comprehensibility to discover data patterns and forecast outcomes, Decision Tree has become a popular technique for prediction purposes. According to Romero et al. (2008) ‘ decision tree models are easily understood because of their reasoning process and can be directly converted into set of IF-THEN rules.’’ Decision Tree was and is still used as a method to evaluate students’ performance; as many studies confirm so. For instance, the studies conducted by the following researchers: Romero et al. (2008), Bunkar et al. (2012), Ramesh et al. (2013), Natek and Zwilling (2014), Mishra et al. (2014), and Jishan et al. (2015), all rely on decision tree to predict students’ performance, dropout, a suitable career based on their behavioural patterns, final grades and so on.

9. Previous research on predicting students’ performance using ML Algorithms:

The integration of machine learning algorithms in education is an emerging field of interest. It enhances the predictive analytics which, in turn, positively affect the decision making process of educators and stakeholders. Dervenis et al conducted a study in which they use date from two Portuguese schools to classify and predict students grades. The data, including academic, social and demographic variables, was taken from UCI repository. The researchers designed their model using the Orange Platform- an open source machine learning and data visualization platform- and utilized three machine learning algorithms; K-NN, SVM (Support Vector Machine), and Random Forest to predict students’ performance in terms of a binary classification; pass or fail. The models were evaluated using metrics of Accuracy, Precision, Recall, and F1 Score. The results highlighted the potential of ML algorithms in forecasting students’ performance based on several academic and socio-economic variables. This study reveals the promising role of ML algorithms in educational settings.

In the context of Morocco, Qazdar et al, conducted a comprehensive study in which they use data from H.E.K high school in Morocco to forecast students’ performance. They collected data from 478 Physics students during the school years: 2015-2016, 2016-2017, 2017-2018. the dataset contained multiple academic factors collected from ‘Massar’ a platform used in Moroccan educational institutions to manage students information. The authors applied different machine learning algorithms, including Random Forest, Naive Bayes, and Decision Trees to build predictive models. To determine the models effective4ness and accuracy, they were assessed based on the metrics of precision, recall, and F1 Score. The findings underline the efficacy of ML algorithms to predict students' performance.

Conclusion:

In this literature review, we investigated the evolving realm of machine learning and educational data mining, focusing on two main algorithms; K-NN and Decision Tree. We clarify the variation between DM and ML, paving the path for our analysis of K-NN and Decision Tree algorithms. While research on data mining and machine learning tend to flourish globally, it remains limited in Morocco, and especially in the field of applied linguistics, which in turn, creates a critical gap and requires research and intervention.

Chapter 2

Chapter 2: Methodology

Introduction:

This section aims to discuss and describe the methodological approach opted for this experimental study seeking to predict students' performance using machine learning algorithms. It is structured into three sub-sections. The first provides a description of the research design adopted, while the second is dedicated to the description of the dataset, and last but not least a sub-section devoted to the analytical phase.

Research design:

The design of this research is experimental in nature as the aim is to utilize machine learning techniques to forecast the performance of students. The community of researchers in EDM often utilizes ''session logs and databases for processing and analyzing student performance prediction using a machine learning algorithm.'' said Albreiki, B et al. (2021). That is to say, students' activity data and past performance information are used to build programs that can predict how well students will perform in the future. Fortunately, such programs are already built and made accessible for us, however, there is need to adjust them to our specific purposes, data and context. In this case, our purpose is to classify students based on success or failure, the data for training the model is taken from Kaggle- an excellent platform that provides trusted data- and the variables upon which the model is trained are suitable for the context of Moroccan education and lifestyle in general. These are further clarified in the upcoming sections.

In this research, two publicly available datasets-from Kaggle- were used to predict students' performance. The two datasets were collected in secondary education in two public schools from 395 students. The dataset attributes contains grades, social backgrounds, demographics, school information and the like. All data were obtained from school reports and student surveys. The datasets have 33 data points for each student, which are detailed further in the table below.

Feature	Description	Type	Values	Descriptive statistic

School	Student's school	Binary	GP/MS	GP= 88%
				MS=12%

Sex	Student' gender	Binary	Female or male	F= 53%
				M=47%

Age	Student' age	Numeric	From 15 to 22 years old	Median= 15y old

Address	Student' home address	Binary	Urban or rural	U=78%
				R=22%

Pstatus	Parents' cohabitation status	Binary	T= living together
			A= apart
				T= 90%
				A=10%

Medu	Mother's education	Numeric	From 0 to 4	-

Mjob	Mother's occupation	Nominal	Teacher, health, services, at home, other	Other=36%
				services=26%

Fedu	Father's education	numeric	From 0 to 4	-

Fjob	Father's occupation	Nominal	Teacher, health, services, at home, other	Other=55%
				services=28%

Guardian	Student's guardian	Nominal	Mother, father, other	Mother=69%
				father=23%

Famsize	Family size	Binary	LE3 (|less than or equal 3) or GT3 (greater than 3)	GT3=71%
				LE3=29%

Famrel	Quality of family relationships	Nominal	From 1 very bad to 5 excellent	1=08 2=18 3=68 4=195 5=106

Reason	Reason for choosing this school	Nominal	Near home, school reputation, course preference, other	Course= 37%
				home=28%

Traveltime	Travel time from home to school	Nominal	1< 15min/ 2 15-30min, 3 30-1h, 4 >1h	1=256 2=107 3=23 4=08
Studytime	Weekly study time	Nominal	1 <2h, 2 2 to 5h/ 3 5 to 10h/ 4> 10h	-
Failures	Number of past class failures	Nominal	n if 1 <= n <3, otherwise 4	0= 312 1=50 2=70 3=16
Schoolsup	After-school support	Binary	Yes or no	-
Famsup	Educational support from family	Binary	Yes or no	-
Activities	Extracurricular activities	Binary	Yes or no	-
Paidclass	Paid classed	Binary	Yes or no	-
Internet	Internet access at home	Binary	Yes or no	-

Nursery	Attended nursery school	Binary	Yes or no	-
Higher	Pursue higher education	Binary	Yes or no	-
Romantic	Has a romantic relationship	Binary	Yes or no	-
Freetime	Free time after school	Nominal	From 1 (very bad) to 5 (excellent)	-
Goout	Going out with friends	Nominal	From 1 (very bad) to 5 (excellent)	-
Walc	Weekend alcohol/drugs consumption	Nominal	From 1 (very bad) to 5 (excellent)	-
Dalc	Workday alcohol/drugs consumption	Nominal	From 1 (very bad) to 5 (excellent)	-
Health	Current health status	Nominal	From 1 (very bad) to 5 (excellent)	-
absences	Number of school absences	Numeric	From 0 to 93	-
G1	Grade of first period	Numeric	From 0 to 20	-
G2	Grade of second period	Numeric	From 0 to 20	-
G3	Final grade	Numeric	From 0 to 20	-

Table 2: main characteristics of the dataset.

Some statistical insights:

Age: the average age is 16 years old.

Absence: the average days of absence is 6, however, we observe outliers with 75 days of absence.

Dalc: the daily alcohol/drugs consumption among students is very low, fortunately.

After data acquisition, cleaning and transforming, we classified the attributes based on the measurement scale to which they belong. The following table demonstrates this in detail.

Categorical variables (ordinal)	Categorical variables (non-ordinal)	Categorical variables with 2 values	Numeric variables
-Medu -Fedu - Travel time - Study time - Failures - Famrel - Freetime - Goout - Dlac - Walc - Health	- Mjob - Fjob - Reason - Guardian -	- School - Sex - Address - Famsize - Pstatus - Schoolsup -Famsup - Paidc - Activities - Nursery - Higher - Internet - Romantic	- Age* - Absence* - G1 - G2 - G3

Table 3: classification of the data.

Preprocessing:

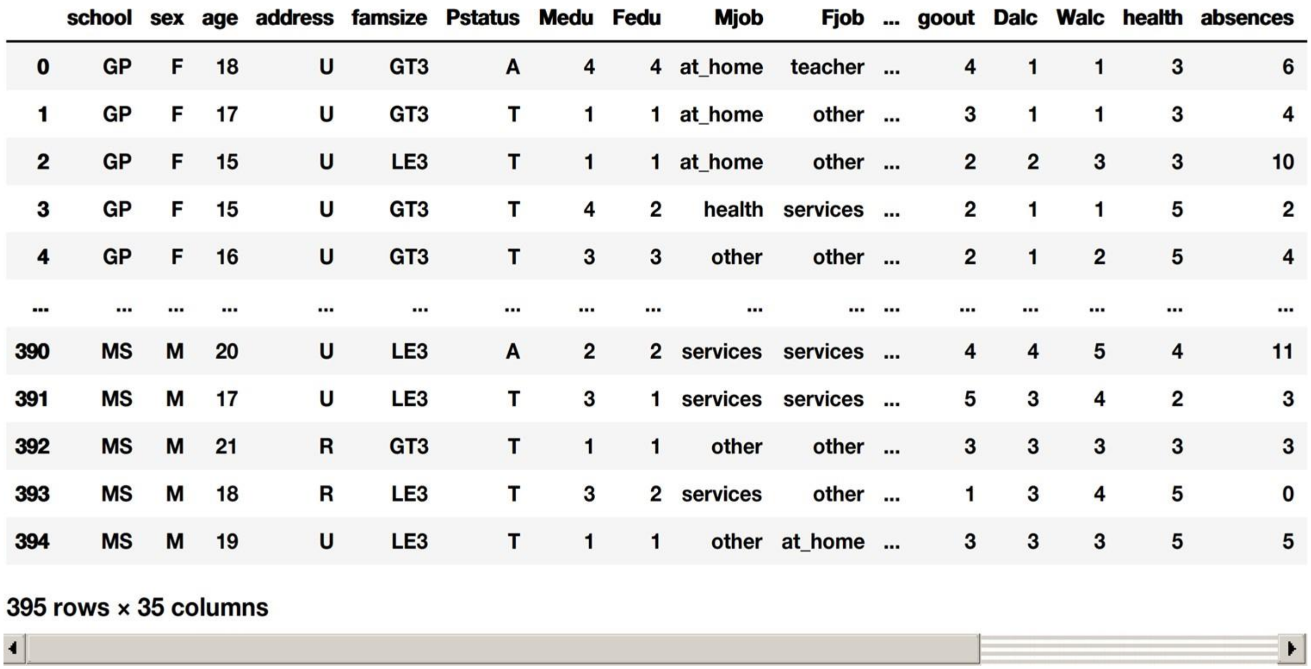
Encoding categorical variables:

Machine learning models typically cannot handle categorical variables directly, except some specific models of course. Thus, we need to convert these categorical values into numerical representations (in form of numbers) that the model can understand. This process is called encoding.

Categorical variables can be ordinal, binary, or nominal. The ordinal variables have categories with an inherent order. For instance, the quality of family relationships, the amount of free time, going out, and health status are all ordinal categories. Binary variables from the other hand have only two categories, such as ‘ yes’’ and ‘ no’’. School support, family support, paid classes, activities, internet, romantic relationship all fall under the category of binary variables. The third type is the nominal, which basically have distinct categories that do not imply an order. Examples include gender, address, school, parents status, mother or father’s job, family size etc.

While encoding the data, each category of the ordinal variables is assigned a numerical value which corresponds to its order. For instance: family relationship values were mapped to 1-5, respectively. For binary variables, the values 0 for ‘ no’’ and 1 for ‘ yes’’ were used. While for nominal variables, One-hot encoding method was used, which is a common method when dealing with categorical data in machine learning. It is basically the conversion of categorical data into a format machine learning algorithms can comprehend through creating a new binary feature for each category. For instance, sex with the categories M (male) and F (female) is converted into two new features: sex-M and sex-F, each observation is assigned a value of 1 or 0 based on the feature to which it corresponds. It is worth mentioning that age and absence were initially numerical variables (e.g., 18years old/ 7 absences) but were discretized into categories ‘‘bins’’ (e.g., 15-17 years/ 0-5 days or low absence, medium absence etc. Thanks to this technique, we can prevent overfitting which occurs when a model memorizes specific details of the training data and performs poorly on new data. For instance, a model can overfit if it learns that a student with a number of 20 absences is likely to fail, and performs poorly when it encounters a student with 19 or 21 absences.

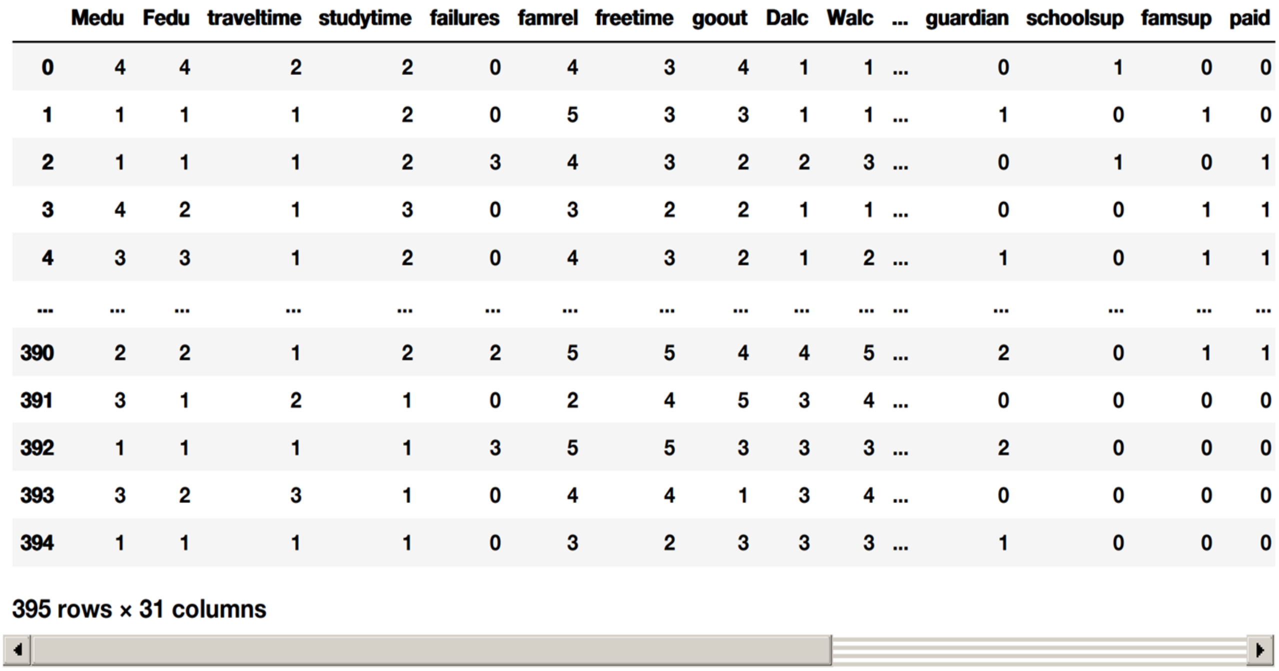
After cleaning the data, a practical step involved its attributes as shown in the table below:



	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	goout	Dalc	Walc	health	absences
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...	4	1	1	3	6
1	GP	F	17	U	GT3	T	1	1	at_home	other	...	3	1	1	3	4
2	GP	F	15	U	LE3	T	1	1	at_home	other	...	2	2	3	3	10
3	GP	F	15	U	GT3	T	4	2	health	services	...	2	1	1	5	2
4	GP	F	16	U	GT3	T	3	3	other	other	...	2	1	2	5	4
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
390	MS	M	20	U	LE3	A	2	2	services	services	...	4	4	5	4	11
391	MS	M	17	U	LE3	T	3	1	services	services	...	5	3	4	2	3
392	MS	M	21	R	GT3	T	1	1	other	other	...	3	3	3	3	3
393	MS	M	18	R	LE3	T	3	2	services	other	...	1	3	4	5	0
394	MS	M	19	U	LE3	T	1	1	other	at_home	...	3	3	3	5	5

395 rows × 35 columns

The table above shows the values as answered by the students. However, as was mentioned that the data should be encoded for the model to interpret it, we encoded it to numerical values as demonstrated in the table below:



	Medu	Fedu	traveltime	studytime	failures	famrel	freetime	goout	Dalc	Walc	...	guardian	schoolsup	famsup	paid
0	4	4	2	2	0	4	3	4	1	1	...	0	1	0	0
1	1	1	1	2	0	5	3	3	1	1	...	1	0	1	0
2	1	1	1	2	3	4	3	2	2	3	...	0	1	0	1
3	4	2	1	3	0	3	2	2	1	1	...	0	0	1	1
4	3	3	1	2	0	4	3	2	1	2	...	1	0	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
390	2	2	1	2	2	5	5	4	4	5	...	2	0	1	1
391	3	1	2	1	0	2	4	5	3	4	...	0	0	0	0
392	1	1	1	1	3	5	5	3	3	3	...	2	0	0	0
393	3	2	3	1	0	4	4	1	3	4	...	0	0	0	0
394	1	1	1	1	0	3	2	3	3	3	...	1	0	0	0

395 rows × 31 columns

NB: this dataset is only used to train the model. The data we wanted to check for failure or success will be used later in an application/website that we developed.

Let us now investigate how the models deal with the given data: KNN and Decision Tree.

K-Nearest Neighbors: as a machine learning algorithm, K-nn can predict students' performance through comparing a new student's data points, the model should be trained on the same data points before. For instance, if a model is trained on variables such as age, sex, absence, and health only, the new student' data should not include new variables that the model is not familiar with. The algorithm then selects the 'K' closest neighbours, and predicts the new student's performance based on the performance of those neighbours it selected before.

Decision trees CART: is a binary tree that recursively splits the dataset until what is left is pure leaf nodes- which is a data with only one type of class. Each node represents a decision point where data is divided, while the leaf nodes represent the final outcomes or predictions. The depth of the tree signifies the length from the root to the deepest leaf.

Metrics of Evaluation:

After selecting the models, we need to ensure their validity, effectiveness and generalizability. To do so, we employ various performance evaluation criteria or metrics to measure the success of the models. Metrics refer to measures used to evaluate the performance of models on specific task. The metrics we utilized are precision, recall, accuracy and AUC-ROC as they are the essential metrics for KNN and Decision tree.

Precision is used to measure the accuracy of positive predictions made by our models, it is calculated as the ratio of true positives to the sum of true positives and false positives as shown below. Precision demonstrates how many instances are actually positive of all the ones classified as positive.

Precision = True positives / (True positives + False positives)

Recall on the other hand measures the ability of a certain model to identify all the true instances. That is to say, how many instances are correctly classified of all the actual positive ones. It is calculated as the ratio of true positives to the sum of true positives and false negatives or simply as:

Recall = True positives / (True positives + False negatives)

In addition to recall and precision, accuracy is another measure for the overall correctness of the predictions made by the models. It is calculates how many instances are correctly classified of all the instances. It is calculated as the ratio of the sum of true positives and true negatives to the total number of instances or as shown below:

Accuracy = (True positives + True negatives) / (TP + TN + FP + FN)

Accuracy is an excellent measure for balanced data. However, and since our data is imbalanced, we needed to utilize balanced accuracy to ensure that one class does not outweigh the other and that the evaluation of the models is fair across all classes. The balanced accuracy is calculated as the following: Balanced accuracy = 1 / 2 (TP/P + TN/N)

In order to measure the classification models' ability to distinguish between classes we used AUC-ROC metric. Aniruddha Bhandari, a Data Engineer, defines AUC-ROC as ‘‘the measure of the ability of a binary classifier to distinguish between classes…the higher the AUC, the better the model's performance at distinguishing between positive and negative classes.'

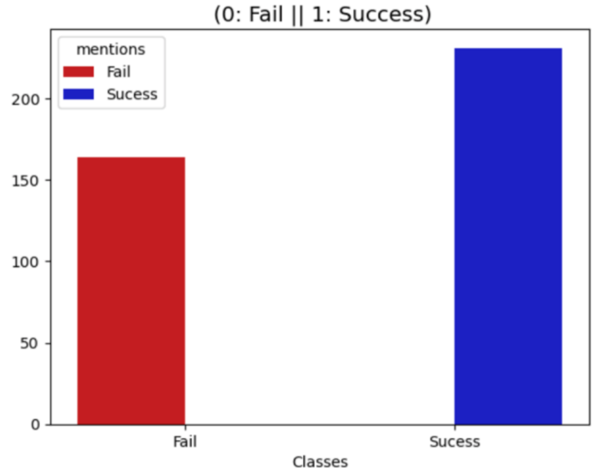
Experimental process:

As commonly used in some countries, the final grade ranges from 0 to 20 in the raw data. 0 represents the lowest grade while 20 is the highest, and since the students' final grades are integers, and the predicted class have to be in categorical forms, we transformed the data into categories which reflect the grading policy. That is to say, we categorized the final grade as pass if it is equal or more than 10, or as fail if it is less than 10.

Pass	**Fail**
= or > 10	< 10

Training a model on an imbalanced dataset is challenging. Imbalanced data is indeed a classification problem in which the number of observations per class is not equally distributed. In our dataset, the majority of the class is ''Success'' with a percentage of 58%, while the rest of the class is ''Fail'' which is a 42% as demonstrated in the graph below.

Number
Of
Occurances



Distribution of classes

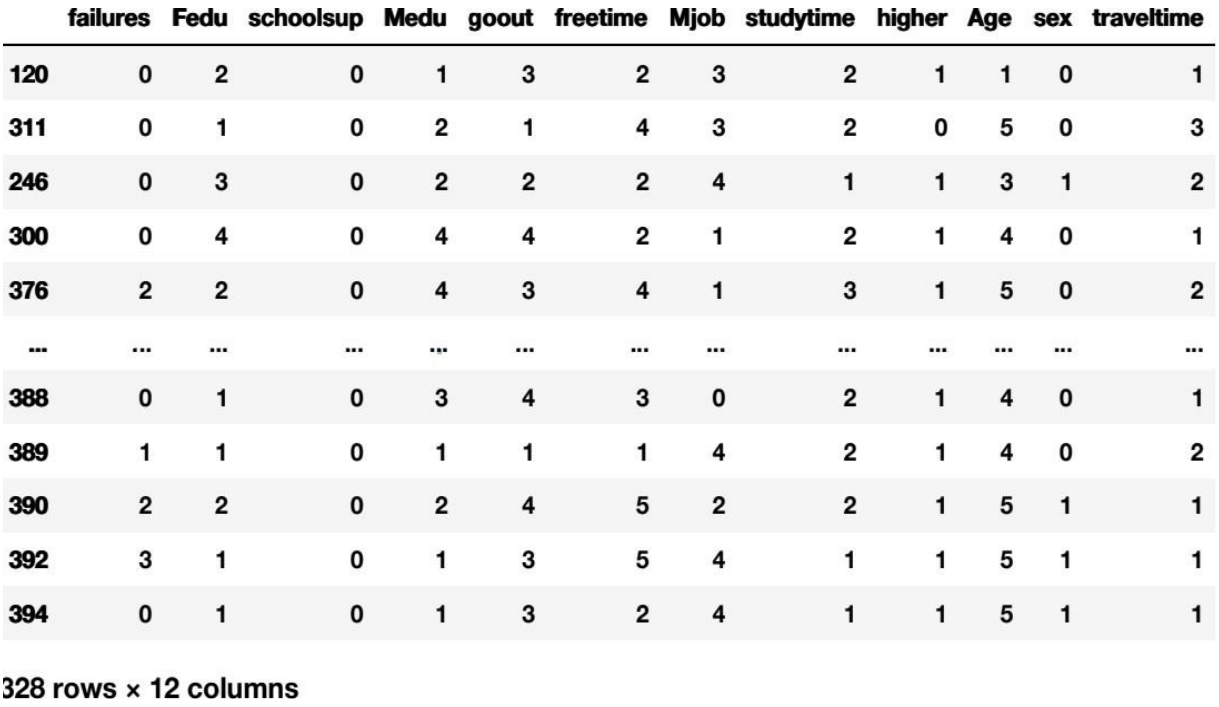
Success: 231

Fail: 164

Total: 395

We addressed the imbalanced data issue by SMOTE and Nearmiss techniques. SMOTE is an oversampling technique which stands for Synthetic Minority Over-sampling Technique. It basically generates new observations or data points for the minority through creating instances that lie between existing samples of the minority class. This helps balance the class distribution in datasets in which one class is the minority. On the other hand, Nearmiss, which is an undersampling technique, creates balance by selecting observations from the majority class that are close to the minority class and consider them as minority.

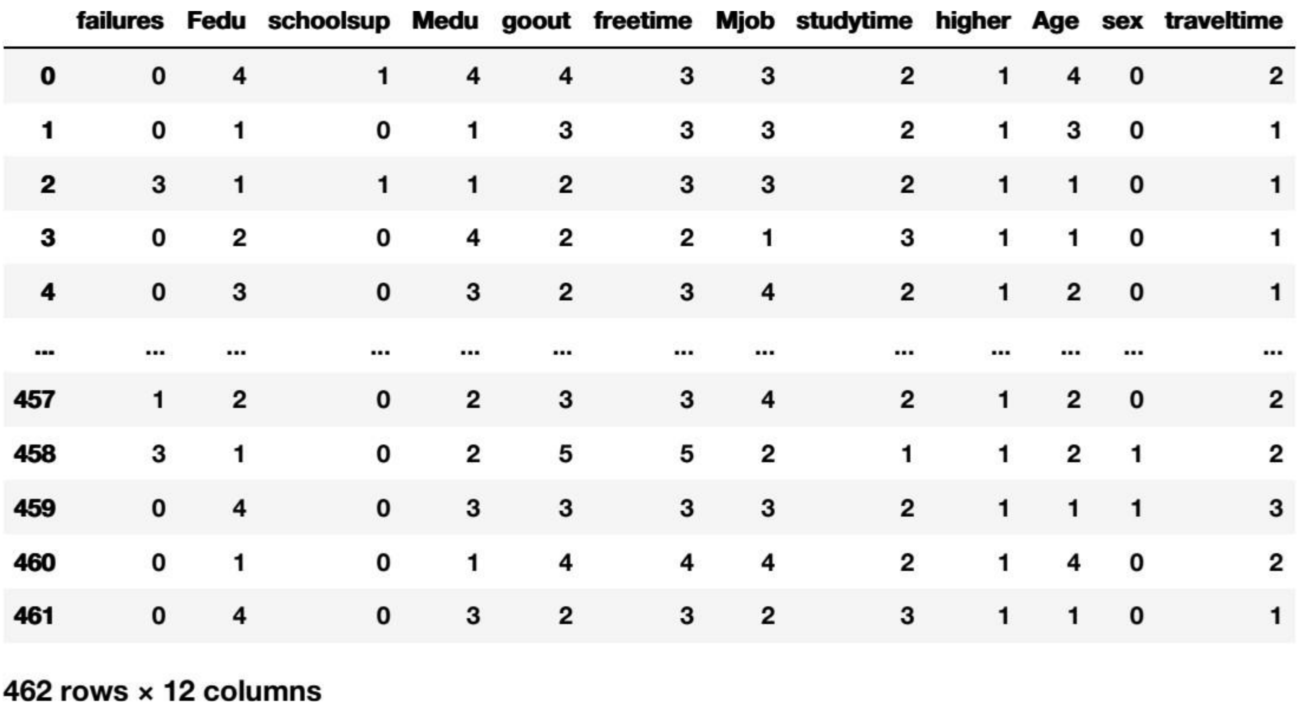
The application of SMOTE and Nearmiss techniques lead to two new datasets:
We obtained a total of 462 data points with SMOTE over-sampling, and 328 with Nearmiss. The tables below are demonstrations of the datasets we obtained:



	failures	Fedu	schoolsup	Medu	goout	freetime	Mjob	studytime	higher	Age	sex	traveltime
120	0	2	0	1	3	2	3	2	1	1	0	1
311	0	1	0	2	1	4	3	2	0	5	0	3
246	0	3	0	2	2	2	4	1	1	3	1	2
300	0	4	0	4	4	2	1	2	1	4	0	1
376	2	2	0	4	3	4	1	3	1	5	0	2
...	...	...	...	...	...	...	...	...	...	...	...	...
388	0	1	0	3	4	3	0	2	1	4	0	1
389	1	1	0	1	1	1	4	2	1	4	0	2
390	2	2	0	2	4	5	2	2	1	5	1	1
392	3	1	0	1	3	5	4	1	1	5	1	1
394	0	1	0	1	3	2	4	1	1	5	1	1

328 rows × 12 columns

New dataset with Nearmiss technique of undersampling



	failures	Fedu	schoolsup	Medu	goout	freetime	Mjob	studytime	higher	Age	sex	traveltime
0	0	4	1	4	4	3	3	2	1	4	0	2
1	0	1	0	1	3	3	3	2	1	3	0	1
2	3	1	1	1	2	3	3	2	1	1	0	1
3	0	2	0	4	2	2	1	3	1	1	0	1
4	0	3	0	3	2	3	4	2	1	2	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...
457	1	2	0	2	3	3	4	2	1	2	0	2
458	3	1	0	2	5	5	2	1	1	2	1	2
459	0	4	0	3	3	3	3	2	1	1	1	3
460	0	1	0	1	4	4	4	2	1	4	0	2
461	0	4	0	3	2	3	2	3	1	1	0	1

462 rows × 12 columns

Dataset with Smote technique of over-sampling

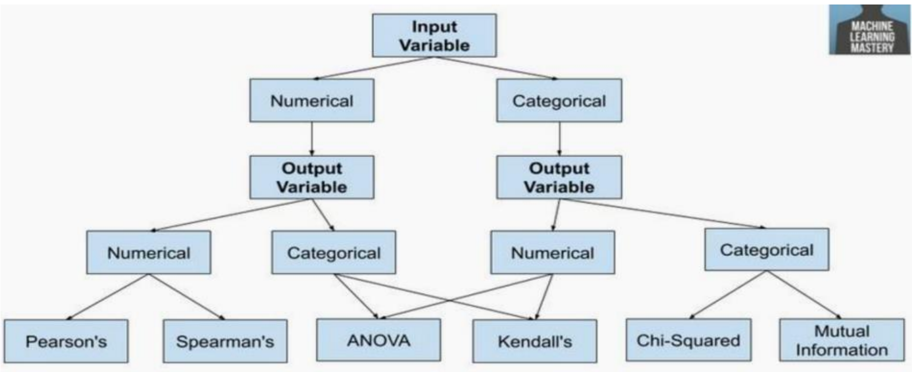
At the end of this stage, we obtained three datasets:

- Raw dataset
- Over-sampling dataset
- Undersampling dataset

and from these datasets, we will apply variable selection.

Variable/ feature selection:

Variable or feature selection is a main concept in machine learning that has crucial impact on the performance of the models. It is basically the selection of the variables that contribute the most to our data. As stated by Trevor LaViale, an ML Solutions Engineer: ‘ The aim of feature selection is to remove irrelevant and redundant features, thereby improving the performance of the algorithm.’’



Selecting variable selection methods

The figure above presents several methods of variable selection. We are only concerned with the categorical input and output. The categorical input refers to variables representing categories or groups such as school type, gender etc. while categorical output is basically the predicted category or class such as, in our case, pass or fail.

As demonstrated in the figure above, Chi-Squared and Mutual Information are the selection methods that best suit our case. The purpose of Chi-Squared is to determine whether there is a significant relation between categorical input features and categorical output. Mutual information on the other hand is used to measure the mutual dependence between input and output variables. It helps identify the most relevant categorical input variables that affect the categorical output.

In addition to Chi-Square and Mutual Information, we use Variance Threshold method to remove all features whose variables does not meet certain threshold. According to Aman Gupta ‘‘variance threshold, by default, removes all zero-variance features, i.e., features with the same value in all samples. We assume that features with a higher variance may contain more useful information.’

The three variable selection methods were applied to the three datasets we obtained before (the raw or brute, undersampling, and over-sampling datasets). After the selection process, each method retained specific percentages of features (20%, 40%, 60%). The table below is a demonstrations of the datasets obtained, including different combinations of variable selection methods along with percentages of the retained features. By the end of the preprocessing phase, we obtained a total of 27 datasets, each one of them represents a distinct subset of selected variables on the methods that were applied and the percentages.

Chi-Square test filter	Mutual Information filter	Variance Threshold filter
rawChi2-20	rawMI-20	rawVAR-20
rawChi2-30	rawMI-30	rawVAR-30
rawChi2-40	rawMI-40	rawVAR-40
overChi2-20	overMI-20	overVAR-20
overChi2-30	overMI-30	overVAR-30
overChi2-40	overMI-40	overVAR-40
underChi20-20	underMI-20	underVAR-20
underChi20-30	underMI-30	underVAR-30

Datasets obtained afterwards

Training of Classification Models:

To evaluate the models (K-NN, Decision Tree) we use the GridSearch algorithm, which is a technique used to find the best set of hyperparameters for a machine learning model. GridSearch evaluates the models based on accuracy, balanced accuracy, precision and AUC-ROC, while including cross-validation, which is also a technique used in machine learning to ensure the reliability and generalizability of the model, and helps in detecting overfitting issues where the model actually performs well on the training data but poorly on unseen data.

The following represent the parameters used for each algorithm:

K-NN:

N-neighbours: number of neighbours to consider. In our case, it ranges from 2 to 11.

Weights: how neighbours are weighted during prediction: uniform (all neighbours are given equal weight) distance (weighted by the inverse of their distance. I.e, nearby points are more likely to belong to the same class)

Metric: hamming (a metric used to compare the similarity or dissimilarity between two binary or categorical data strings)

Decision Tree:

Criterion: gini (simply put, it measures how a good a Decision Tree is at splitting the data)

Splitter: best (a technique to select the best split based on gini criterion to maximize the information gain) random (randomly selects features and select the best random split among them)

Max_depth: specifies the maximum depth allowed for the tree. It ranges from 2 to 5.

Min_samples_leaf: determines the minimum number of samples allowed to be at a leaf node. It ranges from 5 to 15.

Max_features: sets the number of features required when looking for the best split at each node. It ranges from 4 to 8.

Evaluating the Classification Models:

In this study, we utilized Python’s Jupyter Notebook for variable selection and data classification. We used cross-validation method to assess how well our classification models were. This method basically splits the data into 10 equal parts, train the model on 9 parts and then tests it on the rest. Each of the 10 parts is used as a test. This process is repeated 10 times, each part is used once as a validation set.

The table below includes the performance of different K-NN models based on their hyper-parameters (number of neighbours, and weights):

Dataset	N_Neighbors’	Weights	Accuracy %
RAWCHI2_20	3	UNIFORM	61,1987
RAWCHI2_30	10	DISTANCE	60,3885
RAWCHI2_40	7	DISTANCE	59,6248
OVERCHI2_20	6	DISTANCE	68,173
OVERCHI2_30	7	UNIFORM	63,6309
OVERCHI2_40	8	DISTANCE	68,617
UNDERCHI2_20	9	UNIFORM	62,7936
UNDERCHI2_30	9	UNIFORM	62,2254
UNDERCHI2_40	4	DISTANCE	60,0758
BRUTMI_20	3	DISTANCE	61,1987
BRUTMI_30	10	DISTANCE	60,3885
BRUTMI_40	3	DISTANCE	58,934
OVERMI_20	6	DISTANCE	68,173
OVERMI_30	7	UNIFORM	63,6309
OVERMI_40	6	DISTANCE	67,7567
UNDERMI_20	9	UNIFORM	62,7936

Table 5: The best models based on their hyper-parameters and accuracy for KNN.

The table above contains a set of the best data that achieved higher percentages of accuracy. The dataset ‘overCHI2_40’ attained the highest accuracy of 68.61% using 8 neighbours and distance weighting. Similarly, ‘overCHI2_20’ and ‘overMI_20’ both achieved a high accuracy of 68.17% with 6 neighbours and distance weighting. The varying accuracy results across the different hyper-parameters suggests a need for optimization.

The table below includes the performance of different CART models based on their hyper-parameters (max-depth, max-features, min-samples leaf, and split):

Dataset	Max-depth	Max-features	Min_Samples-leaf	Split	Accuracy %
RAWCHI2_20	3	5	5	random	67,9711
RAWCHI2_30	3	5	5	random	67,9711
RAWCHI2_40	3	7	5	best	68,3214
OVERCHI2_20	4	4	5	best	66,6466
OVERCHI2_30	3	4	5	best	66,6605
OVERCHI2_40	4	6	9	best	66,6512
UNDERCHI2_20	2	5	5	best	68,9015
UNDERCHI2_30	3	5	5	random	68,9015
UNDERCHI2_40	2	5	7	best	68,9015
RAWMI_20	3	5	5	random	67,9711
RAWMI_30	3	5	5	random	67,9711
RAWMI_40	3	7	9	best	68,3214
OVERMI_20	4	4	5	best	66,6466
OVERMI_30	3	4	5	best	66,6605

Table 6: The best models based on their hyper-parameters and accuracy for Decision Tree Cart

The table above reveals the performance of various CART models considering their hyper-parameters such as max-depth, max-features, min-samples-leaf, and split criterion. As demonstrated, the dataset ‘underCHI2_20’ reached the highest accuracy at 68.90% using a max-depth of 2, max-features of 5, min-samples of leaf of 5, and ‘best’ as a splitting criterion. Regardless of a slight difference in their hyper-parameters, both ‘underCHI2_30’ and ‘underCHI2_40’ reached a percentage of 68.90%.

Chapter 3

Chapter 3: Results and Discussion

Findings:

This section is dedicated to presenting the results obtained with the two models: KNN and Decision Tree Cart. As far as KNN model is concerned, five best datasets were selected based on three metrics: accuracy, precision and AUC-ROC, and according to their best parameters. The following tables are representations of the datasets along with their metrics and hyperparameters:

- NB: the datasets in the tables below are ranked based on their level of accuracy, precision, and AUC_ROC from the highest percentage to the lowest.

KNN

Dataset	n_neighbours	weights	accuracy	Accuracy %
overVAR_40	2	distance	0.699167	69.9167
overCHI2_40	8	distance	0.68617	68.617
overMI_20	6	distance	0.68173	68.173
overCHI2_20	6	distance	0.68173	68.173
overMI_40	6	distance	0.677567	67.7567

Table 7: Best Accuracy for KNN

As demonstrated in table 7, the ‘overVAR_40’ dataset achieved the highest accuracy (69.92%) when utilizing 2 neighbours and distance weighting. While the other datasets, ‘overCHI2_40’ and ‘overMI_20’ using the same distance for weight and different number of neighbours (8 and 6) reveals slightly lower accuracy with a percentage of 68.61 %. The use of distance weighting across all top datasets emphasizes its crucial role in enhancing the model’s performance through providing more influence and weight to the closer neighbours.

Datasets	n_neighbours	weights	precision	Precision%
overVAR_40	2	uniform	0.718281	71.8281
underCHI2_20	4	uniform	0.697497	69.7497
underMI_20	4	uniform	0.697497	69.7497
overCHI2_20	6	distance	0.68173	68.173
overCHI2_40	8	distance	0.68617	68.617

Table 8: Best Precision for KNN

Similarly to table 7, the dataset ‘overVAR_40’ scores the highest precision at 71.83% using 2 neighbours and uniform weighting. Unlike accuracy, the best hyper-parameter for weight is uniform as it contributes to achieving more accurate positive predictions. The datasets ‘overCHI2_20’ and ‘OVERchi2_40’ achieved lower percentages of precision (68.17 and 68.61) due to the selection of large number of neighbours and distance weighting.

Dataset	n_neighbours	weights	sensitivity	AUC_ROC%
overVAR_40	2	distance	0.723981	72.3981
overCHI2_40	8	distance	0.68617	68.617
overChi2_20	6	distance	0.68173	68.173
overMI_20	6	distance	0.68173	68.173
overVAR_30	4	distance	0.679579	67.9579

Table9: Best AUC-ROC for KNN

Correspondingly, the dataset ‘overVAR_40’ achieved the best score of 72.40%. AUC_ROC (sensitivity) suggests that the model has a good discrimination ability to distinguish between the positive and negative classes. The hyper-parameters of ‘overVAR_40’; a number of neighbours that does not exceed two, and distance weighting are the optimal setup for our data to get the best predictions.

Analysis and Discussion:

K-Nearest Neighbors:

overVAR_40:

After analysing the results obtained from KNN model across five different datasets and based on their best selected hyper-parameters, we observe that the KNN model with 'overVAR_40' dataset consistently outperforms the other datasets across all metrics; accuracy, precision, and AUC_ROC.

Metrics of performance and (K):

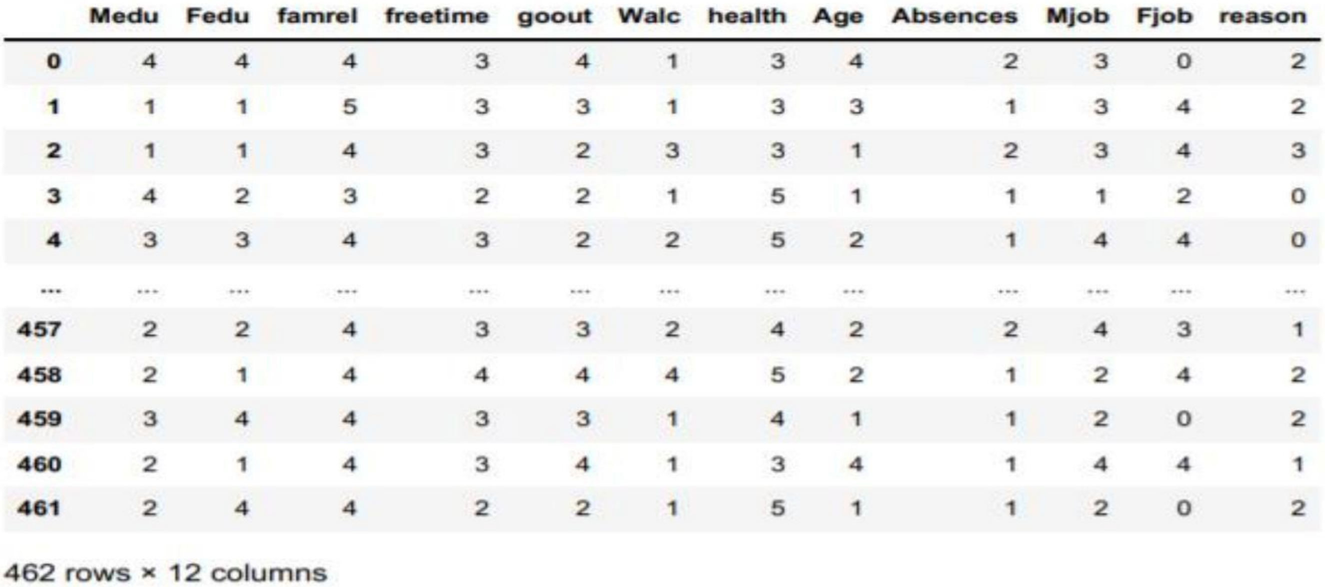
We noticed that the metrics of performance(accuracy, precision, and AUC_ROC) tend to decrease as the number of neighbours (k) increases. This decrease occurs due to the fact that considering too many neighbours can weaken the decision-making process. When KNN does so, it may consider neighbours who are not particularly relevant to the point being classified. These less prominent neighbours can mislead the classification, which lower the overall performance.

Accuracy, precision and AUC_ROC:

The highest accuracy (69.92%) was achieved by the (overVAR_40) dataset with 2 neighbours and distance weighting. As the number of neighbours increases from 2 to 8, we noticed that the accuracy decreases from 69.92% to 67.76%. The same pattern is observed in precision and AUC_ROC. The same dataset achieved the highest precision of 71.83% with 2 neighbours and uniform for weighting, and the highest sensitivity (AUC_ROC) for a value of 72.40% and distance weighting. It is obvious that, as far as KNN model is concerned, a smaller value of K results in better accuracy. The following table is a demonstration of the best dataset across the three metrics with their best hyper-parameters.

Best dataset	metric	n_neighbours	weights	Value%
overVAR_40	accuracy	2	distance	69.92
overVAR_40	precision	2	uniform	71.83
overVAR_40	sensitivity	2	distance	72.40

Table 10: Best dataset for KNN



	Medu	Fedu	famrel	freetime	goout	Walc	health	Age	Absences	Mjob	Fjob	reason
0	4	4	4	3	4	1	3	4	2	3	0	2
1	1	1	5	3	3	1	3	3	1	3	4	2
2	1	1	4	3	2	3	3	1	2	3	4	3
3	4	2	3	2	2	1	5	1	1	1	2	0
4	3	3	4	3	2	2	5	2	1	4	4	0
...	...	...	...	...	...	...	...	...	...	...	...	...
457	2	2	4	3	3	2	4	2	2	4	3	1
458	2	1	4	4	4	4	5	2	1	2	4	2
459	3	4	4	3	3	1	4	1	1	2	0	2
460	2	1	4	3	4	1	3	4	1	4	4	1
461	2	4	4	2	2	1	5	1	1	2	0	2

462 rows × 12 columns

Dataset that achieved the best score in KNN

Decision Tree CART:

Similarly to KNN, CART model was evaluated across different datasets using the performance metrics: accuracy, precision, and AUC-ROC, and based on their best-selected hyper-parameters. The following tables are representations of the datasets along with their metrics and hyperparameters:

Dataset	max_depth	max_features	min_samples_leaf	Accuracy%
underCHI2_20	2	5	5	68,9015
underMI_20	2	5	5	68,9015
underCHI2_40	2	5	7	68,9015
underMI_40	2	5	7	68,9015

Table 11: best accuracy in CART

Decision Tree CART model consistently achieves a higher percentage of accuracy at 68.90%. across diverse datasets. The accuracy percentage remain stable in all the datasets presented in the table, which in turn, reflects the model's ability to correctly classify.

Dataset	max_depth	max_features	min_samples_leaf	Precision%
underCHI2_20	2	4	5	83,6061
underMI_20	2	4	5	83,6061
underCHI2_30	2	5	5	82,6667
underMI_30	2	5	5	82,6667
brutCHI2_30	3	5	5	81,2846

Table12: best precision in CART

The metric of precision demonstrates significant results, with datasets 'underCHI2_20' and 'underMI_20' achieving the higher percentages of precision (83.60%). This reflects the model's ability to reduce false positives, and identify positive instances correctly.

Dataset	max_depth	max_features	min_samples_leaf	AUC-ROC%
underMI_30	4	5	9	72,9121
underMI_20	4	6	5	72,7426
underCHI2_20	4	6	5	72,7426
underCHI2_30	4	5	10	72,6753
brutMI_30	4	5	5	72,3193

Table 13: best AUC_ROC in CART

The discriminatory ability of CART model is higher with a top AUC_ROC of 72.92% with the dataset ‘underMI_30’. This metric reflects the model’s effectiveness in making correct predictions.

Analysis of CART results:

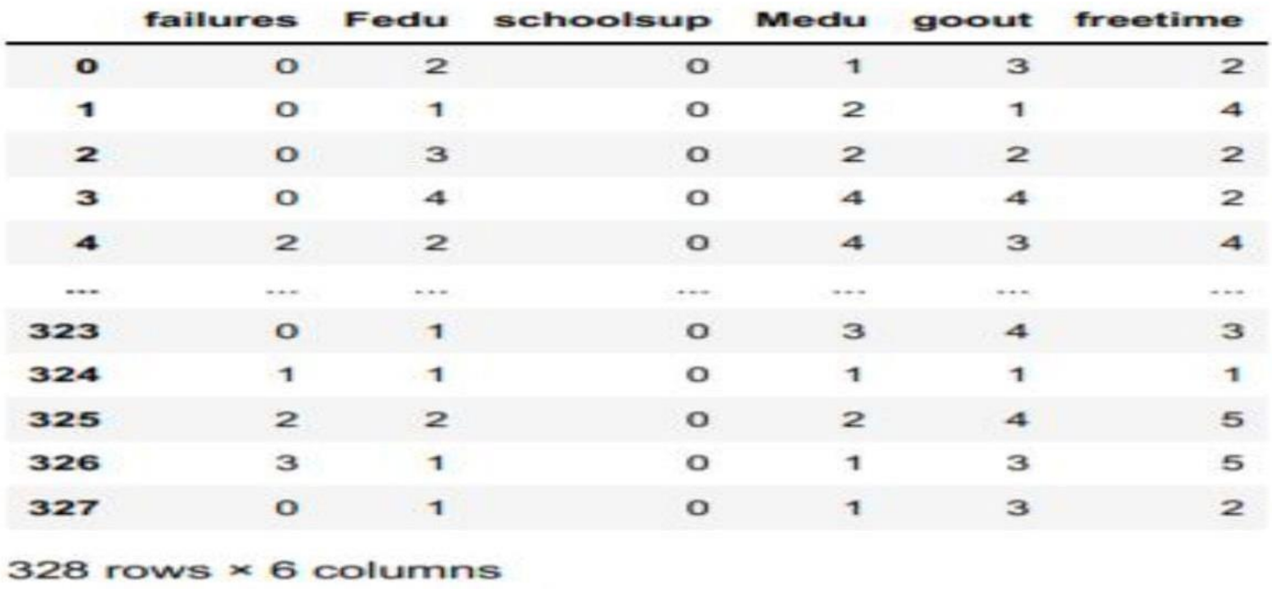
After testing the performance of Decision Tree CART model across different datasets, we obtained promising results in the three metrics; accuracy, precision, and sensitivity. The datasets 'underCHI2_20, underMI_20, underCHI2_40, and underMI_40' achieved high percentages of accuracy (68.90%) with similar hyper-parameters for 2 as max-depth setting, 5 for max-features, and a default min-samples-leaf.

The same dataset 'underCHI2_20' that scores a high accuracy percentage of 68.90%, achieves the highest score of precision as well as sensitivity, with a percentage of 83.60% for precision and 72.74 % for AUC_ROC. The high precision was achieved with a max_depth setting of 2, 4 as a max_feature, and 5 as the minimum samples per leaf. While the sensitivity's settings were 4 for max_depth, 5 for max_features and 9 as the minimum samples leaf. Attaining such a high score in sensitivity validates the model's efficiency and reliability in tasks requiring classification.

The dataset 'underCHI2_20' emerges as the best model of Decision Tree CART, and can be represented as the following:

Dataset	Accuracy%	Precision%	AUC_ROC%
underCHI2_20	68,9015	83,6061	72,7426

Table 14: Best model of CART



	failures	Fedu	schoolsup	Medu	goout	freetime
0	0	2	0	1	3	2
1	0	1	0	2	1	4
2	0	3	0	2	2	2
3	0	4	0	4	4	2
4	2	2	0	4	3	4
...	...	...	...	...	...	...
323	0	1	0	3	4	3
324	1	1	0	1	1	1
325	2	2	0	2	4	5
326	3	1	0	1	3	5
327	0	1	0	1	3	2

328 rows × 6 columns

The dataset that achieved the best score in CART

Discussion:

The results we obtained from KNN and CART models offer insightful information regarding the prediction of students performance. The aim behind training the two models ‘KNN and CART’ is not to compare them and select the best, however, we can utilize them both for the same purpose, and obtain more accurate and precise results.

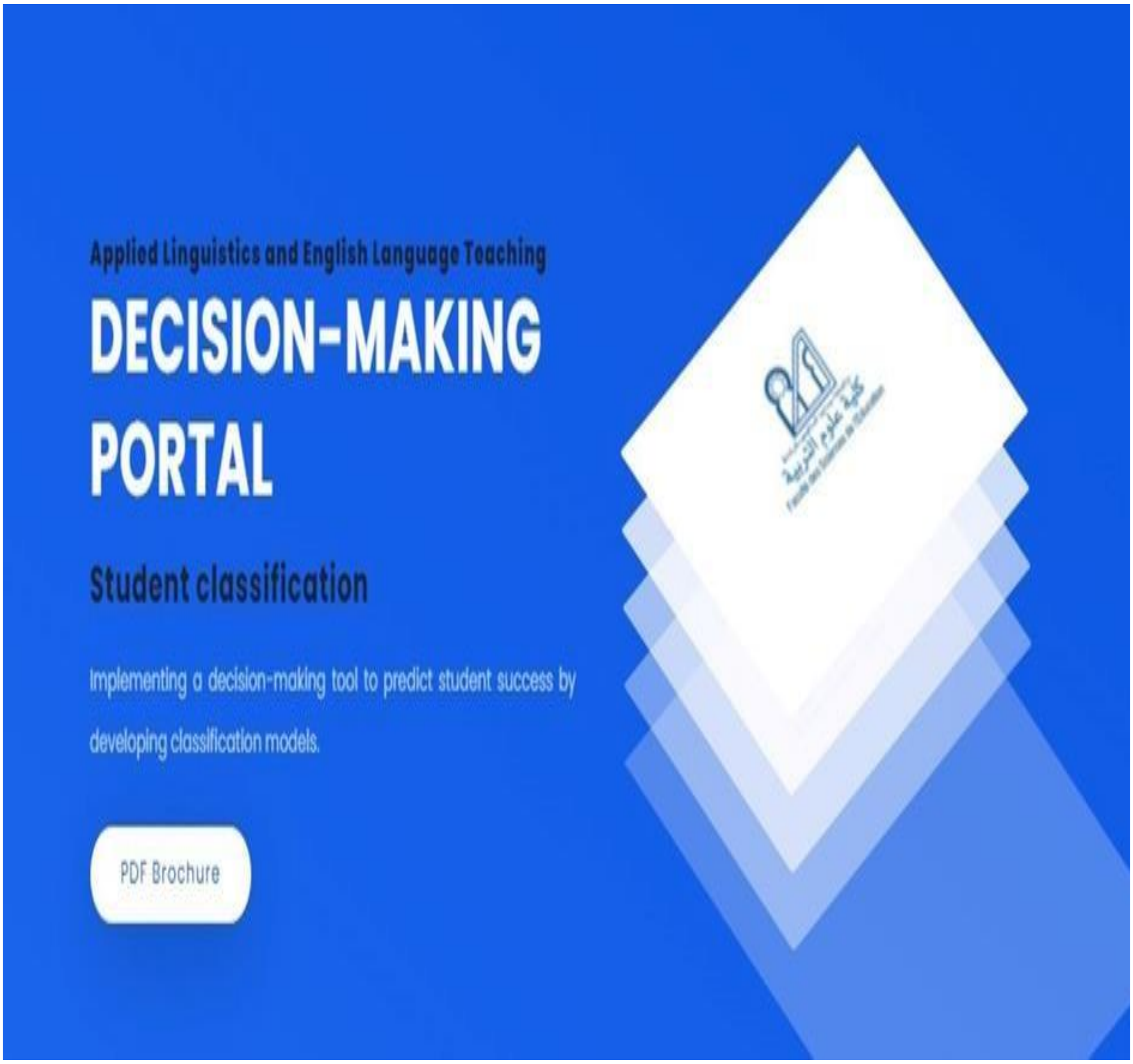
The CART model with ‘underCHI2_20’ in particular demonstrated the highest precision at 83.61%, while KNN with ‘overVAR_40’ dataset reached a percentage of 72.40% of sensitivity, reflecting their validity and accuracy in identifying students who are more likely to fail or to succeed. This superior precision confirms that CART and KNN models are indeed useful in educational settings where the aim is to accurately classify students.

The findings highlight the importance of hyper-parameter tuning to optimize the performance of the model. As far as KNN is concerned, a small number of neighbours yield better results, indicating the necessity of balancing the size of neighbours to keep a precise classification. CART model from the other hand, demonstrated the need for adjusting the Max-depth, Max_Features, and Min_Samples_Leaf to enhance the model over all generalizability to real-world datasets.

After all the process and stages we went through, from training the models to evaluating them, we can confidently utilize the model on new data points. Fortunately, we created a website dedicated only to checking datapoints on our models KNN and CART to facilitate the whole process and make it accessible for everybody to use. Teachers, educators and schools can, thus, benefit form a personal and easy access which in turn enables them to keep track of students progress.

Implementation of KNN and Decision Tree CART:

To implement the classification models; KNN and Decision Tree CART, we developed an interactive website. It offers a user-friendly interface and enables users to enter new data points and obtain accurate predictions from both models quickly and effortlessly. The website is basically a demonstration of the practical application of KNN and CART to real-world data. The following images are screenshots taken from the website, we evaluate two new data points using the two different models.

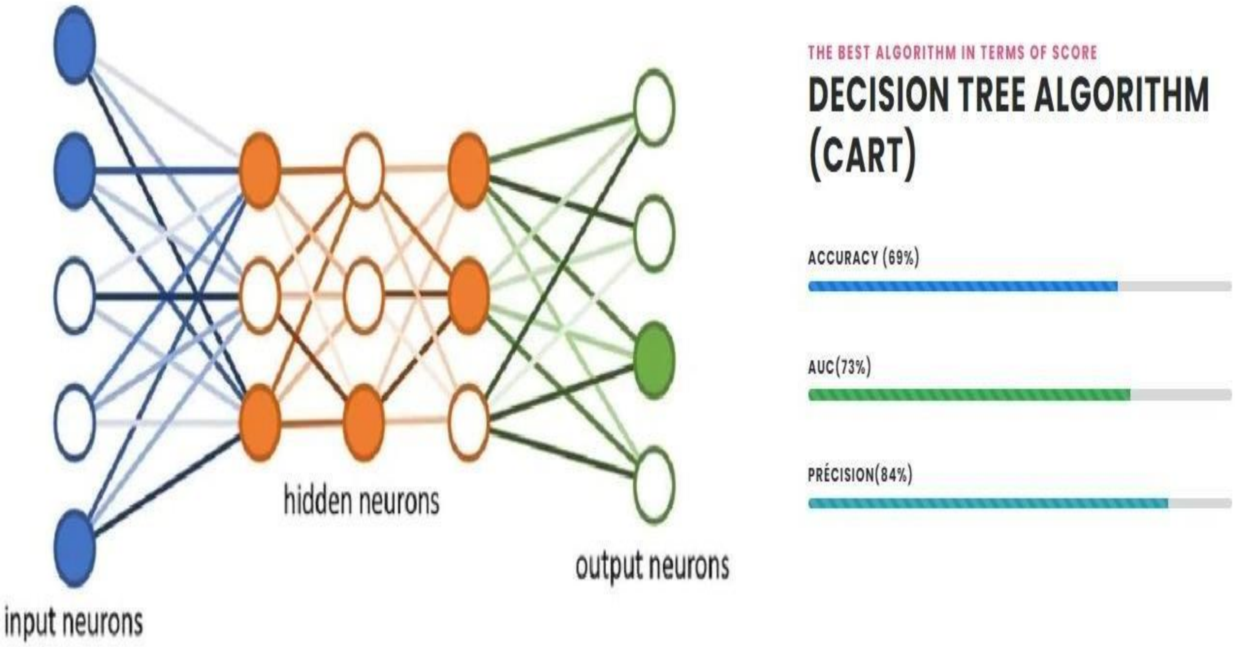


Applied Linguistics and English Language Teaching

DECISION-MAKING PORTAL

Student classification

Implementing a decision-making tool to predict student success by developing classification models.



THE BEST ALGORITHM IN TERMS OF SCORE

DECISION TREE ALGORITHM (CART)

ACCURACY (69%)

AUC(73%)

PRÉCISION(84%)

PROJECT IMPLEMENTATION

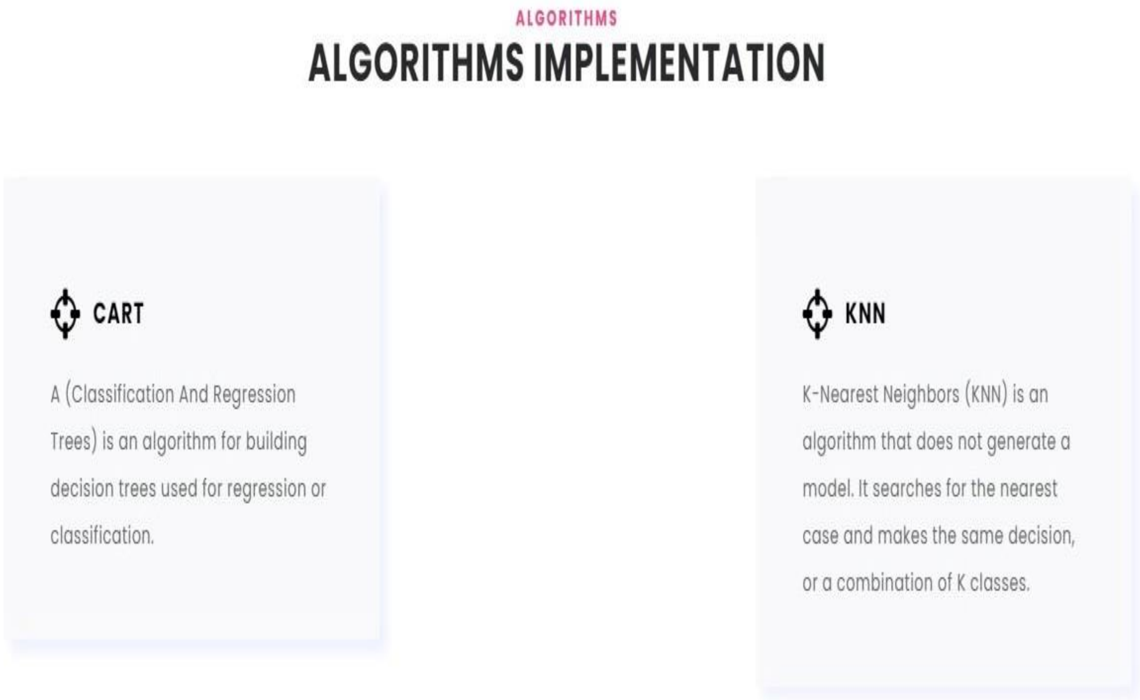
TECHNOLOGIES USED

MODELING AND EVALUATION (MACHINE LEARNING)

In project implementation, Python language along with its various Machine Learning libraries was utilized.

WEB DEVELOPMENT

The web application was developed using Python for the back end, specifically Django, and React JS for the front end.



ALGORITHMS

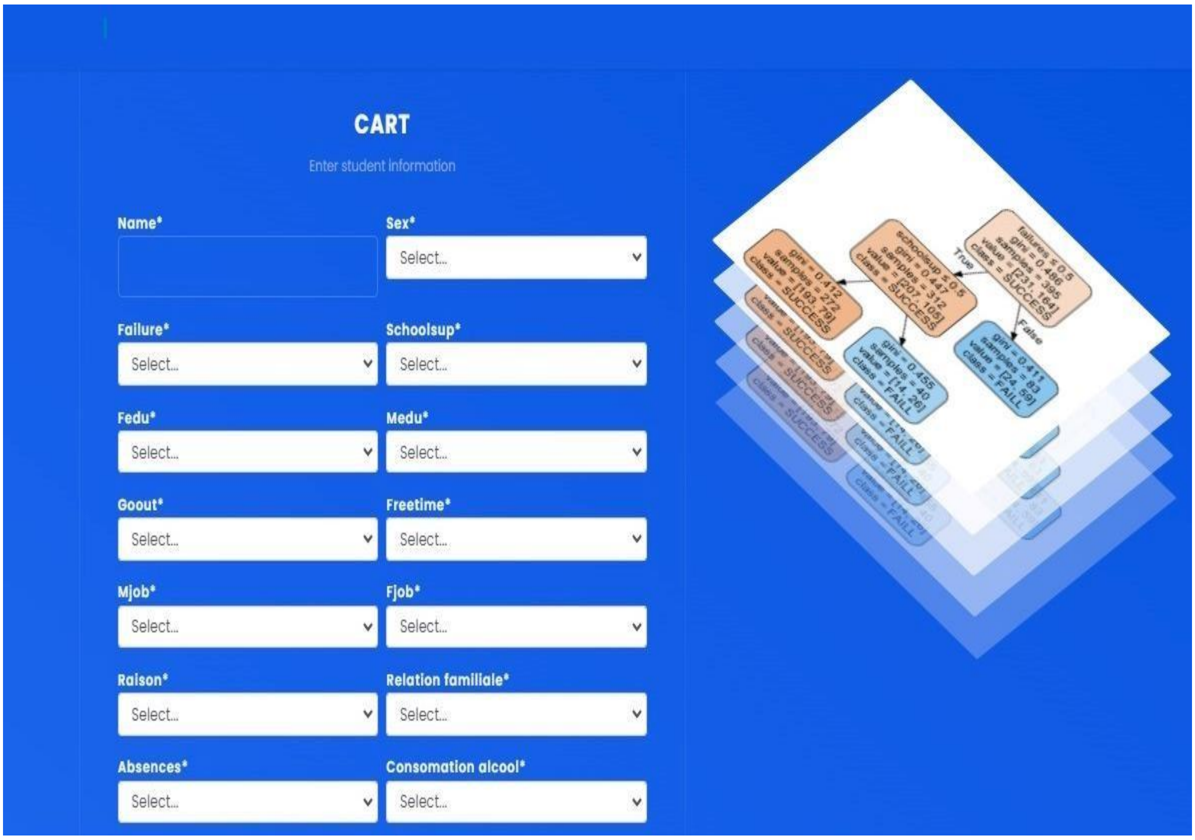
ALGORITHMS IMPLEMENTATION

CART

A (Classification And Regression Trees) is an algorithm for building decision trees used for regression or classification.

KNN

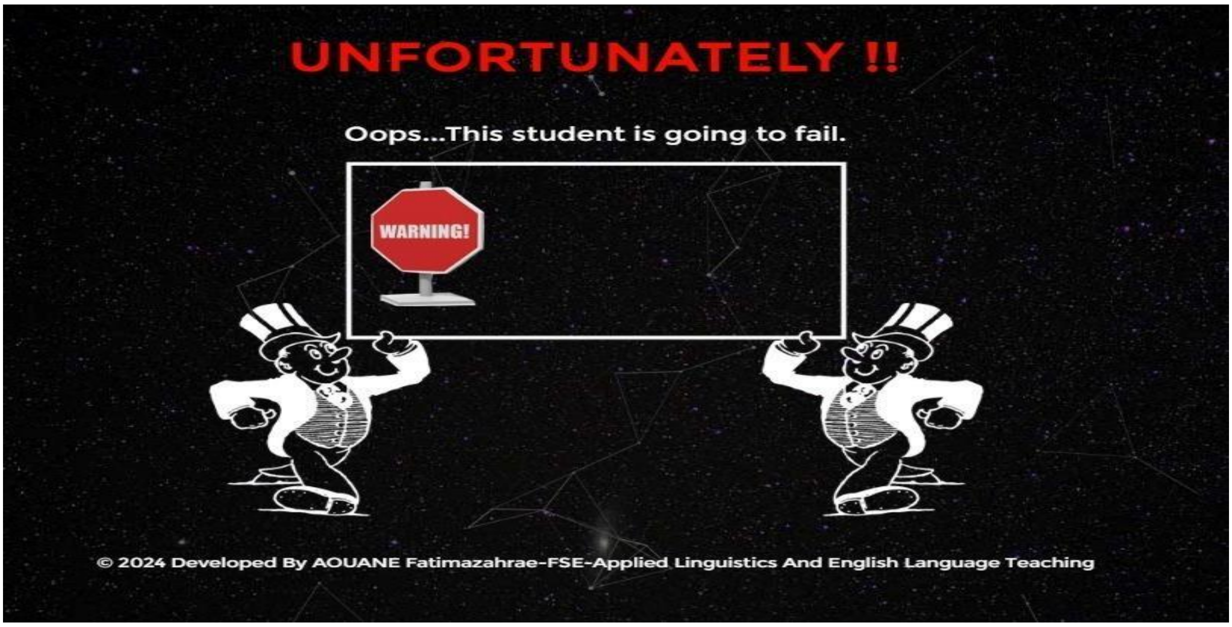
K-Nearest Neighbors (KNN) is an algorithm that does not generate a model. It searches for the nearest case and makes the same decision, or a combination of K classes.



CART

Enter student information

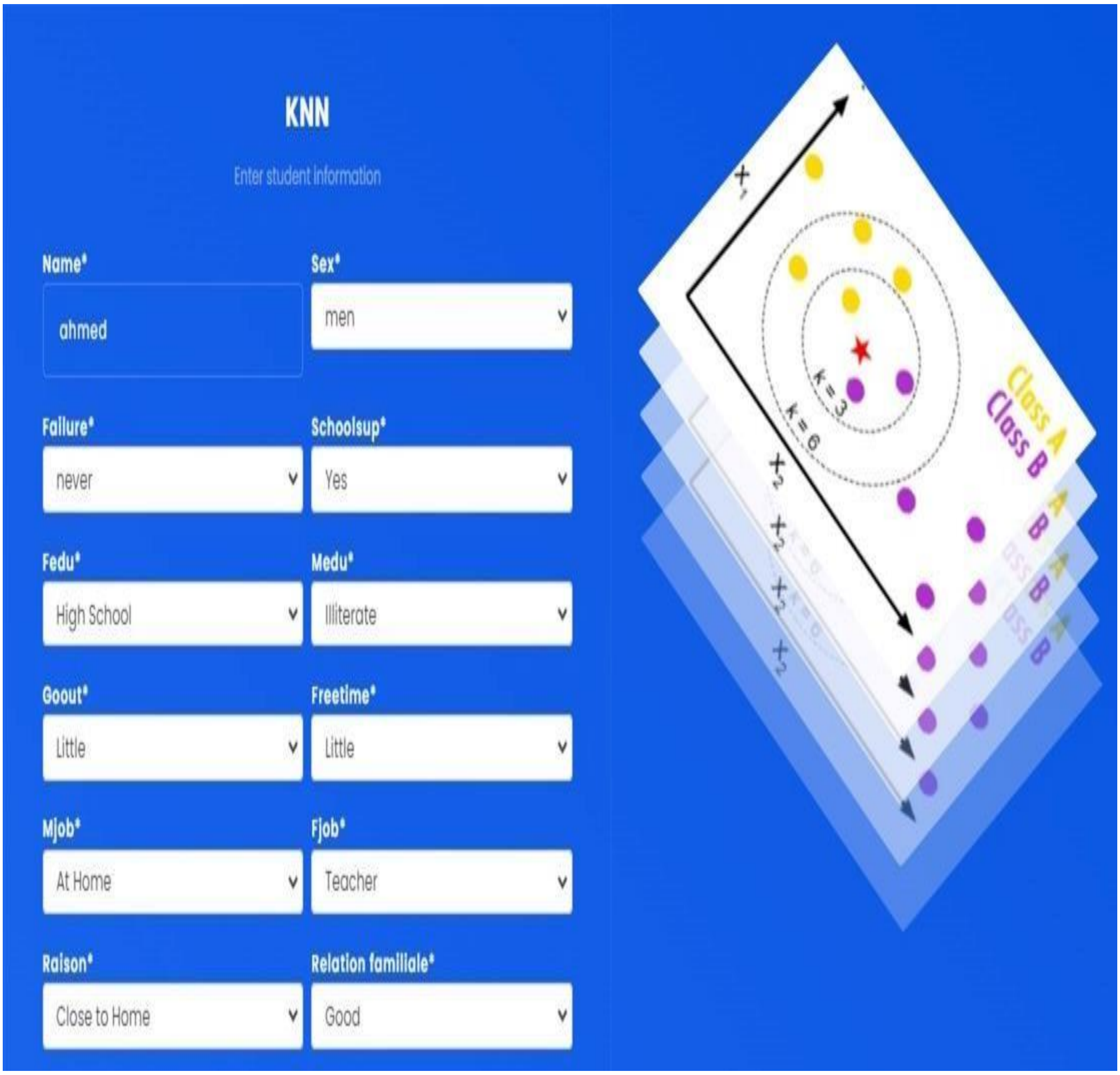
Name*	Sex*
	Select...
Failure*	Schoolsup*
Select...	Select...
Fedu*	Medu*
Select...	Select...
Goout*	Freetime*
Select...	Select...
Mjob*	Fjob*
Select...	Select...
Raison*	Relation familiale*
Select...	Select...
Absences*	Consomation alcool*
Select...	Select...



UNFORTUNATELY !!

Oops...This student is going to fail.

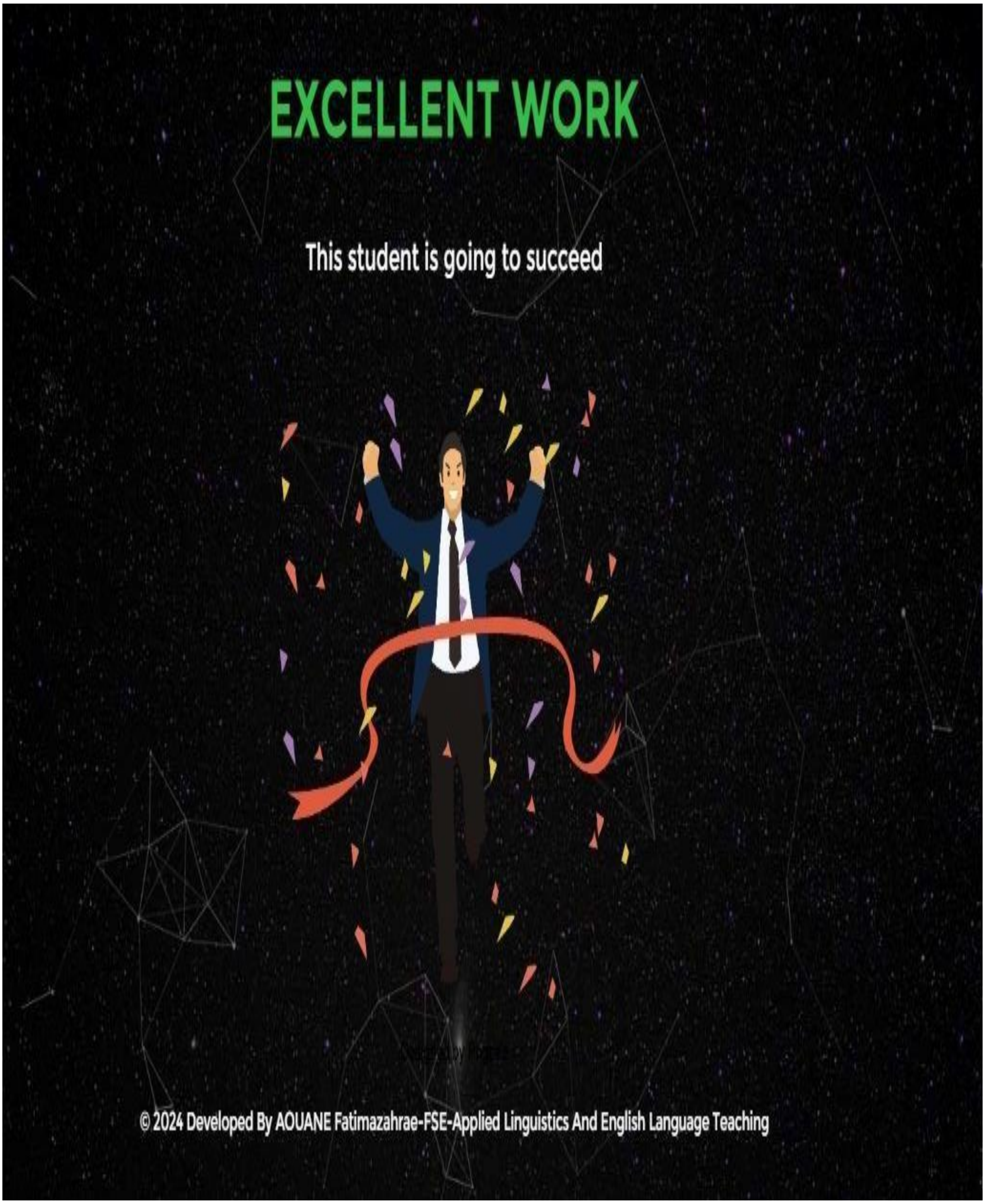
© 2024 Developed By AOUANE Fatimazahrae-FSE-Applied Linguistics And English Language Teaching



KNN

Enter student information

Name* ahmed	Sex* men
Failure* never	Schoolsup* Yes
Fedu* High School	Medu* Illiterate
Goout* Little	Freetime* Little
Mjob* At Home	Fjob* Teacher
Raison* Close to Home	Relation familiale* Good

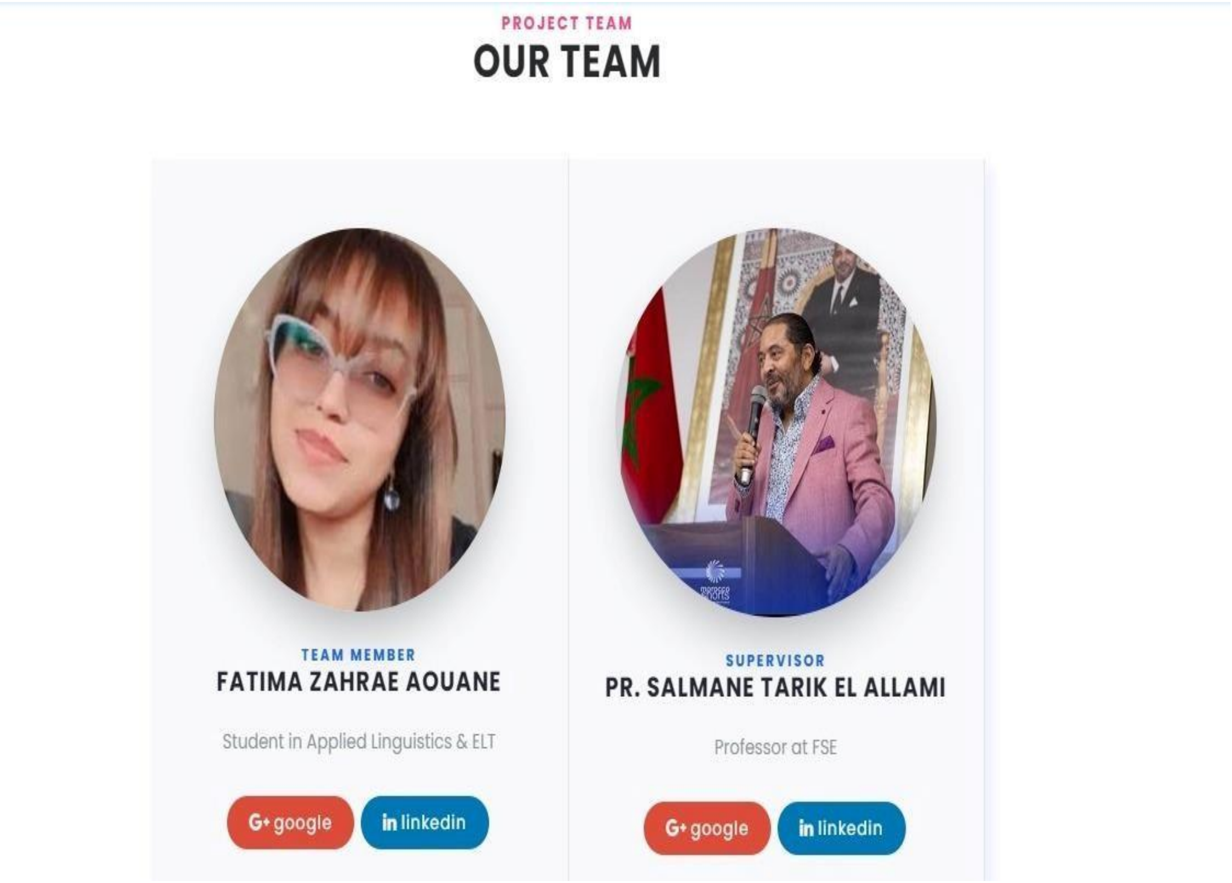


EXCELLENT WORK

This student is going to succeed

© 2024 Developed By AOUANE Fatimazahrae-FSE-Applied Linguistics And English Language Teaching

As demonstrated in the images above, the structure of the website is pretty simple and accessible for everyone to use. It only requires filling out the information needed based on which algorithm is selected, either KNN or Decision Tree CART, then the algorithms quickly yield the results as either the student is going to fail or will succeed.



PROJECT TEAM

OUR TEAM

TEAM MEMBER

FATIMA ZAHRAE AOUANE

Student in Applied Linguistics & ELT

SUPERVISOR

PR. SALMANE TARIK EL ALLAMI

Professor at FSE

Limitations and Challenges:

The main challenges of this research rise from the fact that machine learning is not an easily accessible science that can be well understood in a limited period of time, and as a student of applied linguistics, concepts of machine learning are not introduced and, therefore, creates a need for these fields to benefit from each other and to finally enhance the techniques and tools used in education. Moreover, and despite the researcher's efforts to rigorously select the best models for performance classification purposes, the choice of KNN and CART over other models of classification reflects the nature of the models per se. That is to say, KNN and CART are known for their simplicity and interpretability, although other models like neural networks or ensemble methods such as Random Forest might yield better results but they require more expertise, time and resources. Moreover, there is another limitation which is the fact that this study does not provide any strategies or techniques to intervene and help in decision making, basically due to time and resources constrains. However, it opens the door for more research and studies to be conducted regarding the same issue.

Practicality to the Moroccan context:

The implementation of machine learning algorithms; K-NN and Decision Tree CART to forecast students performance offers practical benefits that can significantly impact education world-widely and Morocco in particular. There has been a need to generate and store students' information in electronic datasets since the launch of the first initiatives to integrate ICTs in Morocco. Such a large amount of data is a treasure for data analysts to reveal hidden patterns, transform data into knowledge, and guide the decision-making. Through implementing data-driven techniques to predict students' success or failure, policy makers and stakeholders can identify students in need of support and provide targeted interventions to guide and support them accordingly. This predictive ability does not only enhance the efficiency of the educational interventions but also permits schools to effectively allocate their resources and customize the teaching strategies to meet the diversity of students' needs, creating a more personalized form of learning and enhancing the overall quality of education in Morocco.

General Conclusion:

This study explored the realm of predictive analytics within the educational settings, aiming to use advanced machine learning algorithms, notably KNN (K-nearest neighbours) and Decision Tree CART (Classification and Regression Trees) to predict students' academic performance. These algorithms were selected for their ability to deal with complex data patterns and generate precise predictions. Through a careful process of models selection, data selection, models training, evaluation and implementation, we succeeded in predicting students at risk of failing, and those who are more likely to succeed.

The findings of this study demonstrate promising results which, in turn, indicates the fact that the selected machine learning models; K-NN and CART are capable of providing precise and accurate predictions regarding students' performance. Decision Tree CART model achieved a high accuracy score of 68.90% in various datasets 'underCHI2_20' 'underMI_20', and impressive precision scores of 83.60 with underCHI2_20' and 'underMI_20' datasets, and the same dataset achieved a notable sensitivity score of 72.74%. K-NN from the other hand, attained high scores with the same dataset of 'overVAR_40'. It achieved a percentage of 69.92% of accuracy, a percentage of 71.83% of precision, and a sensitivity of 72.40%. These results further emphasizes the models reliability and accuracy to identify students as either will succeed or at risk of failing and, therefore, require immediate interventions and additional support.

These machine learning algorithms are solely means for an end. We started from the need that exists in all educational settings which is the inability to identify students at risk of failing and students who will succeed, and therefore, the interventions we could make were limited. However, and by means of machine learning, educators and policy makers can make informed decisions to improve students learning outcomes as well as educational practices. However, it is worth mentioning that utilizing machine learning algorithms in educational settings poses crucial ethical considerations regarding bias, fairness, and the implementation of these models should be carefully addressed to avoid all possible inequality for all students.

Generally, this research contributes to the growing body of knowledge on predictive analytics in education and demonstrates the significance of benefiting from other fields for the common good of education. It prompts questions on how pedagogical interventions can support students if the variables affecting their performance are mainly socio-economic and demographic.

References:

Ahajjam, T., Moutaib, M., Aissa, H., Azrour, M., Farhaoui, Y., & Fattah, M. (2022). Predicting students' final performance using artificial neural networks. Big Data Mining and Analytics, 5(4), 294-301.

Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student performance prediction using machine learning techniques. Education Sciences, 11(9), 552. https://doi.org/10.3390/educsci11090552

Berland, M., Baker, R. S., & Blikstein, P. (2014). Educational data mining and learning analytics: Applications to constructionist research. Technology, Knowledge and Learning, 19, 205-220.

Bidgoli, B. M., Kashy, D., Kortemeyer, G., & Punch, W. (2003). Predicting student performance: An application of data mining methods with the educational web-based system lon-capa. In Proceedings of ASEE/IEEE Frontiers in Education Conference

Bunkar, K., Singh, U. K., Pandya, B., & Bunkar, R. (2012). Data mining: Prediction for performance improvement of graduate students using classification. In Wireless and Optical Communications Networks (WOCN), 2012 Ninth International Conference on (pp. 1-5). I EEE.

Dervenis, C., Kyriatzis, V., Stoufis, S., & Fitsilis, P. (2022, September). Predicting Students' Performance Using Machine Learning Algorithms. In Proceedings of the 6th International Conference on Algorithms, Computing and Systems (pp. 1-7).

Gray, G., McGuinness, C., & Owende, P. (2014). An application of classification models to predict learner progression in tertiary education. In Advance Computing Conference (IACC), 2014 IEEE International (pp. 549-554). IEEE.

Jishan, S. T., Rashu, R. I., Haque, N., & Rahman, R. M. (2015). Improving accuracy of students final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique. Decision Analytics, 2(1), 1-25.

Mayilvaganan, M., & Kalpanadevi, D. (2014). Comparison of classification techniques for predicting the performance of students academic environment. In Communication and Network Technologies (ICCNT), 2014 International Conference on (pp. 113-118). IEEE.

Mishra, T., Kumar, D., & Gupta, S. (2014). Mining students’ data for prediction performance. In Proceedings of the 2014 Fourth International Conference on Advanced Computing &

Communication Technologies, ACCT ’14 (pp. 255-262). IEEE Computer Society. https://doi.org/10.1109/ACCT.2014.10

Mohamad, S. K., & Tasir, Z. (2013). Educational data mining: A review. Procedia - Social and Behavioral Sciences, 97(6), 320–324. https://doi.org/10.1016/j.sbspro.2013.10.240

Mohammed, A. D., Khalil, N., Mohssine, B., Rachida, I., Soufiane, B., & Mohammed, T. (2019). Using students’ data to improve the quality of the education in Moroccan institution. International Journal of Open Information Technologies, 7(11), 59-6

Natek, S., & Zwilling, M. (2014). Student data mining solution–knowledge management system related to higher education institutions. Expert Systems with Applications, 41(14), 6400-6407.

Qazdar, A., Er-Raha, B., Cherkaoui, C., & Mammass, D. (2019). A machine learning algorithm framework for predicting students performance: A case study of baccalaureate students in Morocco. Education and Information Technologies, 24, 3577-3589.

Ramesh, V., Parkavi, P., & Ramar, K. (2013). Predicting student performance: A statistical and data mining approach. International Journal of Computer Applications, 63(8), 35-39

Romero, C., Ventura, S., Espejo, P. G., & Hervas, C. (2008). Data mining algorithms to classifyh students. In Educational Data Mining 2008.

What is Data Mining? Retrieved from https://bootcamp.rutgers.edu/blog/what-is-data-mining

K-Nearest Neighbor Algorithm. Retrieved from https://arize.com/blog-course/knn-algorithm-k-nearest-neighbor/#:~:text=The%20aim%20of%20feature%20selection,improve%20performance%20in%20several%20ways

Feature Selection Techniques in Machine Learning. Retrieved from https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning

AUC-ROC Curve in Machine Learning. Retrieved from https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/#:~:text=The%20Area%20Under%20the%20Curve,the%20positive%20and%20negative%20classes.