peoples democratic and republic of algeria

ministry of high education and scientific research

**IBN KHALDOUN UNIVERSITY - TIARET**

**MASTER THESIS**

introduced to:

mathematics and computer science faculty

computer science departement

for obtaining :

**MASTER DECREE**

SPECIALITY: software engineering

**By**

**KHELIL BENAISSA**

**SAHD SID AHMED**

About:

# An interactive learning analytics tool to support higher education stakeholder to more explain and interpret predictive student's failure or success

Submited o, 26/06/2023 in Tiaret in front of the jury:

| | | |
|---|---|---|
| Mr chadli Abdelhafid | MCA TIARET UNIVERSITY | RESIDENT |
| MR TALBI OMAR | MCA TIARET UNIVERSITY | Supervisor |
| MR LAKHDARI AICHA | MCA TIARET UNIVERSITY | EXAMINER |

# ACKNOWLEDGMENTS

*Our thanks go in the first place to GOD the Almighty for having helped us to complete this modest work.*

*We express our sincere gratitude to our supervisor Mr. TALBI Omar for his constant monitoring, for his great scientific qualities and who has always pushed us to do better, thanks to his precious advice and guidance.*

*Our warmest thanks also go to the president and the members of the jury for agreeing to examine and evaluate our work.*
*We humbly dedicate this work to our parents, who made it all easier with their constant prayers and support.*

*We also want to thank REZZAG Bilal, AMARA Amine for their support.*

*We would like to thank all the Professors and teachers who have participated in our training since our first cycle of studies until the end of our university cycle.*

*We thank all those who have directly or indirectly contributed to the realization of this work.*

# Abstract

**Learning Analytics has great potential for improving student learning and prediction making about the probability of their success or failure.**
**However, these machine learning algorithms generally do not explain these predictions. Even if a machine learning model works well, ignoring why it makes certain decisions excludes the human touch, which in turn negatively impacts the stakeholders' trust.**
**Also, knowing the "why" can help the teaching staff decipher the data and the reason why a particular prediction was made.**
**In this study, we are interested in providing the educators with an interactive learning analytics tool that allows them to be actors in this 'loop' of analytics and predictions, thus having a proactive role in deciding a student's future.**

**Keywords:**
**Learning Analytics,decipherability, prediction...**

# Summary

**Higher education institutions are interested in the success and failure of students to ensure that they have the support they need to achieve their academic objectives.**
**The aim of learning analytics is to analyse and understand students' data in order to enhance their learning results.**
**The primary goal our study is to provide education experts with an interactive learning analytics tool that allows them to participate actively in the analytics and prediction loop.**
**The proposed technology will improve learning analytics' explainability and interpretability.**
**The application seeks to make machine learning models clear and more accessible allowing educators to understand why a specific prediction was made.**

# Acronymes

| | |
|---|---|
| **LA** | Learning Analytic |
| **KDD** | Knowledge Discovery in Databases |
| **EDM** | Educational Data Mining |
| **ML** | machine learning |
| **MOOCs** | Massive Open Online Courses |
| **LMS** | Learning management system |
| **VLE** | Virtual learning environment |
| **HE** | Higher Education |
| **HEI** | Higher Education Institutions |

# Contents

# List of Figures

# Introduction

# Introduction

Learning analytics is an emerging field that has the potential to significantly reduce errors and increase accuracy and precision in many fields which include but are not limited to business intelligence and action analysis. web analysis, academic analysis, and educational data mining.
Learning analytics has proved the most valuable in educational data mining as it can provide insight into a student's performance, identify possible areas of weakness, and predict the probability of success and failure by examining the collected data. That can be provided manually or collected online.
Machine learning is an assembly of algorithms that use computational methods to "learn" information directly from data without relying on a predetermined equation as a model. It has three main branches: reinforcement learning, unsupervised learning, and supervised learning.
One of the types of supervised learning is classification, which uses an algorithm known as a decision tree.

## Context

Learning Analytics is mostly used in the collection and analysis of students' data to improve their learning outcomes. Through the use of machine learning techniques, Learning Analytics can analyse large amounts of data, including students' academic performance and behaviour on learning platforms.
For the purpose of creating machine learning models that ensure that education stakeholders can effectively use the insights provided by these models to achieve their purposes.

## Problem Statement

Machine learning techniques generally do not explain these predictions. Even if a machine learning model works well, ignoring why it made a certain decision leads education stakeholders to lack trust. Also, knowing the "why" can help the teaching staff know more about the problem, the data, and the reason why a particular prediction was made.

## The objectives

In this study, we are interested in providing the educators with an interactive learning analytics tool that allows them to be actors in this 'loop' of analytics and predictions. By deciphering the model and the reasons why a particular prediction was made, thus having a proactive role in deciding a students' future.

## Methodology

In this project, we are planning on deciphering the model that is usually a "black box" into a "white box" allowing us to see the path of prediction and altering its framework for more accuracy and control.

## Outline

This graduation project is composed of three chapters. The first chapter addresses the importance of adequate support for students in higher education and how predictions can improve it. Familiarise ourselves with learning analytics.
The second chapter defines our objectives and how we're going to accomplish them.
The third chapter illustrates which tools were used to realise our project and how they work.
We end the project by giving an assessment and a few perspectives to build on this work in the future.

# Chapter *I*

## State of Art

# Chapter I

# State of Art

## I.1 Introduction

Predicting student success and failure in higher education has become an increasingly important area of research and practice. Higher education institutions are invested in ensuring that students are successful and have the support they need to achieve their academic and career goals."Higher education institutions are generally interested in the success of students during their study, predicting of student performance also aids educational planners and decision makers and administrators to adequately plan for changes in student population in any direction, the overall success of educational institutions can be measured by the success of its students" [12] . The prediction of the student's success helps the education organization to supply the student with additional assessment; this process also enhances the development of the education system in educational institutions [21].

This chapter aims to be a state of art of this study, it comes in six sections. In the first section, we present the field of learning analytic and its life cycle. In the second section we outline the higher education sector and its stakeholders. In the third section we describe data mining, its process, educational data mining and its aims. In the fourth section we present machine learning fields, supervised learning, we introduce decision trees as a classification technique and their underlying algorithms. In the fifth section we discuss interpretability, explainability, and their relationship. Finally, in section six, we cover the main findings, related work that may have been conducted using these methodologies. We will close this chapter with a conclusion in which we will draw attention to the learning analytic tool's potential by giving stakeholders further explanations.

## I.2 Learning analytics

Learning analytics is defined as "the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" [8].
Learning analytics is an emerging field in which sophisticated analytic tools are used to

improve learning and education. It draws from, and is closely tied to, a series of other fields of study including business intelligence, web analytics, academic analytics, educational data mining, and action analytics [10].

### I.2.1 LA life cycle

There are a multitude of factors that have motivated interest in learning analytics. One motivating factor for the increased interest in learning analytics is the general trend towards increased accountability at all levels of education.
There is more demand on educational institutions across the nation to explain what and how their pupils are learning. Online courses now have their own accreditation criteria, which puts even more pressure on them. One way to document student success is through the use of learning analytics, which also offers tools to promote the kinds of continual improvement that accrediting organisations are looking for [8].
Learning analytics life cycle includes four main stages shown in Fig I.1 [18]:

- **Generation of data:**This process begins in the learning environments, such as MOOCs, LMSs, or any other VLEs, where various stakeholders are present.
- **Data storage:**Students leave a lot of tracable information in their wake. Students are both creators and consumers of data.
- **Analysis:** Analytical techniques look for hidden patterns in datasets used in education. Various analytics methods exist. They were primarily classified by the authors into quantitative and qualitative analysis methodologies.
- **Act:** the analysis's findings should be put into practice. Action in this phase includes prediction, intervention, advice, personalisation, and reflection.

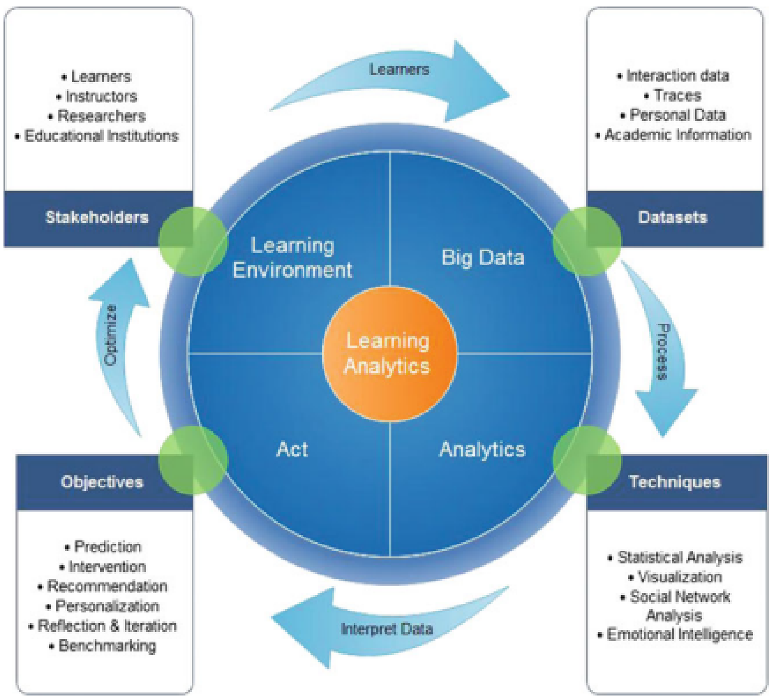


Figure I.1: La life cycle.

On a global scale, decision-makers in HE institutions now place a high priority on student recruitment, management, and retention. Due to the financial losses, lower graduation rates, and poorer school reputation in the eyes of all stakeholders, boosting student retention is of particular importance, and the understanding of the cause of and/or prediction of attrition has come into the spotlight [18].

## I.3 Higher education

Higher education is a rich cultural and scientific asset which enables personal development and promotes economic, technological and social change. It promotes the exchange of knowledge, research and innovation and equips students with the skills needed to meet ever changing labour markets. For students in vulnerable circumstances, it is a passport to economic security and a stable future [32].
Institutions of higher learning are essential to the growth of nations. Higher education institutions are focused on the advancement of practically every facet of life because their main purpose is to create and share information [22].
Education in general and higher education, in particular, is the subject of a great deal of study. Higher education is the backbone of any society. It is the quality of higher education that decides the quality of human resources in a country. Higher education includes college and university teaching and learning process towards which students march to attain the higher educational qualification. Higher education imparts in-depth knowledge and understanding so as to advance the students to new frontiers of knowledge. Higher education also provides opportunities for lifelong learning, allowing people to upgrade their knowledge and skills from time to time based on societal needs [33].

### I.3.1 stakeholder

A stakeholder is defined as “the most cited in the literature a stakeholder in an organization is (by definition) any group or individual who can affect or is affected by the achievement of the organization’s objectives" [22].

### I.3.2 Stakeholders in education

Stakeholders in education include anyone with an interest in an educational system’s success. This includes those who are directly involved (such as parents, teachers, and students) and those indirectly impacted (such as government officials, local business leaders, and volunteers). The list of stakeholders in education involves internal stakeholders and external stakeholders in education, which means those that work within and those that work outside of the school on a day-to-day basis, respectively. Students are internal and primary stakeholders in education, as they are directly impacted by the educational system the most. Yet, all stakeholders play an important role, as each stakeholder is part of a team that works for the success of educational goals. This makes stakeholders in education important [37], as is shown in fig I.2 [18].

| Stakeholder | Objectives, benefits and perspectives |
|---|---|
| Learner | Support the learner with adaptive feedback, recommendations, response to his or her needs, for learning performance improvement |
| Educators | Understand students' learning process, reflect on teaching methods and performance, understand social, cognitive and behavioral aspects |
| Researchers | Use the right EDM technique which fits the problem, evaluation of learning effectiveness for different settings |
| Administrators | Evaluation of institutional resources and their educational offer |

Figure I.2: Overview of the stakeholders .

## I.4 Data mining

Data mining, also called Knowledge Discovery in Databases (KDD), is the field of discovering novel and potentially useful information from large amounts of data [3].
Many industries, including marketing, medical, real estate, customer relationship management, engineering, web mining, etc., can benefit from the use of data mining principles and techniques [30].
Data mining is the process of extracting and discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal of extracting information (with intelligent methods) from a data set and transforming the information into a comprehensible structure for further use. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD [34].
Data mining is an essential step in the knowledge discovery in databases (KDD) process that produces useful patterns or models from data shown in FigI.3 . KDD and data mining have various definitions. KDD is the term used to describe the entire process of extracting useful information from data. Finding new patterns from a multitude of data in databases by concentrating on the algorithms to extract relevant knowledge is known as "data mining." [28].

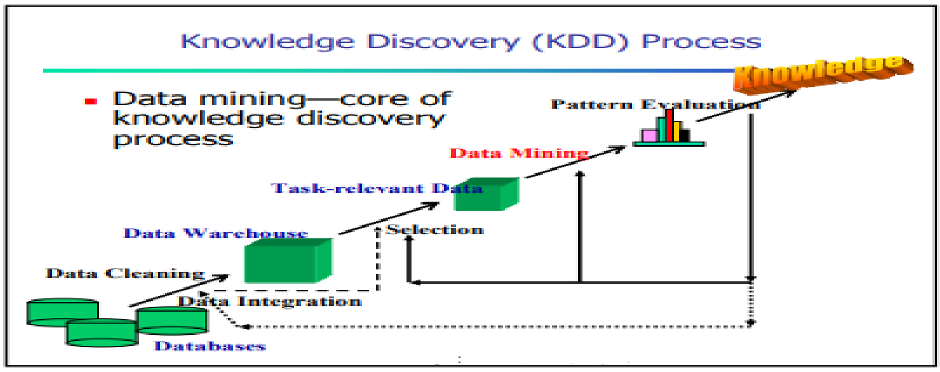


Figure I.3: Data Mining Process .

### I.4.1 KDD process

KDD process consists of iterative sequence methods as follows [28]:

1. Selection: Choosing data from the database that is pertinent to the analytical activity.
2. Preprocessing: Eliminating noise and incorrect information, integrating data from several sources
3. Data transformation: converting data into the right formats for data mining.
4. Data mining: Selecting an algorithm for data mining that is appropriate for the data pattern, extracting data patterns
5. Interpretation/Evaluation: Converting the useful patterns into terms that people can comprehend, and interpreting the patterns into knowledge by deleting redundant or irrelevant patterns.

### I.4.2 The functions of data mining

Data mining has six main functions [28]:

1. Classification is finding models that examine and split a data item into a number of specified classes.
2. Regression is relating a data item to a real-valued prediction variable
3. Clustering is choosing a limited number of clusters or groups to characterize the data
4. Dependency Modeling (Association Rule Learning) is Finding a model that describes important connections between variables
5. Deviation Detection (Anomaly Detection) is Finding the biggest changes in the data
6. Summarization is finding a brief overview for a small amount of data

The possible use of data mining methods to predict variables affecting students' performance using data sets from university students. To estimate a student's success or failure, use three sets of variables [31]:

a. Student history (identity, socioeconomic background, academic direction, age, and gender)

b. The student's participation in their studies (participation in cultural events, meeting with teachers, etc.)

c. The student's perception (views on the course, teachers, and academic context to judge their impact on the students' academic achievement).

### I.4.3 Educational data mining

In recent years, there has been increasing interest in a field of study known as educational data mining, which uses data mining to look into scientific issues inside educational research. The field of scientific research known as "educational data mining" (also known as "EDM") is focused on the creation of techniques for making advances within the special types of data that are generated in educational institutions and applying those techniques to learn more about students and the environments in which they are conducted [36].
Educational Data Mining is an emerging discipline focuses on creating techniques for examining the specific and growing quantities of data that are generated in educational settings and using those techniques to learn more about students and the environments in which they learn. Whether educational data is taken from students' use of interactive learning environments, computer-supported collaborative learning, or administrative data from schools and universities, it often has multiple levels of specific structure, which often need to be determined by properties of the data itself, instead than in advance. Issues of time, sequence, and context also play important roles in the study of educational data (educationaldatamining.org). A newly developed data mining method called "educational data mining" can be used on information linked to the field of education [30].
In order solve educational issues, Educational Data Mining (EDM), an interdisciplinary field of study, applies machine learning, statistics, Data Mining (DM), psycho-pedagogy, information extraction, mental health, and recommender systems methods and techniques to various educational data sets. EDM is involved with evaluating data produced in a learning environment utilizing various systems for the purpose to create models that will enhance learning and institutional effectiveness [9].

### I.4.4 The main Goals of educational data mining

EDM has both pure research purposes, such as gaining a greater knowledge of educational happenings, as well as applied research objectives, such as enhancing the learning process and directing students' learning. These objectives may require a unique set of measurement tools and are challenging to quantify [26].
The main Goals of educational data mining are [1].

- Predicting students' future academic conduct by developing student models that include specific data on knowledge, motivation, reflection, and attitudes of students.
- Identifying or enhancing domain models that describe the material to be learnt and the ideal pedagogical flow
- Researching the outcomes of various pedagogical supports that learning software can offer.
- Developing computational models that include student, domain, and software pedagogy models to advance scientific understanding of learning and learners.

According to (Hasan, Raza, et al 2020) [14] some of the important activities are recognized in educational datamining, as shown in fig I.4:

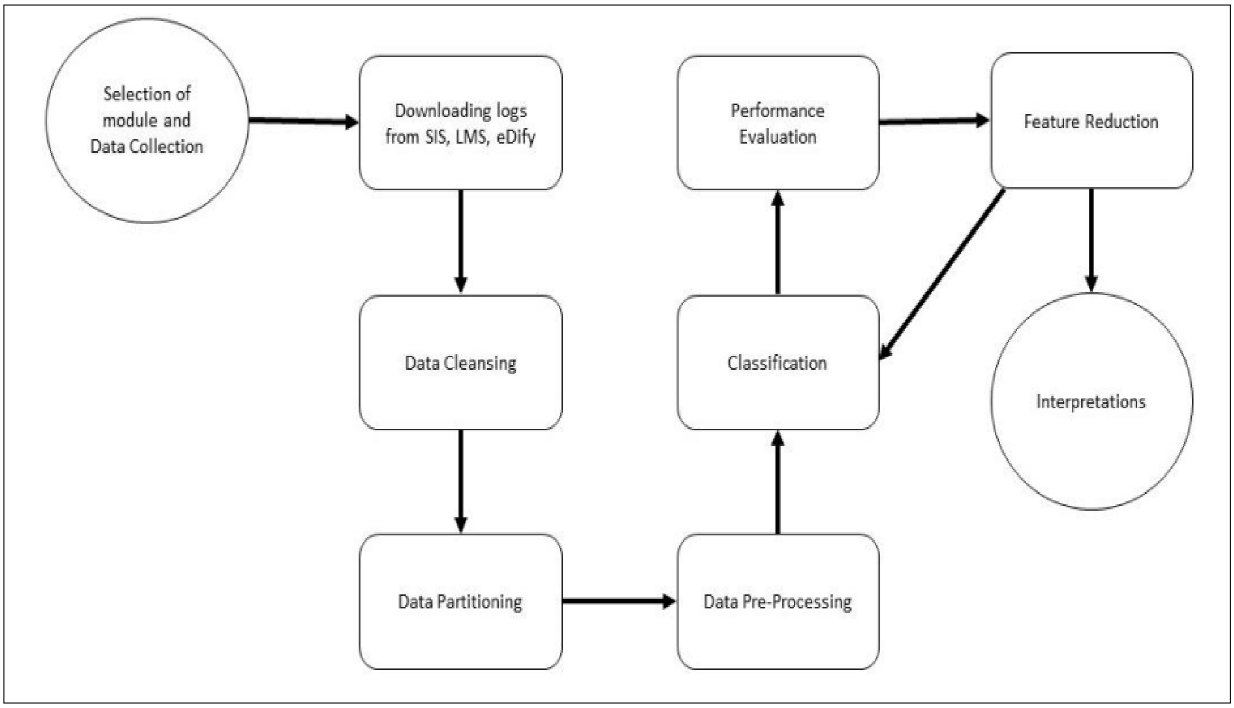


Figure I.4: The cycle of applying educational data mining in research.

Such a group of interdependent disciplines can be found with the use of education data mining. Students can use this information to determine which academic fields will be most important in the future [23], as seen in Fig I.5.


Figure I.5: EDM and Machine Learning

### I.4.5 Data mining work methodology

Data mining work methodology start from the problem definition which means description in the data mining work approach, then Data collecting from student databases is discussed. Data is arranged so that preprocessing is unnecessary, and after that, we use the association, classification, and clustering data mining techniques before evaluating the results [28]. As seen in Fig I.6.

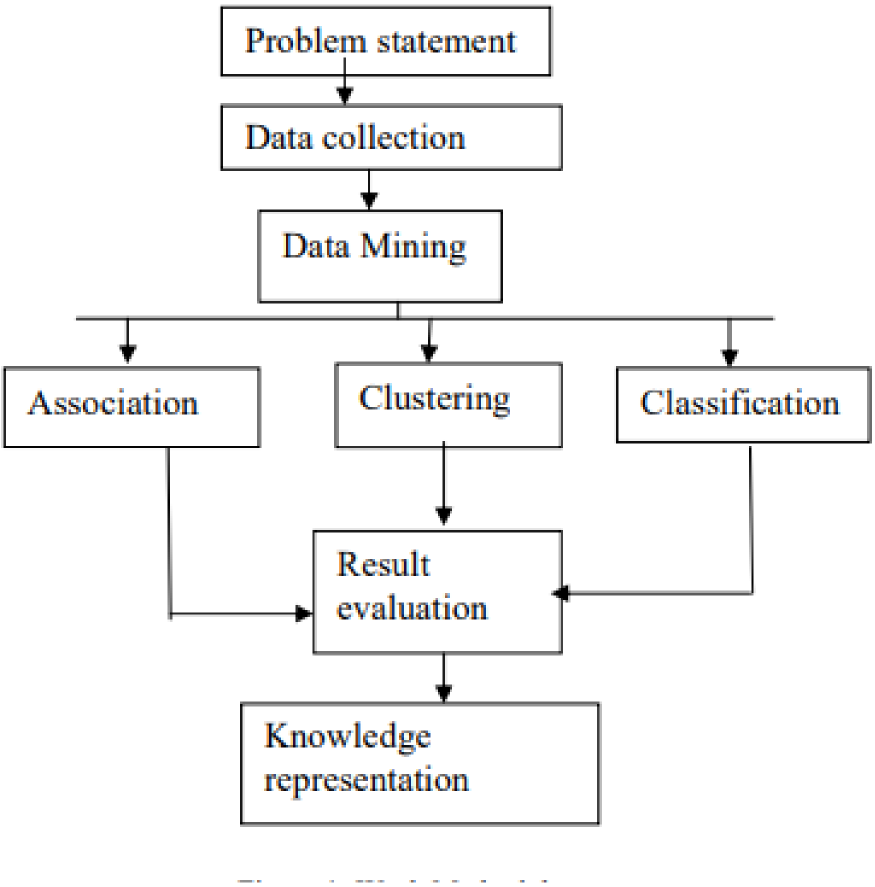


Figure I.6: Work methodology.

## I.5 Machine learning (ML)

The study of algorithms and statistical models that computer systems employ to complete a task without being explicitly programmed is known as machine learning (ML) [19]. The challenge of creating computers that learn automatically through use is addressed by machine learning. It is one of the technical disciplines that is expanding the fastest today [16]. Machine learning has the ability to deduce complicated associations from data, machine learning has consistently shown astounding predictive potential [4].
To learn from the data, machine learning is used. It is applied to train computers how to handle data more effectively. Sometimes, even after viewing the data, we are unable to interpret the information it contains. In that situation, we use machine learning [19].

The Fig I.7 below represents machine learning fields [35].

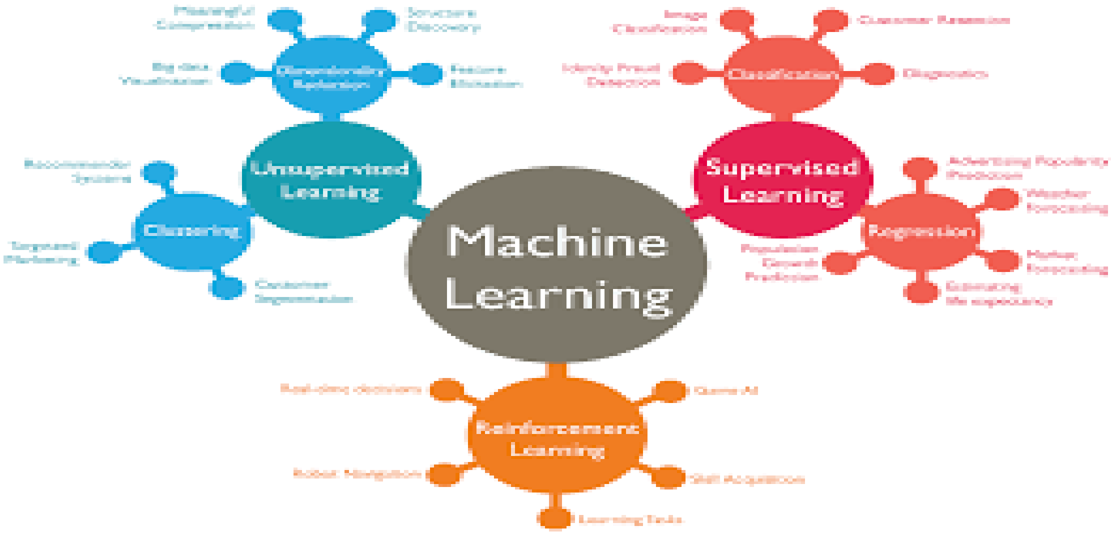


Figure I.7: Machine Learning Fields.

Machine learning is having an important influence on many fields of technology and science; examples of recent success stories include robotics and autonomous vehicle control, voice and natural language processing, neuroscience research, and applications in computer vision [16].

### I.5.1 Supervised learning

Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It develops a function from marked training data made up of a collection of actual practice examples. Algorithms that require outside help are known as supervised machine learning algorithms [19].
The workflow of supervised machine learning algorithms is given in Fig I.8 below [19]:

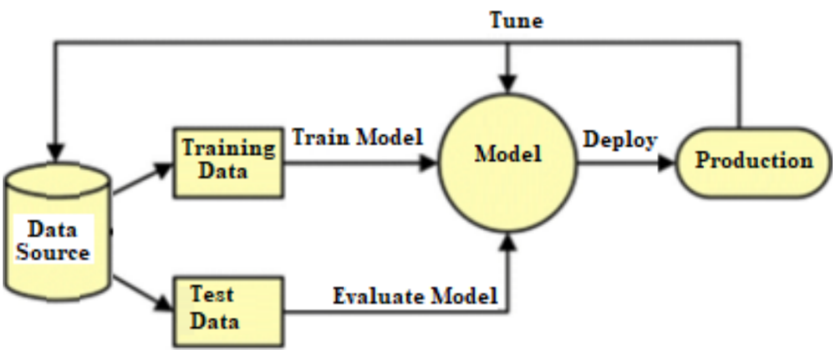


Figure I.8: Supervised Learning Workflow.

### I.5.2 Classification method

Classification is one of the most common application domains of data mining. Before being employed, the classifier model's classification performance is determined on a separate test set. For the purpose of developing a classifier model in classification, there are plenty of alternative techniques and algorithms that can be used. Decision tree algorithms, support vector machines, artificial neural networks, discriminant analysis, logistic regression, Bayesian belief networks, and rule-based systems are a few of the more well-known ones [2].
A well-known machine learning-based data mining method is classification. In simple terms, classification is used to place each item in a set of data into a predetermined class or group.
A rule-based classifying derives a collection of rules that illustrate connections between the class name and the data set's attributes. It classified objects according to a set of IF-THEN rules. While classification rules are prediction rules for characterizing future situations, association rules are characteristic rules that characterize the current situation [29].
The study of Surjeet and Saurabh (2012) [30] confirmed that decision trees produce classification rules which are very easy to intercepts compared to other classification methods and machine learning algorithms.
Decision trees have been heavily utilized by the majority of scholars due to their ease of use and comprehension in revealing little or large data structures and predicting values [21].
Decision trees are particularly successful for supervised learning since they are easier to understand than other types of algorithms [17].

### I.5.3 Decision tree

A decision tree is a tree structure that resembles an organizational diagram, with each internal node represented by a rectangle and each leaf node represented by an oval. There are always at least two internal child nodes. Splits are present in every internal node and are used to evaluate the value of an expression of the attributes. The test's distinct results are marked on the arcs that connect an internal node to its descendant. Each leaf node is assigned a class label [30].
A decision tree is a tree structure that resembles a flowchart, where each core node indicates a test on an attribute, each branch shows the test's result, and leaf nodes signify cases or class distributions of data. These choices produce the categorization rules for a dataset. For every route leading from the root to a leaf, a rule can be written. The class prediction is stored in the leaf node. J48,ID3, CART and others are a few of the decision tree classifiers [27].
A decision tree is a graph that displays options and their outcomes as a tree. The edges of the graph indicate the conditions or rules for making decisions, whereas the nodes in the graph represent an event or a choice. There are nodes and branches in every tree. Each node represents a set of features that needs to be categorized, and each branch indicates a possible value for the node [19].

### I.5.4 The goal of a decision tree

The goal of a decision tree algorithm is to repeatedly split the observations into segments that are incompatible with each other until there is no more division that affects the statistical or quality metrics. The most well-known impurity measures, such as Information Gain,

Gain Ratio, and Gini Index, are used to determine the homogeneity of instances in a node of the tree. Typically, Gini Index is utilized in Classification and Regression Trees (CART), while Information Gain is typically employed in Iterative Dichotomiser (ID3), Gain Ratio in C4.5, and C5.0 (the successors of ID3) [2]. As seen in Fig I.9.

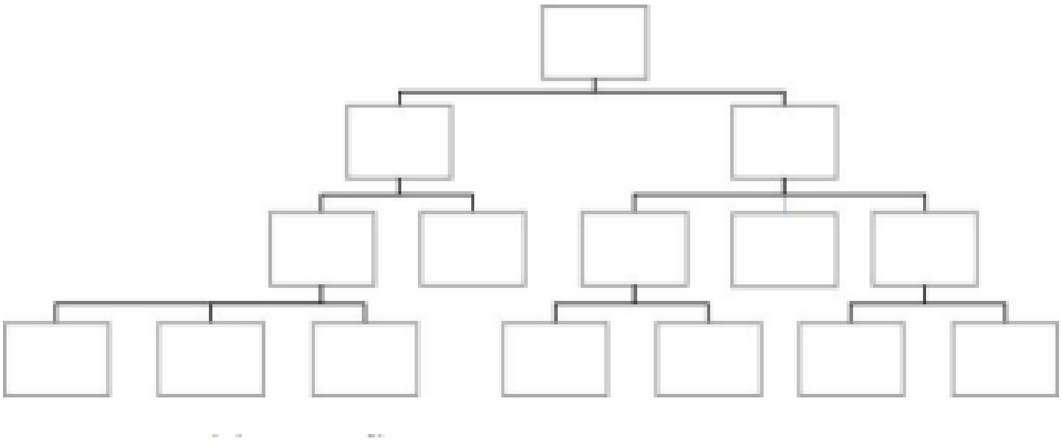

Figure I.9: Decision tree diagram.

Decision trees have the advantage of defining rules that are easily understood and interpreted by users, requiring minimal data preparation, and working well with numerical and categorical variables variables [21]. Decision trees are frequently used to gather information for decision-making [30].

### I.5.5 Algorithms of decision tree

The three widely used decision tree learning algorithms are: ID3, C4.5 and CART [30].

- **ID3 (Iterative Dichotomiser 3)**
  This decision tree algorithm was first presented by Quinlan Ross in 1986. It utilizes Hunt's algorithm. Two phases go into building the tree. Pruning and tree construction are the two stages. ID3 selects the splitting characteristic using the information gain measure. When creating a tree model, only categorical characteristics are accepted. In the presence of noise, it does not produce accurate results. Pre-processing is required to get rid of the noise [30].
  A decision tree induction algorithm is called ID3 - Iterative Dichotomiser 3. J. Ross Quinlan created the ID3 algorithm for the first time at the University of Sydney. Using data, the ID3 algorithm generates decision trees. This approach uses supervised learning and is trained using examples from several classes. The algorithm needs to be capable of predicting the class of a new object after training. ID3 makes advantage of the statistical trait of flexibility to identify which attributes are most essential. The quantity of information in an attribute is measured by sensitivity. The decision tree, which will be used to test upcoming cases, is constructed in this manner [27].

- **C4.5**
  This algorithm is a successor to ID3 developed by Quinlan Ross. It is also based on Hunt's algorithm. Building a decision tree using C4.5 involves handling both categorical and continuous attributes. C4.5 divides the attribute values into two groups based on the chosen threshold such that all values over the threshold are treated as one child and the remaining values are treated as another child in order to manage continuous attributes. Missing attribute values are also handled by it. Gain Ratio is used by C4.5 as an attribute selection metric while creating a decision tree. When there are numerous result values for an attribute, it eliminates the bias in information gain [30].
  J48 - is an open-source Java implementation of the C4.5 algorithm in the Weka data mining tool. This algorithm was developed by Ross Quinlan. Using recursive data partitioning, the program C4.5 builds a decision tree from a set of identified input data. The depth-first approach is used to grow the decision. C4.5 may produce decision trees that can be utilized for classification. The program takes into account every test that may be used to divide the data set and chooses the test that provides the most information gain. Every internal node examines the state of a particular characteristic, and each branch of the tree indicates a study result. The leaves at the tip of the tree's branching define a class to which examples belong [27].

- **CART**
  Breiman's classification and regression trees are referred to as CART. Hunt's algorithm is also the foundation of it. To create a decision tree, CART handles both categorical and continuous attributes. It manages values with errors. To create a decision tree, CART uses the Gini Index as an attribute selection metric. In contrast to ID3 and C4.5, CART generates binary divides. As a result, binary trees are produced. Like ID3, C4.5, the Gini Index metric does not make any probabilistic requirements. To increase accuracy, CART uses cost complexity pruning to eliminate the decision tree's unstable branches [30].
  Simple Cart is a non-parametric decision tree learning method that, depending on whether the dependent variable is categorical or numeric, generates classification trees. Despite being extremely similar to C4.5, CART (Classification and Regression Trees) differs in that it does not construct rule sets and enables numerical target variables (regression). The feature and threshold used by CART to build binary trees are those that produce the most information gain at each node. Without the use of a stopping rule, trees are grown to their maximum size and then, using cost-complexity pruning, are cut back to the root. The split that contributes the least to the overall effectiveness of the tree on training data is the next split to be trimmed. Tree performance is always evaluated using independent test data, and tree selection only takes place following such an assessment [27].

## I.6 Explainability and interpretability

The need of explaining and/or interpreting artificial intelligence (AI) and machine learning systems is well acknowledged [24]. The necessity to create understandable machine learning systems leads to the development of comprehensible and interpretable ML approaches. ones that a human intellect can understand [20].

### I.6.1 Explainability

Explainability loosely refers to any technique that helps the user or developer of ML models understand why models behave the way they do [6]. There is a new line of research in interpretability, which is simply defined as the science of understanding what a model did or might have done, as a first step towards developing explanation mechanisms [11]. Explainability is a general term for technologies that enable a stakeholder to comprehend and, if necessary, question the reasons behind model results [5].

Below are some definitions provided by [5]:
• Explainability gives stakeholders a reduced understanding of how a model function to confirm whether the model serves its original purpose.
• Explainability is for a specific stakeholder in a specific environment with a defined purpose, and it seeks to answer a stakeholder's explanatory needs while bringing the way they think closer to a model's behavior.
• Explainability enables human interaction with ML algorithms to provide superior decisions than either could alone.
The problem of explainability has again come into focus, though the term interpretability [24].

### I.6.2 Interpretability

Doshi-Velez and Kim define interpretability of ML systems as “the ability to explain or to present in understandable terms to a human” [20]. Interpretable machine learning systems offer explanations for their outputs in addition to choices or predictions. By pointing out when the request is appropriate and when it is not, explanations can promote confidence and safety [17].

### I.6.3 The goals of interpretability

A few of the goals that could be achieved with interpretability inspired by [20]:
• **Trust:** trust and interpretability are frequently mentioned together. Knowing "how often a model is right" and "for which examples it is right" are the building blocks of trust.
•**Causality:** When using an interpretable model or an explanation method to develop hypotheses about causal links between variables, causality may be beneficial.
• **Reliability:** according to the reliability basis, ML systems ought to be adaptable to noisy inputs and (acceptable) domain transformations.
• **Fairness:** When ML algorithms are used in social, economic, or medical decision-making, fairness is crucial.
• **Privacy:** privacy can be of concern in systems relying on sensitive personal data. Interpretations and explanations, like fairness, can aid in determining if user privacy is protected.

### I.6.4 Relation between explainability and interpretability

While explanation methods seek to respond to the question, "What else can the model tell me?" interpretability demands, "How does the model work?", [20]. We believe that interpretability is not enough on its own. Black-box methods must be explainable for people to believe them [11] , This is why "Explainable models are interpretable by default, but the reverse is not always true" (the same source). There is "a clear line between interpretable and explainable ML: The goal of interpretable machine learning is to create models that are naturally interpretable, whereas explainable ML aims to offer additional explanations for current black box models, models that are proprietary or incomprehensible to humans" [20].

## I.7 Related work

Narayan Prasad and Subarna Shakya (2022) [7] predicted students results and performance using three decision tree algorithms: ID3, C4.5 (J48), and CART (Classification and Regression Tree) with other classification algorithms: Random Forest (RF), K-nearest Neighbours (KNN), Support Vector Machine (SVM), and Artificial Neural Network (ANN). This study has used two different datasets containing around 1,000 records. This study showed that the algorithm CART has the highest performance and accuracy compared to all other algorithms, while decision tree C4.5 and Random Forest have good accuracy with more execution time.

Muluken Alemu Yehuala (2015) [31] conducted a study on student performance using student data, using a questionnaire comprised of 42 questions distributed to first-year students using data mining techniques such as decision trees, neural networks, and Bayesian networks. The students were from two different academic institutes, and the datasets included student records and their GPA. Weka application software was used to build the model. This study showed that the main attributes considered by different literatures are gender, examination result, number of courses given in a semester, college, department, and number of students in class.

It also offered recommendations to stakeholders in universities to help them make decisions and improve students' performance. The study also confirmed that we can use data mining techniques to determine students' failures and successes.

Surjeet and Saurabh (2012) [30] conducted a study to identify weak students and help them get better marks. They applied the decision tree algorithms C4.5, ID3, and CART to students who were going to fail or succeed and predicted their performance in the final exam. The results showed that the C4.5 had the highest accuracy, with an acceptable level of accuracy for ID3 and CART. The study confirmed that decision trees produce classification rules that are very easy to interpret compared to other classification methods, and machine learning algorithms can produce predictive models to improve students learning.

Hamoud, Alaa, Ali Salah Hashim, and Wid Akeel Awadh (2018) [12] conducted an interesting study about predicting students 'performance in higher education. According to them, this step will help educational stakeholders make good decisions that will be to the benefit of students and that will improve learning in higher education.

They have collected a total of 161 questionnaires about students data after choosing a number of factors that are important to affect students success and failure in order to build a model that can be used by staff and academics to decide which attributes can increase the success of students. They have applied decision tree algorithms such as J48 and Random Tree to the dataset, and they've used Weka tools to construct the model. The results showed that decision tree algorithms are the best solution to predict students success and failure, and attributes such as age, work, gender, stage, and others are the most affected by the

accuracy of the decision tree.
Pallathadka, A. Wenda, E. Ramirez-Ass, et al. (2021) [23] did an interesting comparative using machine learning algorithms such as Nave Bayes, ID3, C4.5, and SVM applied to a dataset that had 33 attributes and 649 instances. The results showed the accuracy of those algorithms; it obviously appeared that SVM had the highest accuracy, about almost 90%, followed by Nave Bayes at about 75%, C4.5 at 70%, and ID3 at about 62%. They've identified the path of a framework for student performance prediction, which started with the dataset student, then the preprocessing of it, which means removing all the noise from the data. After using the classification machine learning algorithms and getting the classification result, they had the prediction of student performance.
In conclusion, they get our intention that the teacher's success is measured by the performance of his students, and the teacher must focus on pupils who need them most.
Mustafa Agaoglu (2016) [2] did an interesting comparison of variables between classifiers by importance values based on the student's perception; four different classification techniques were used, including decision tree algorithms.
The data was collected from departments of Marmara University, Istanbul, Turkey, and consisted of responses of students to a real course evaluation questionnaire using accuracy, precision, recall, and specificity performance metrics. A total of 2850 evaluation scores were obtained; 70% of them were used to train the classifier models, and the other 30% were used to test the data. The confusion matrix was used to show the distribution of predictions for C5.0, CART, SVM, ANN-Q2H, ANN-Q3H, and ANN-M. The results showed that C5.0 was the best in performance according to accuracy by SVM, and CART was the worst. This study also showed the importance of data mining models in course evaluation, and there are many measurement instruments used in instructors 'performance evaluation.

### I.7.1 Main results of related work

The studies mentioned in the related work focus on predicting student performance using various machine learning algorithms such as decision tree algorithms (ID3, C4.5, CART), Random Forest, K-nearest Neighbors, Support Vector Machine, and Artificial Neural Network. The studies used different datasets and attributes, including gender, examination result, number of courses given in a semester, college, department, number of students in class, age, work, stage, and others. The results showed that decision tree algorithms are efficient and produce classification rules that are easy to interpret compared to other classification methods. In particular, CART was found to have the highest performance and accuracy, while C4.5 and Random Forest showed good accuracy with longer execution time. The studies recommended using data mining techniques to determine students' failure and success and providing recommendations to stakeholders in universities to help them make decisions and improve students' performance. The studies also highlighted the importance of machine learning models in course evaluation and the measurement instruments used in instructors' performance evaluation.
As we saw in the related work, there are a lot projects about using machine learning techniques, at the opposite the projects which explain these predictions are not enough.

## I.8 conclusion

Through the interactive learning analytics tool and other channels of distribution, education specialists will have the ability to contribute their knowledge and offer extra explanation and interpretation of learning analytics predictions to stakeholders in higher education.
In the way we look educational stakeholders are having troubles to understand these predictions which can lead to wrong perceptions, and it may create issues in the path of improving learning. Overall, the studies related we are likely to focus on the use of machine learning techniques and interactive learning analytics tools in education and the need for explainability and interpretability in the predictions made by these models.

# Chapter *II*

## Design

# Chapter II

# Design

## II.1 Introduction

This project aims to increase the credibility, reliability, and understanding of learning analytics among higher education stakeholders by integrating knowledge of machine learning models and human experience. By helping education professionals modify, interpret, and evaluate learning analytics predictions.
We chose the decision tree method from the analytical field of supervised machine learning because it gives the best prediction result among the remaining methods because it looks at each data point individually instead of the set as a whole and because of its ease of understanding.
Fig II.1 shows an example of a decision tree that, when viewed from the outside, can't be tracked.

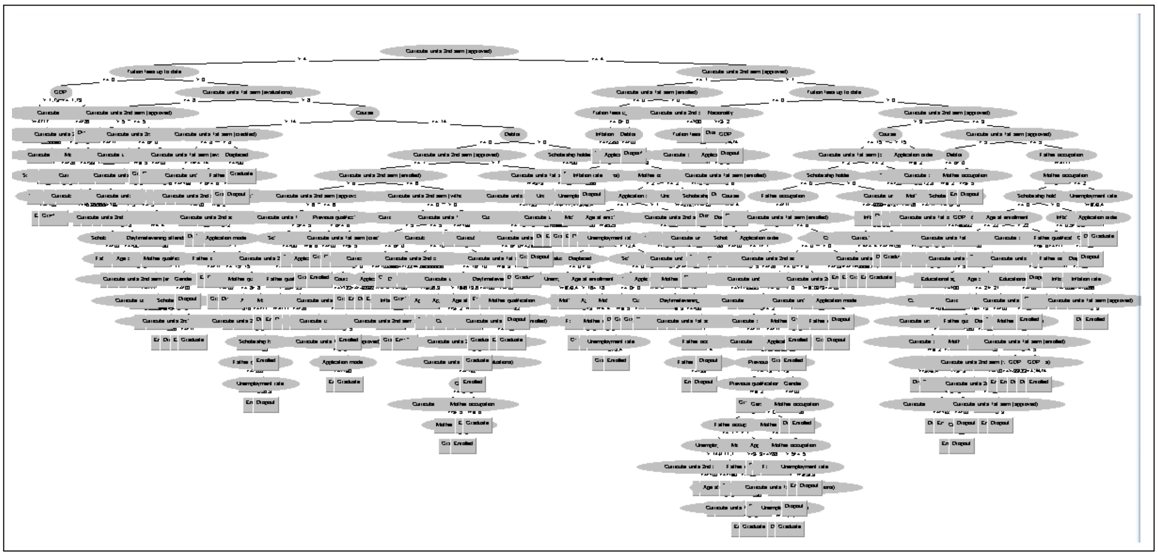

Figure II.1: Decision tree example.

This chapter is composed of three sections and a conclusion. We will first tackle the main issues of a standard decision tree. The sources of the compiled data. The structure of the tool to be made Conclusion of the chapter.

## II.2 Decision tree setbacks

Decision trees have a few issues that discomfort stakeholders and teaching staff.

1. **Inability to show the prediction path:** The decision tree gives an output without showing the path taken.

2. **Inability to parkour in large trees :** The bigger the tree, the harder it is to follow a single path.

3. **Inalterability of the tree :** The tree's structure cannot be modified.

4. **Indivisibility of the tree :** The tree cannot be subdivided to read each node alone.

5. **Inability to be shared :** Only the model's result can be shared between staff, and it has to be generated from scratch each time.

## II.3 Data collection

The data collection for the evaluation is obtained from an open online repository at www.Kaggle.com. This data was used as an example to generate a decision tree model.
This dataset can be used to understand and predict student dropouts and academic outcomes. The data includes a variety of demographic, socio-economic, and academic performance factors related to the students enrolled in higher education institutions. The dataset provides valuable insights into the factors that affect student success and could be used to guide interventions and policies related to student retention..

### II.3.1 Data Description

- **Attributes.**
- **Marital status:** The marital status of the student (Categorical).
- **Application mode:** The method of application used by the student (Categorical).
- **Application order:** The order in which the student applied (Numerical).
- **Course:** The course taken by the student (Categorical).
- **Daytime or evening attendance:** Whether the student attends classes during the day or in the evening (Categorical).
- **Previous qualification:** The qualification obtained by the student before enrolling in higher education (Categorical).
- **Nationality:** The nationality of the student (Categorical)
- **Mother's qualification:** The qualification of the student's mother (Categorical).
- **Father's qualification:** The qualification of the student's father (Categorical).
- **Mother's occupation:** The occupation of the student's mother (Categorical).
- **Father's occupation:** The occupation of the student's father (Categorical).

- **Displaced:** Whether the student is a displaced person (Categorical).
- **Educational special needs:** Whether the student has any special educational needs (Categorical).
- **Debtor:** Whether the student is a debtor (Categorical).
- **Tuition fees up to date:** Whether the student's tuition fees are up to date (Categorical).
- **Gender:** The gender of the student (Categorical).
- **Scholarship holder:** Whether the student is a scholarship holder (Categorical).
- **Age at enrollment:** The age of the student at the time of enrollment (Numerical).
- **International:** Whether the student is an international student (Categorical).
- **Curricular units 1st semester (credited):** The number of curricular units credited by the student in the first semester. (Numerical)
- **Curricular units 1st semester (enrolled):** The number of curricular units enrolled by the student in the first semester. (Numerical)
- **Curricular units 1st semester (evaluations):** The number of curricular units evaluated by the student in the first semester (Numerical).
- **Curricular units 1st semester (approved):** The number of curricular units approved by the student in the first semester. (Numerical)

## II.4 Architectural design

In this part, we will explain how the tool we are about to create works.

**Firstly :**

The entry for this tool will have two options:

- For the new decision tree model, the data set supporting the classification models (features and objectives) will be presented in.csv or .arff format, and from it a decision tree model will be generated.
- For an existing model, upload it as a file.

In the last part of this party, the application will save an instance from the model on the home page.

**Secondly :**

The user can select an existing model in the page home for editing, deleting, or showing propriety.

**Thirdly :**

When it select editing model will be can:

- Show the tree visualizer, select any decision node for splitting the tree, and show only the node selected graph.
- Enable the modification of the tree by: Add node with condition
  - Delete node or delete all child of one node.
  - Modify a condition or a child node.

- The prediction will have two choices: input an instance of the data separated by a comma for all features (direct predict) or enter each feature in the allocated cell in the feature table (manual predict).

  When the prediction finishes, show the class of prediction and the ability to show the path of prediction in the tree (auto Parkour).

- Manual parkour shows the current node, asks for input of the value of the selected feature, and predicts the next node until the leaf node (target).

- Save the model as a file on the PC.

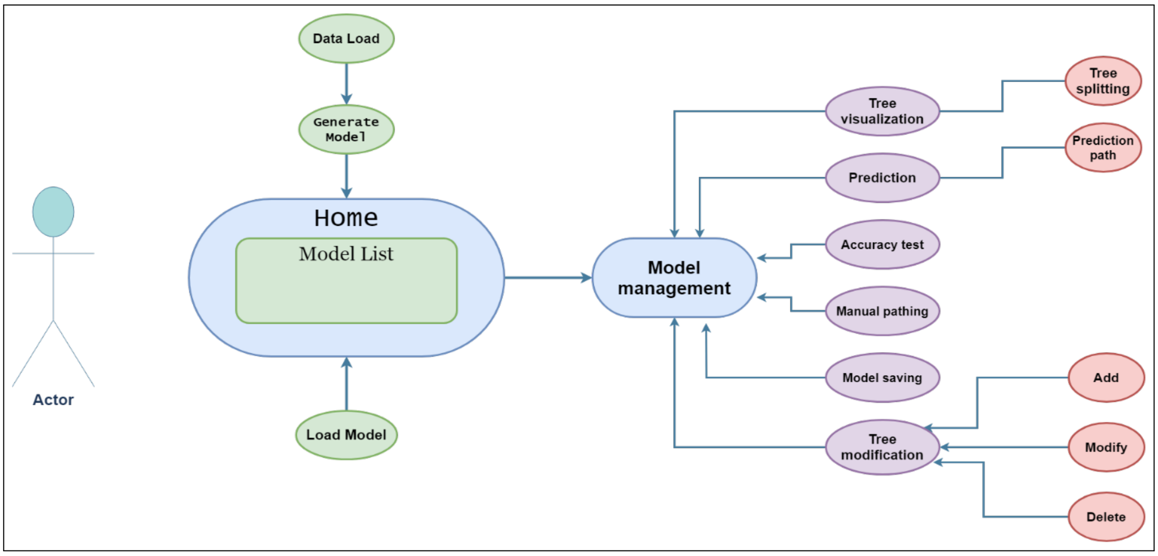


Figure II.2: Model of the learning analytic tool.

## II.5 Conclusion

In this chapter we discussed the setbacks of the standard decision tree and proposed theoretical solutions.

# Chapter $III$

## Realization

# Chapter III

# Realization

In this project we elaborate on the environment and tools used. Showing the results obtained after complete this GP.

## III.1 Environment and tools

In this project we used various tools and technologies to create our desktop application.

### III.1.1 IDEs

**NetBeans:** NetBeans is an integrated development environment for Java. NetBeans allows applications to be developed from a set of modular software components called modules. NetBeans runs on Windows, macOS, Linux and Solaris.[1]

Figure III.1: Logo of NetBeans

[1]For more information, please visit https://netbeans.apache.org/wiki/index.html

### III.1.2 Language

**Java:**Java is a general-purpose computer-programming language that is concurrent, class-based, object-oriented and specifically designed to have as few implementation dependencies as possible.[2]

Figure III.2: Logo of Java

### III.1.3 Technologies

**Machine learning:** Machine learning is s set of algorithm and tools to help and teach computer how to learn and how to take decision based on data.

Figure III.3: Logo of Machine learning.

[2]For more information, please visit https://www.java.com/en/

### III.1.4 Libraries

**Weka:** Weka is a collection of machine learning algorithms for data mining tasks. It contains tools for data preparation, classification, regression, clustering, association rules mining, and visualization.[3]

Figure III.4: Logo of WEKA.

### III.1.5 Library

The pre-existing libraries do not help with explaining the decision tree so we had to make our own based on the J48 algorithm from Weka.

Comprised of two classes:

1. **Node class:**
   It presents the tree's nodes (decision nodes and leaf nodes). Its attributes are:

   - ID : Integer.
   - Label : string.
   - Children : Array<Node>.
   - Isleaf : Boolean .
   - Type : String.
   - Values : Array<String>.

   Its functions are:

   - Constructor : Node (Int id, String label)
   - Getters and setters

[3]For more information, please visit https://www.cs.waikato.ac.nz/ml/weka/#: :text=Weka%20is%20a%20 collection%20 of,association%20rules%20mining%2C%20 and%20 visualization

- AddChild(String edgeLabel, Node child)
- DeleteChild(Node child)
- DeletAllChild()
- ModifieChildEdgeLabel(Node child, String newEdgeLable)
- GetEdge(Node child)
- GetAllNode()

2. **DecisionTree class:**
   Its attributes are:

   - Root : Node
   - Tree : J48
   - Testdata : Instances
   - Name : String
   - Graph : String
   - Data : Instances
   - Modelpath : String
   - SizeId : Integet
   - SizeData : Integer
   - Accuracys : double[ ]

   Its functions are:

   - Constructor : Node () → Load Dataset and generate model.
   - Getters and setters
   - toGraph() : String → model to graph format dotFile
   - selectNode(int id) : Node
   - nextNode(Node CurrentNode , String value) : Node → predict the node
   - predect(String[] input) ; void
   - getVisualiser(TreeDisplayListener L , int id) : TreeVisualiser →show tree
   - saveModel() : void.
   - loadModel() : DecisionTree.

## III.2 Desktop application

We ended up with a desktop application that resolves the aforementioned issues in the second chapter. It consists of:

1. **Home page**
   Shows the list of models the user is currently working on.

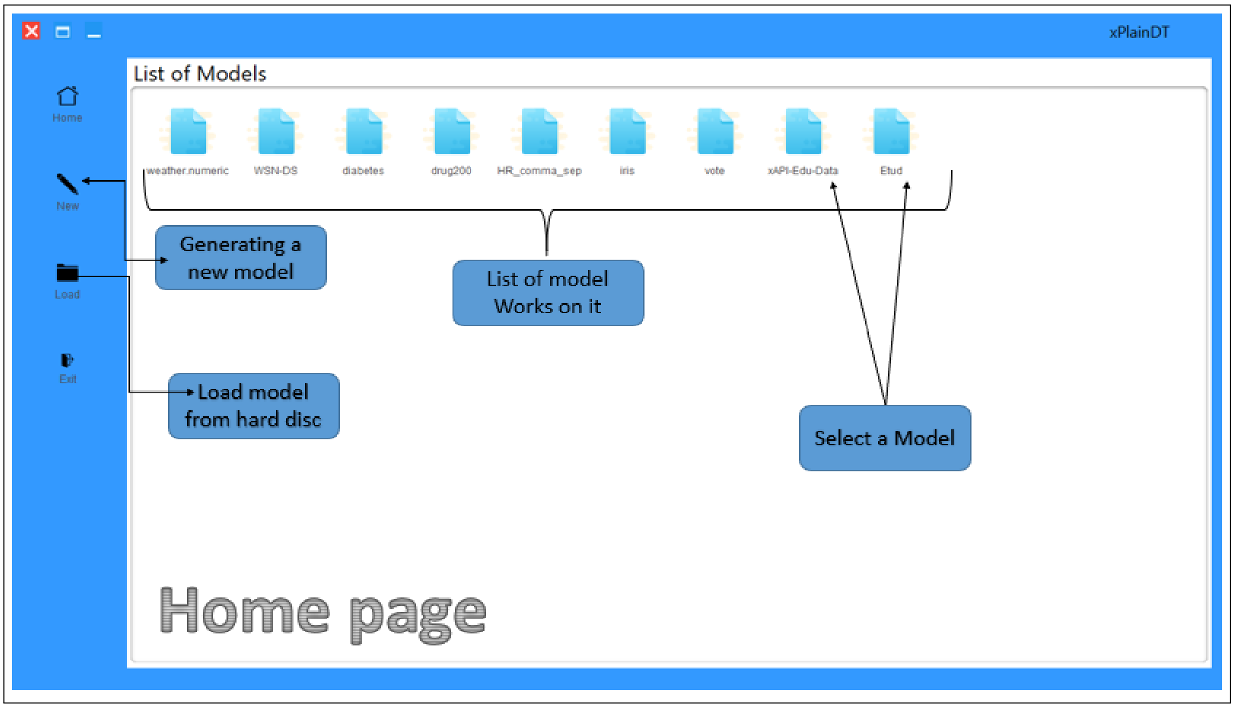


Figure III.5: Home Page.

2. **Model management page**
   Can be accessed through the list of models, after generating a new model, or by loading an existing model from the hard disc.

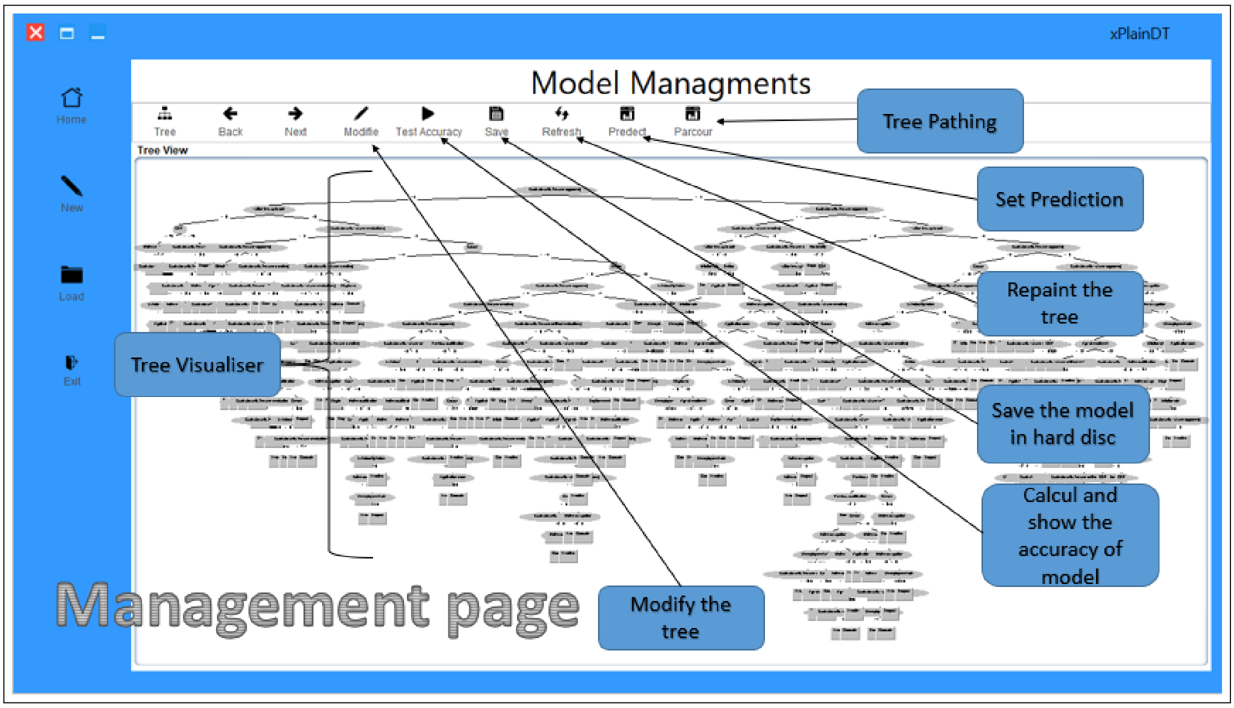


Figure III.6: Management page.

3. **Modify the tree**

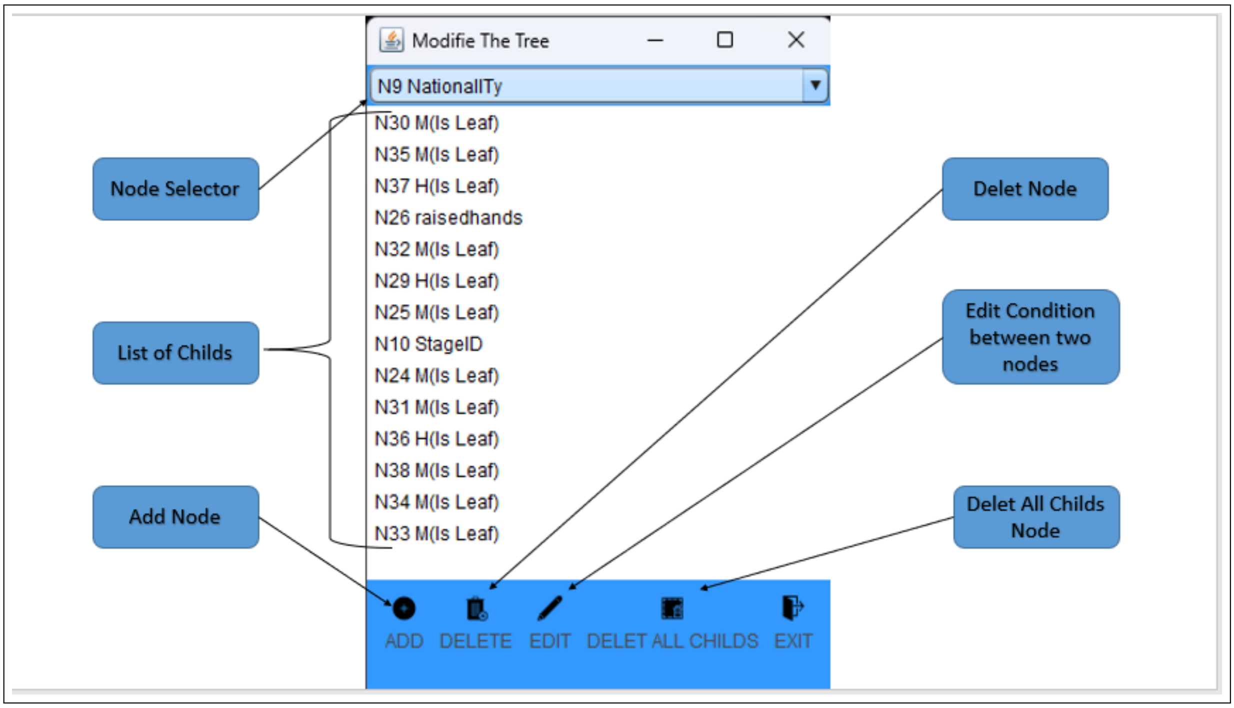


Figure III.7: Modify the Tree.

4. **Accuracy test**

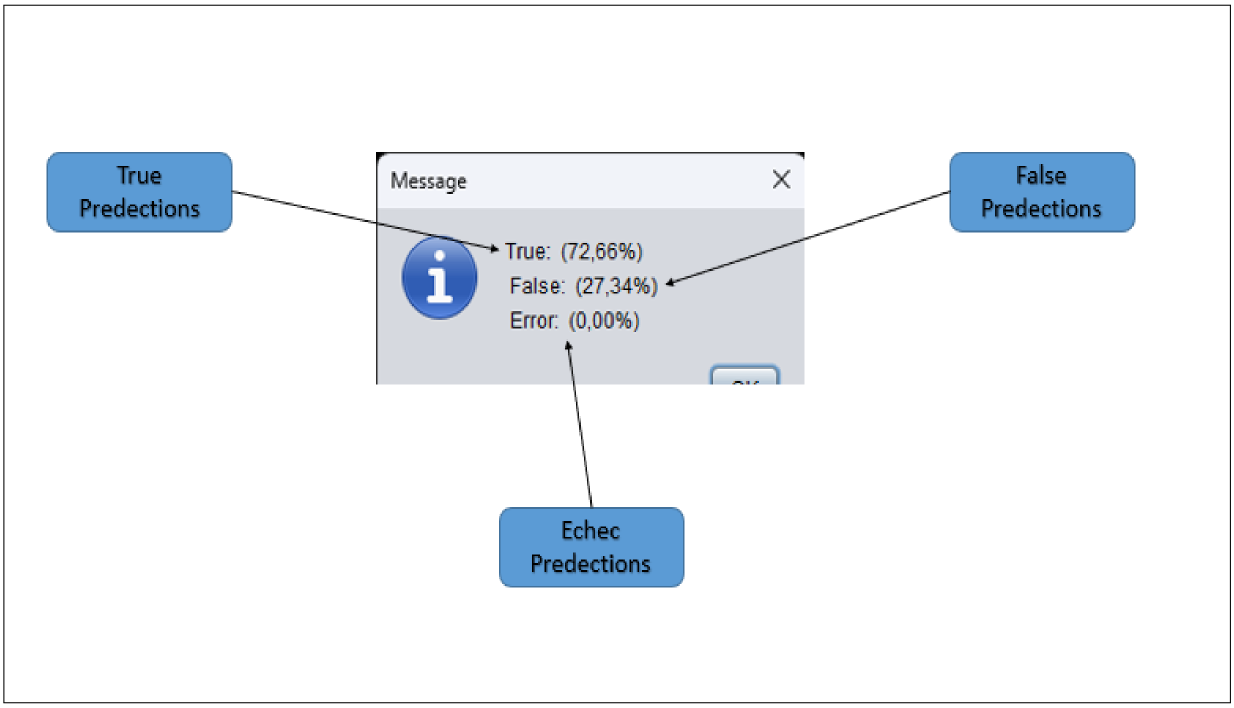


Figure III.8: Test Accuracy.

5. • **1-Predict outcome**
   **a) Manual:**

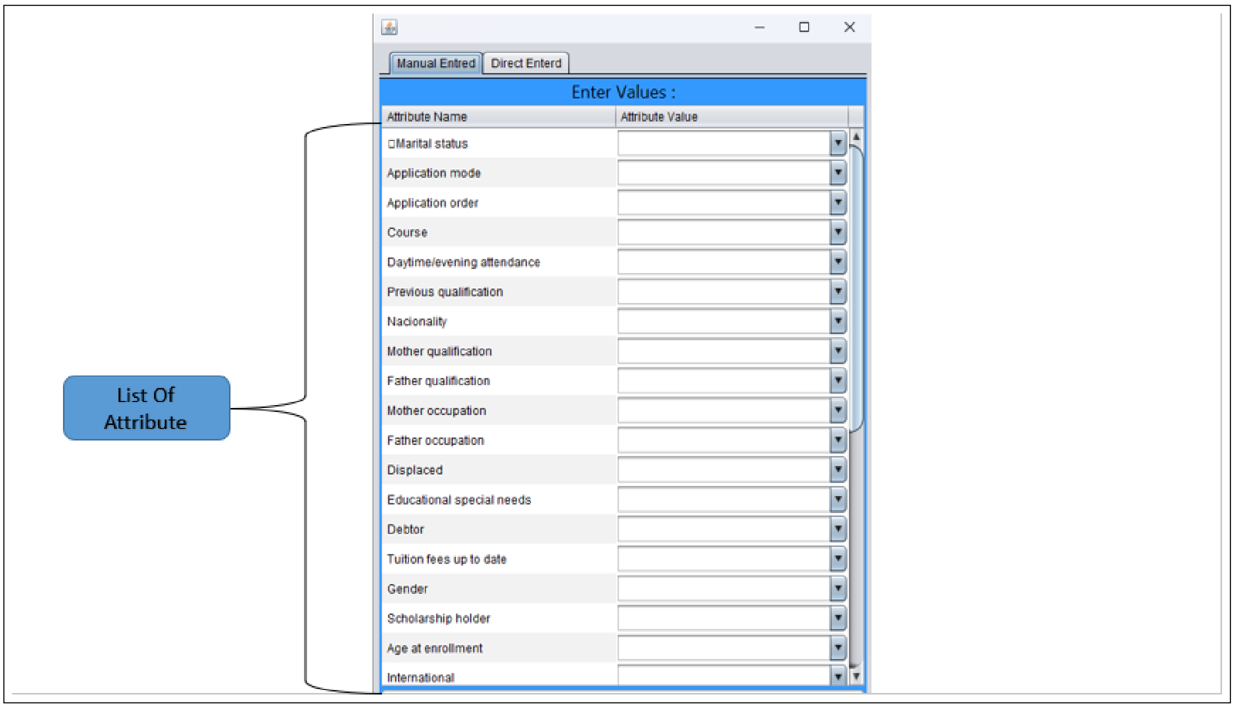


Figure III.9: Manual Predict.

**b) Direct:**

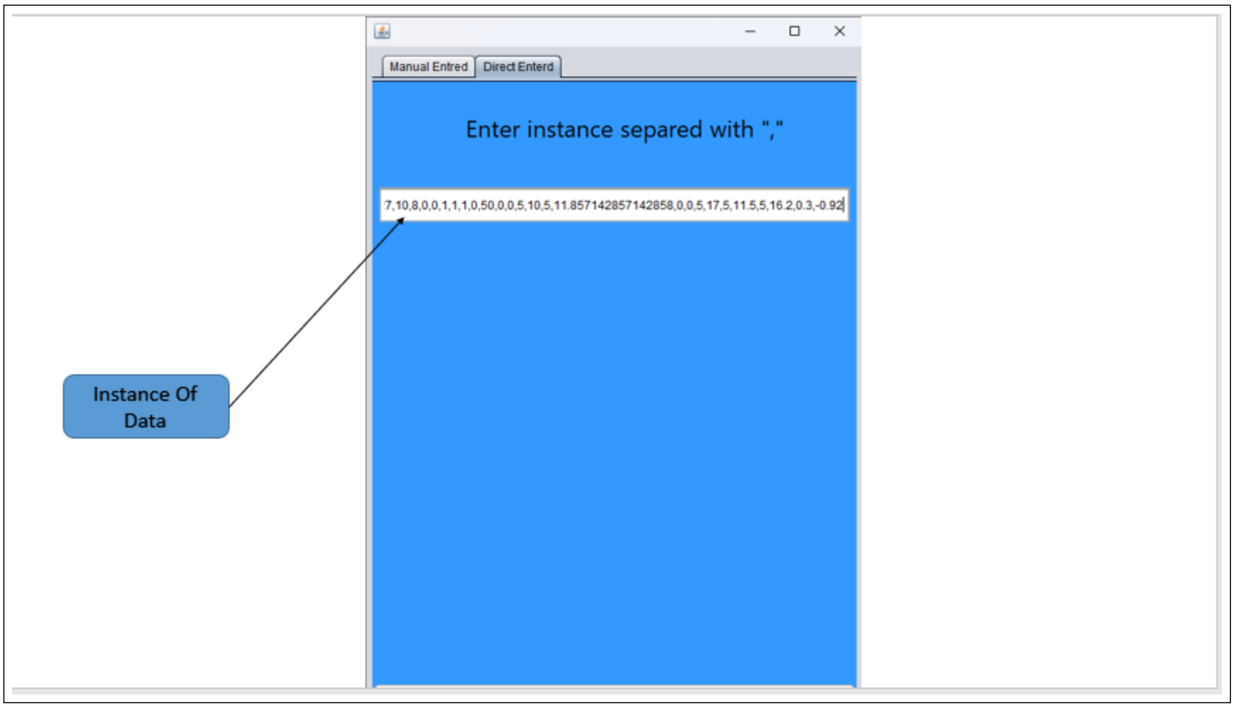


Figure III.10: Direct predict.

- **2-Result of predict:**

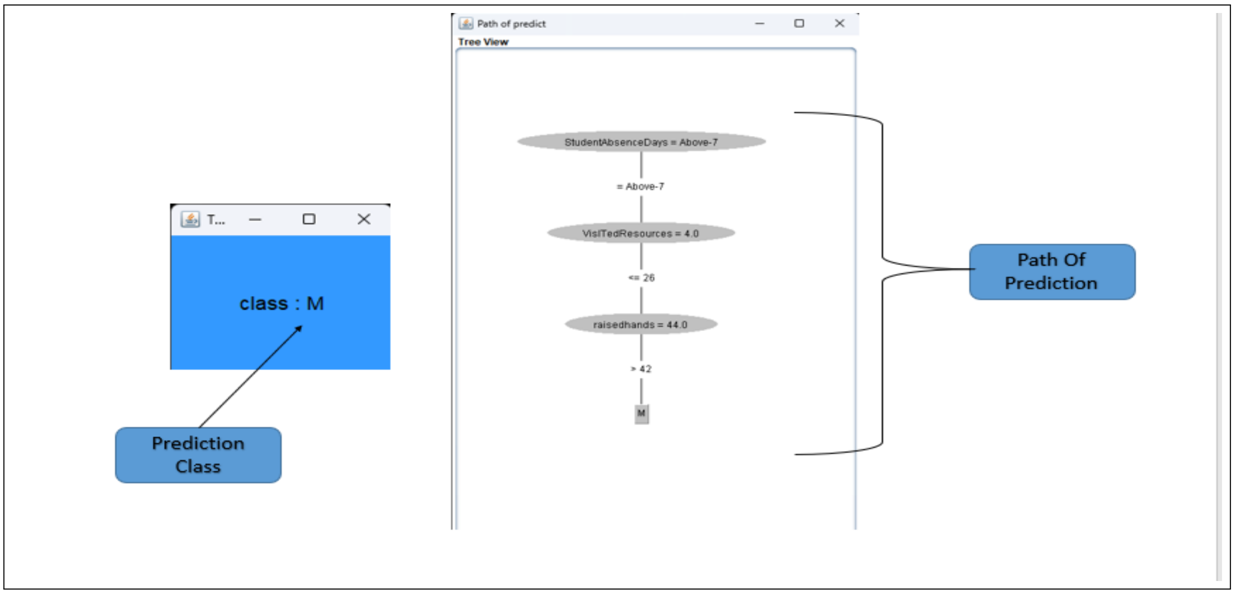


Figure III.11: Result of predict.

6. **Tree pathing:**

It starts from the current node until the last class.

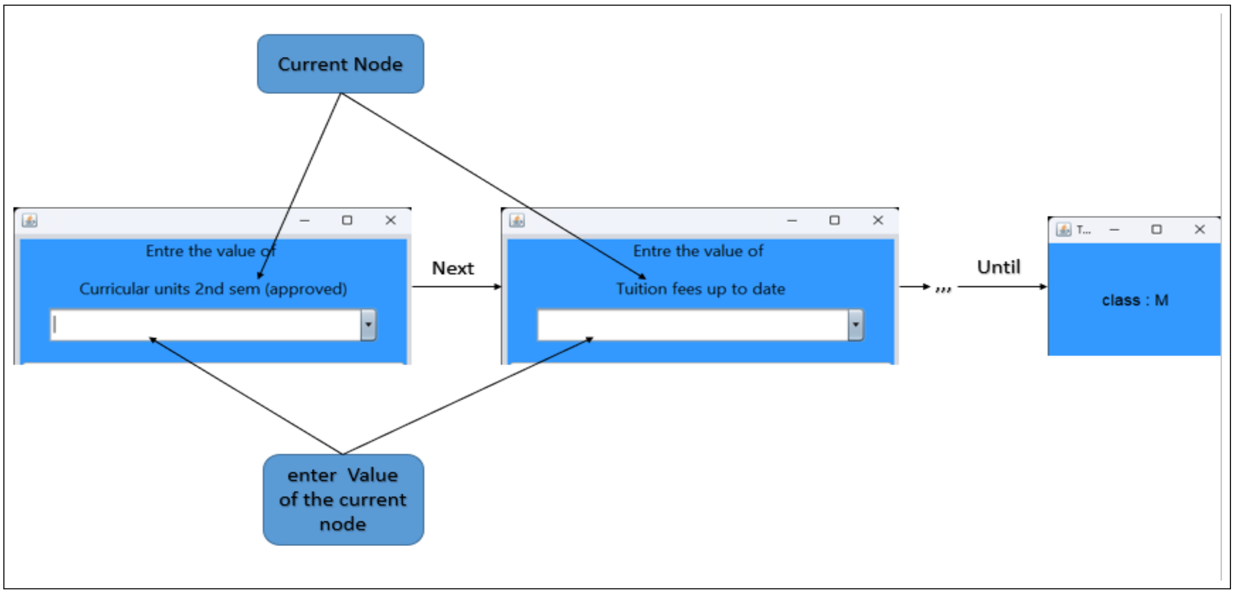


Figure III.12: Path of Tree.

## III.3 Conclusion

This application has solved many issues with decision trees mentioned in the second chapter. While providing better performance.

# General conclusion

In this study, we are interested in making decision trees more comprehensible and editable, allowing for more control over outcomes, more accurate results, and most of all, more trust on the stakeholders' part. We now present the detailed work report and the future perspectives of it.

## Work report

In the first chapter, we emphasise the importance of learning analytics in aiding the educational system through predictions and its shortcomings, such as the inability to edit or access the decision trees.
In the second chapter we discussed the shortcomings of the standard decision tree. Where we got our data from. Suggest a structure for a tool that allows the management of models of decision trees.
In the last chapter we illustrated the open-source tools used to create this application and the private library made to accommodate our needs. We presented our desktop application (Learning Analytic Tools) and how it works.

## Perspective and future work

While the research on developing an interactive learning analytics tool based on decision tree models has many advantages, it also has limitations and room for future work. Some of these limitations and areas for further research include:

1. Only higher education is the focus of the study; consequently, other educational levels may not be applicable to the results.

2. While the study admits that machine learning methods do not explain their predictions, it does not give an exhaustive explanation of these methods, which can limit the comprehension of the results.

3. The study does not offer a comprehensive examination of the interactive learning analytics tool created, which may limit its usefulness in real-world situations.

4. The study does not address any ethical problems with employing machine learning in education, such as potential biases and privacy issues.

Future work of this study could involve:

- The study might look into the possibilities of expanding the tool's usage to a larger population of students and educators and evaluating the tool's effectiveness in real-world educational settings.

- The study could involve a comprehensive evaluation of the effect of the tool on student learning outcomes such as grades, retention rates, and engagement.

- The creation and application of the interactive learning analytics tool may incorporate ethical issues in the future, such as protecting the privacy of user data.

- To address the issue of a lack of confidence in machine learning-based predictions, the study could investigate the possibilities of new machine learning models that are more explainable and transparent in their decision-making processes.

- Expand this study into other fields of machine learning such as random forest and neural network

## Conclusion

The application of learning analytics and decision tree models in education has the potential to significantly improve student learning outcomes and provide important data to educators and other stakeholders. With the help of the interactive learning analytics tool developed in this work, educators will have a potential way to understand the model's decision-making process and give more explanations and interpretations to stakeholders. The effectiveness of the tool lies in its ability to display true, false, and wrong predictions, which enables it to evaluate the performance of students and identify areas where they may need additional support.

# Bibliography


[1] Abazeed, A., & Khder, M. (2017). A Classification and Prediction Model for Student's Performance in University Level. J. Comput. Sci., 13(7), 228-233.

[2] Agaoglu, M. (2016). Predicting instructor performance using data mining techniques in higher education. Ieee Access, 4, 2379-2387.

[3] Baker, E., & McGaw, B. (2010). International encyclopedia of education. Elsevier Academic.

[4] Begley, T., Schwedes, T., Frye, C., & Feige, I. (2020). Explainability for fair machine learning. arXiv preprint arXiv:2010.07389.

[5] Bhatt, U., Andrus, M., Weller, A., & Xiang, A. (2020). Machine learning explainability for external stakeholders. arXiv preprint arXiv:2007.05408.

[6] Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M., & Eckersley, P. (2020). Explainable machine learning in deployment. Proceedings of the 2020 conference on fairness, accountability, and transparency,

[7] Dahal, N. P., & Shakya, S. (2022). A Comparative Analysis of Prediction of Student Results Using Decision Trees and Random Forest. Journal of Trends in Computer Science and Smart Technology, 4(3), 113-125.

[8] Dietz-Uhler, B., & Hurn, J. E. (2013). Using learning analytics to predict (and improve) student success: A faculty perspective. Journal of interactive online learning, 12(1), 17-26.

[9] Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. Ieee Access, 5, 15991-16005.

[10] Elias, T. (2011). Learning analytics. Learning, 1-22.

[11] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA),

[12] Hamoud, A., Hashim, A. S., & Awadh, W. A. (2018). Predicting student performance in higher education institutions using decision tree analysis. International Journal of Interactive Multimedia and Artificial Intelligence, 5, 26-31.

[13] Han, J., Kamber, M., & Pei, J. (2012). Data mining concepts and techniques third edition. University of Illinois at Urbana-Champaign Micheline Kamber Jian Pei Simon Fraser University.

[14] Hasan, R., Palaniappan, S., Mahmood, S., Abbas, A., Sarker, K. U., & Sattar, M. U. (2020). Predicting student performance in higher educational institutions using video learning analytics and data mining techniques. Applied Sciences, 10(11), 3894.

[15] Hooda, M., Rana, C., Dahiya, O., Shet, J. P., & Singh, B. K. (2022). Integrating LA and EDM for improving students Success in higher Education using FCN algorithm. Mathematical Problems in Engineering, 2022.

[16] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

[17] Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S., & Doshi-Velez, F. (2019). An evaluation of the human-interpretability of explanation. arXiv preprint arXiv:1902.00006.

[18] Leitner, P., Khalil, M., & Ebner, M. (2017). Learning analytics in higher education—a literature review. Learning analytics: Fundaments, applications, and trends: A view of the current state of the art to enhance E-learning, 1-23.

[19] Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9, 381-386.

[20] Marcinkevičs, R., & Vogt, J. E. (2020). Interpretability and explainability: A machine learning zoo mini-tour. arXiv preprint arXiv:2012.01805.

[21] Ofori, F., Maina, E., & Gitonga, R. (2020). Using machine learning algorithms to predict students' performance and improve learning outcome: A literature based review. Journal of Information and Technology, 4(1), 33-55.

[22] Ömer, A., Emily, R., & MITCHELL, L. (2015). Stakeholders in US higher education: An analysis through two theories of stakeholders. Bilgi Ekonomisi ve Yönetimi Dergisi, 10(2).

[23] Pallathadka, H., Wenda, A., Ramirez-Asís, E., Asís-López, M., Flores-Albornoz, J., & Phasinam, K. (2023). Classification and prediction of student performance data using various machine learning algorithms. Materials today: proceedings, 80, 3782-3785.

[24] Preece, A., Harborne, D., Braines, D., Tomsett, R., & Chakraborty, S. (2018). Stakeholders in explainable AI. arXiv preprint arXiv:1810.00184.

[25] Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting Student Dropout and Academic Success. Data, 7(11), 146.

[26] Romero, C., & Ventura, S. (2010). Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (applications and reviews), 40(6), 601-618.

[27] Ruby, J., & David, K. (2014). Predicting the Performance of Students in Higher Education Using Data Mining Classification Algorithms-A Case Study. Int. J. Res. Appl. Sci. Eng. Technol, 2, 173-180.

[28] Silwattananusarn, T., & Tuamsuk, K. (2012). Data mining and its applications for knowledge management: a literature review from 2007 to 2012. arXiv preprint arXiv:1210.2872.

[29] Singh, R., Tiwari, M., & Vimal, N. (2013). An empirical study of applications of data mining techniques for predicting student performance in higher education. Int. J. Comput. Sci. Mob. Comput, 2, 53-57.

[30] Yadav, S. K., & Pal, S. (2012). Data mining: A prediction for performance improvement of engineering students using classification. arXiv preprint arXiv:1203.3832.

[31] Yehuala, M. A. (2015). Application of data mining techniques for student success and failure prediction (The case of Debre Markos university). International journal of scientific & technology research, 4(4), 91-94.

[32] https : //www.unesco.org/en/higher-education/need-know.

[33] https : //www.lisedunetwork.com/higher-education/.

[34] https: //en.wikipedia.org/wiki/Data_mining.

[35] https://www.wordstream.com/blog/ws/2017/07/28/machine-learning-applications.

[36] https://educationaldatamining.org/.

[37] https://study.com/academy/lesson/what-is-a-stakeholder-in-education-definition-examples.html.