STUDENT ACADEMIC PERFORMANCE PREDICTION VIA ARTIFICIAL INTELLIGENCE USING MACHINE LEARNING ALGORITHMS

HATİCE NAZLI BASTEM

NOVEMBER 2021

ÇANKAYA UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

DEPARTMENT OF COMPUTER ENGINEERING

MASTER’S THESIS

STUDENT ACADEMIC PERFORMANCE PREDICTION VIA ARTIFICIAL INTELLIGENCE USING MACHINE LEARNING ALGORITHMS

HATİCE NAZLI BASTEM

NOVEMBER 2021

ABSTRACT

STUDENT ACADEMIC PERFORMANCE PREDICTION VIA ARTIFICIAL INTELLIGENCE USING MACHINE LEARNING ALGORITHMS

BASTEM, Hatice Nazlı
M.Sc., Computer Engineering Department
Supervisor: Assist. Prof. Dr. Roya CHOUPANI

NOV 2021, 72 pages.

The academic success of students in schools is valued by both students, teachers, and families. For this reason, performance prediction plays a significant role in students' life. With performance prediction, unsuccessful students can be directed to increase their success, study programs can be created, supportive course resources can be suggested, or elective courses can be selected. In this study, the academic success of the student can be predicted with machine learning methods. This study made use of dataset collected from student's knowledge from two schools in Portugal from Kaggle web site. We used three different algorithms to make performance prediction. These are Decision Tree, Random Forest and Logistic Regression. 30% of the dataset was used as test data. The remaining 70% data was used as training data. By using three algorithms, the confusion matrix, accuracy, recall, precision and auc values are obtained. It has been concluded that which algorithm is more successful on which amount of data. decision tree algorithm gives the best accuracy rate with max depth 2 value with 649 student data. The random forest algorithm gives the best accuracy with 649 student data. The logistic regression algorithm gives the best accuracy with 395 student data.

Keywords: Performance Prediction, Machine Learning, Accuracy, Algorithm, Decision Tree, Random Forest, Logistic Regression.

ÖZ

MAKİNE ÖĞRENMESİ ALGORİTMALARINI KULLANARAK YAPAY ZEKA YOLUYLA ÖĞRENCİ AKADEMİK PERFORMANS TAHMİNİ

BASTEM, Hatice Nazlı

Yüksek Lisans, Bilgisayar Mühendisliği Bölümü

Tez Danışman: Dr. Öğr. Üyesi Roya CHOUPANI

Kasım 2021, 72 sayfa

Öğrencilerin okuldaki akademik başarıları hem öğrenciler hem öğretmenler hem de aileler tarafından önemsenmektedir. Bu nedenle performans tahmini, öğrencinin yaşamında önemli bir rol oynamaktadır. Performans tahmini ile başarısız öğrenciler başarılarını artırmaya yönlendirilebilir, çalışma programları oluşturulabilir, destekleyici ders kaynakları önerilebilir veya seçmeli dersler seçilebilir. Bu çalışmada öğrencinin akademik başarısı makine öğrenmesi yöntemleri ile tahmin edilebilmektedir. Bu çalışmada, Kaggle web sitesinden Portekiz'deki iki okuldan öğrencilerin bilgilerinden toplanan veri seti kullanılmıştır. Performans tahmini yapmak için üç farklı algoritma kullandık. Bunlar Karar Ağacı, Rastgele Orman ve Lojistik Regresyondur. Veri setinin %30'u test verisi olarak kullanılmıştır. Kalan %70'lik veri ise eğitim verisi olarak kullanılmıştır. Üç algoritma kullanılarak, karışıklık matrisi, doğruluk, geri çağırma, kesinlik ve auc değerleri elde edilir. Hangi algoritmanın hangi miktarda veri üzerinde daha başarılı olduğu sonucuna varılmıştır. Karar ağacı algoritması, 649 öğrenci verisi için maksimum derinlik 2 değeri ile en iyi doğruluk oranını verir. Rastgele orman algoritması, 649 öğrenci verisi ile en iyi doğruluğu verir. Lojistik regresyon algoritması, 395 öğrenci verisi ile en iyi doğruluğu verir.

Anahtar Kelimeler: Performans Tahminleme, Makine Öğrenmesi, Doğruluk, Algoritma, Karar Ağacı, Rassal Orman, Lojistik Regresyon.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my dear advisor, Asst. Prof Dr. Roya CHOUPANI, who provided me with all kinds of help and sacrifice during my studies and expanded my horizons in academic studies.

I would like to thank Asst. Prof Dr. Abdül Kadir GÖRÜR cordially for his valuable comments, continuous guidance and counseling.

I wish to thank the examining committee for their kindness during the presentation of this thesis.

I must express my profound sincere and gratitude to my mother, Fatma Nüket SÜER, and my father, Murat KUŞ, who have brought me to where I am today and have always supported me without sparing their efforts.

It is a pleasure to express my special thanks to my spiritual sister Zeynep BİÇER for her encouragement, help and continuous support.

Finally, my sincere acknowledgement also goes to my husband, Tevfik Uğur BASTEM, who stood by me unconditionally during my most stressful times and gave me support and strength with his presence.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ML	**:**Machine Learning
AI	**:**Artificial Intelligence
ANN	**:**Artificial Neural Networks
SVM	**:**Support Vector Machines
kNN	**:**K-Nearest Neighbor Algorithm
DT	**:**Decision Tree
LR	**:**Logistic Regression
RF	**:**Random Forest
TN	**:**True Negative
TP	**:**True Positive
FN	**:**False Negative
FP	**:**False Positive
ROC	**:**Receiver Operating Characteristics
AUC	**:**Area Under The Curve
etc	**:**Et Cetera, Other Similar Things

CHAPTER 1
INTRODUCTION

In today's world, technology is well advanced. Like many situations, the development of technology has provided many advantages and disadvantages for people. The important thing is that people take advantage of technology. For instance, it was very difficult to access information in the old times. Information was obtained from many books by reading line by line. However, now, we can quickly access information via the internet in just a few minutes. In another situation, student records at the school were written into the registry books throughout the days. Later, these registry books were searched for hours to get information about a student. On the other hand, nowadays, all student records are kept on computers and when we need any student's information, we can get that knowledge in a short time via technology. These examples are just some of the advantages of technology. Of course, when all the data like the data I mentioned in these examples are considered, the first thing that comes to mind is the concept of big data. Thanks to the use of information systems, these data can be stored and managed easily. However, the transformation of these data into meaningful information is just as important as its storage and management. When the data are scarce, it is very easy to extract the necessary information from this data, to obtain new information, to make decisions, to make definitions and to generate predictions. However, it is very difficult for the human brain to perform these operations on big data. People have a hard time analyzing and understanding this massive data. At this point, the concept of artificial intelligence, machine learning and data mining enter our lives.

In the study, based on the student performance in education, how artificial intelligence, machine learning and data mining techniques can be used in performance prediction has been examined in detail, supported by studies in the literature, and the algorithms have been evaluated in line with the findings.

1.1. PROBLEM STATEMENT

The academic success of students in schools is valued by both students, teachers, and families. Every family is ready to do whatever it takes to make their child successful. But sometimes, the academic success of students does not increase due to unpredictable results. For this reason, performance prediction plays a significant role in students' academic life. After encountering students' performance problems, bringing delayed solutions (reactive approaches) to these problems may damage students and it may be late for solutions. For this reason, generating solutions with predictive information (proactive approaches) can prevent these problems. With performance prediction, unsuccessful students can be directed to increase their success, study programs can be created, supportive course resources can be suggested, or elective courses can be selected.

The academic success of the student can be predicted with data mining and machine learning methods. Huge amounts of information are used to make this prediction. With this huge information, it is difficult for the human brain to analyze and make the right decision for student performance. For this reason, machine learning and data mining techniques are frequently used in analyzing data and extracting high-level information. By using the students' various information, performance prediction can be realized with less error rate than previous studies. Accordingly, "Student Academic Performance Prediction Via Artificial Intelligence Using Machine Learning Algorithms" was chosen as the research subject.

1.2. AIM OF THE THESIS

The aim of this thesis is to predict the academic success of students by using machine learning methods based on classification with factors that affect the education process. Thus, it can be ensured that students tend to the areas where they will be successful. As I mentioned in the Literature Review, these studies made estimates using machine learning methods. In general, more than one method was tried and the most successful algorithm was determined. All of these studies cover student-related issues such as prediction of the students' academic performance, students' graduation grades, the academic performance of the first year, course success of high school students, the test scores of students, student's exam results, learner progression, the student's performance at the final examination and, at-risk students for success in school life. As dataset in these studies, demographic's, educational and personality information were used. Gender, age, race, educational status of mother and father, secondary school information can be given as an example to the demographic’s information. Computer skill, name of courses taken, high school GPA, the student's year-end success average, absenteeism information, lab work, class performance, attendance, assignment, sessional performance, grade of mathematics tests can be given as an example to the educational information. Anxiety, exhaustion, academic motivation, general self-efficacy, academic self-efficacy, interest, reading level, aptitude, personality and motivation learning strategies can be given as an example to the personality information. Predictions have been made with many data mining methods by using the data types mentioned in these studies. For instance, artificial neural network, decision tree, multilayer perceptron, random forests, k-nearest neighborhood, naive bayes classifier, C4.5 decision tree, logistic regression and support vector machine were used for prediction.

In our thesis, we will predict the academic performance of students by using data such as demographic, educational and personality information and many machine learning algorithms as in other studies. Unlike other studies, we aim to predict academic success by finding the most successful algorithm with less error rate.

1.3. THESIS ORGANIZATION

In the first part of the thesis, basic information, problem situation and the purpose of the thesis are given in order to gain a general point of view to the thesis.

The organization of other departments is presented below:

In Chapter 2, the concepts of artificial intelligence, machine learning and algorithms are defined and each component they have is explained.

In Chapter 3, the literature related to the thesis has been examined.

In Chapter 4, prediction processes have been defined.

In Chapter 5, performance prediction of students has been predicted.

In Chapter 6, experimental setup has been explained and the results of the thesis are discussed.

In Chapter 7, whether the research has achieved its purpose and the results are explained. In addition, forward-looking application areas and suggestions were discussed.

CHAPTER 2
BACKGROUND

Many techniques have been developed for the analysis of big data on the internet especially. These techniques are generalized as data mining and can be grouped under the heading of data mining as machine learning algorithms. Data mining is applied to extract information from large amounts of data. On the other hand, machine learning is a field related to artificial intelligence that aims to make a machine perform a task on its own and with the best performance. In other words, machine learning is a technique that has been developed from the field of artificial intelligence and forms the basis of data mining. This section includes fundamental information about artificial intelligence, machine learning and data mining.

2.1. ARTIFICIAL INTELLIGENCE

In 1950, Alan Turing stated in an article he published whether machines could think or not. Thus, the subject of artificial intelligence was raised. But the real name father of artificial intelligence is John McCarthy, who held an academic conference on the subject in 1956.

A wide variety of quotes have been made by many scientists about artificial intelligence. One of the most remarkable of these quotes is by Edward Fredkin. He says; "There are three major events in history. The first is the formation of the universe, the second is the beginning of life, and the third is the emergence of artificial intelligence."[1]

A few definitions that answer the question of what is artificial intelligence are given below.

- "The art of creating machines that perform functions that require intelligence when performed by people" (Kurzweil, 1990) [2]

- "The branch of computer science that is concerned with the automation of intelligent behavior" (Luger and Stubblefield, 1993) [3]
- According to Russel & Norving, artificial intelligence definitions can be examined in four categories. These are; “Systems that think like humans. Systems that act like humans. Systems that think rationally. Systems that act rationally” [4]

In other words, AI is the artificialization of natural intelligence and transferring it to machines through various methods of simulation.

In the 1950s and 60s, artificial intelligence studies were on puzzle solving, playing chess and checkers, proving theorems, answering simple questions, and image classification. Today's artificial intelligence studies are about resembling some of human cognitive abilities and even doing better than humans do.[5]

Artificial intelligence studies are based on the human brain; In addition to giving products in different areas of daily life, it is also used for purposes such as prediction, classification, clustering.

In our daily life, an event can be interpreted by different people in different ways and different inferences can be made from the event. Different technologies came to the fore when computers were asked to interpret events like humans and to make inferences from the event. Therefore, nowadays, there are more than 60 artificial intelligence technologies. Some of these are mentioned below.

- Expert Systems
- Fuzzy Logic
- Genetic Algorithms
- Artificial Neural Network and Machine Learning [6]

Artificial neural networks from these techniques are networks that operate according to the working principles of nerve cell networks in the biological central nervous system. [7]

Machine learning, which also includes artificial neural networks, is shown as an artificial intelligence technique. It can even be considered the most popular artificial intelligence technique.

First, I will briefly mention Expert Systems, Fuzzy Logic and Genetic Algorithms. Then, I will mention Machine Learning, which is the main subject of the thesis, in detail under the title 2.2 Machine Learning.

2.1.1 Expert Systems

Expert systems, as the name indicates, are systems that find a solution just as experts in a subject find a solution to a problem on this subject. Expert systems use their knowledge to get a job done, make a decision, and solve a problem. [8]

There are four components of an expert system. These components are; user interface and inference engine, knowledge base and working memory. I will explain a few components, but you can see all the components in Figure1, which belongs to Rychener. Tasks such as obtaining information, debugging the knowledge base, testing the knowledge base, running test cases, drawing summary results, explaining the steps leading to a conclusion and evaluating the performance of the system are performed with the user interface. At the center of the diagram, the main computing engine searches the knowledge base for information and makes inferences according to the problem. [9]

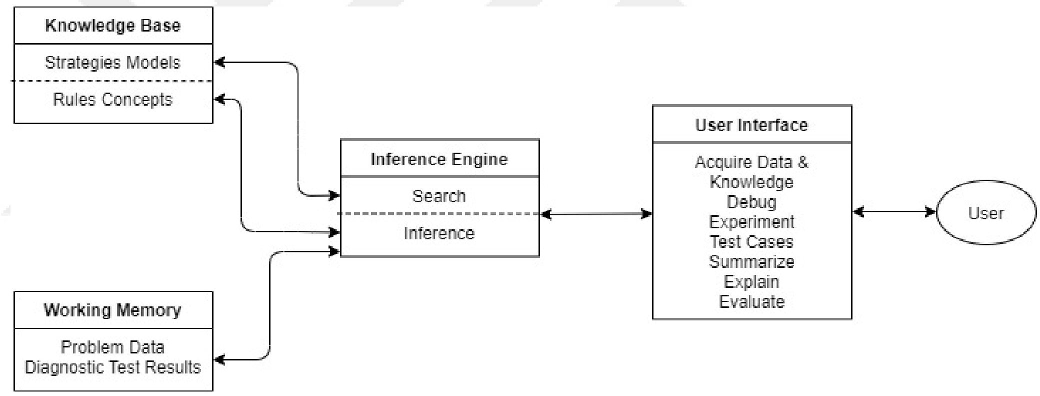


Figure 1 Components of an expert system

In short, expert systems search the suitable knowledge for the problem encountered in the knowledge base and make the most appropriate inferences for the user to reach her goal.

2.1.2 Fuzzy Logic

The concept of fuzzy logic came into our lives in 1965 with the article titled "Fuzzy Sets" written by Prof. Lotfi Askar Zadeh. In his article, Zadeh talked about the fuzzy set theory that sets with imprecise boundaries are formed. [10] The work in question is important because it challenges Aristotelian logic. In classical logic, clearly separated intervals are used, but not in fuzzy logic. Fuzzy logic uses multiple ranges nested with defined functions, rather than intervals separated by strict lines. Fuzzy sets theory is the approach of creating systems that can think, make decisions, generate solutions to problems, and make choices like a human.

Fuzzy logic is a mathematical discipline based on fuzzy set theory. Fuzzy logic works according to intermediate values such as too long-long-medium-short-too short, temperature-less cold-Cold-too cold, instead of long-short, hot-cold, Fast-Slow, black-white.

The working structure of the fuzzy logic system is given in Figure 2 [11].

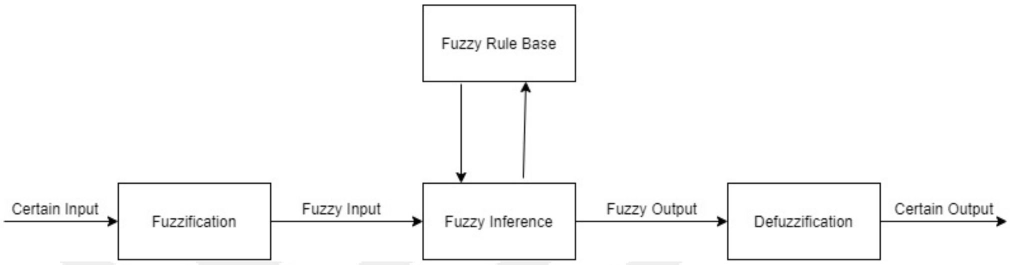


Figure 2 Working structure of the fuzzy logic system

First of all, definite inputs are evaluated with functions and blurred. Then, inference is made according to the chosen inference method and using the rule base and the fuzzy result obtained is clarified and converted into a classical number. [11] In short, fuzzification is turning the numbers given as input data into fuzzy variables. On the other hand, defuzzification is the conversion of fuzzy variables back to numerical values.

The areas where fuzzy logic is used are given below.

- Operation control of subways
- Cameras focusing on the image
- Adjusting air conditioners, washing machines, vacuum cleaners
- Control of automobile engines
- Missile control
- Recognition of characters and objects

2.1.3 Genetic Algorithms

Genetic algorithms for the two purposes of abstracting and explaining the adaptive processes of natural systems and designing artificial systems software that preserve the important mechanisms of natural systems were developed by John

Holland, his colleagues, and his students. [12] Genetic algorithms are used when selecting parameters to optimize the performance of a system. Genetic algorithms give successful results especially in problem types where the solution space is large, discontinuous and complex. Genetic algorithms are used in many areas such as optimization problems, mechanical learning, economic and social system modeling, financial modeling applications, marketing problems, problems in production/operations, assembly line balancing problems, scheduling problem, facility layout problem, assignment problem, cellular production problem, system reliability problem, transportation problem and vehicle routing problem. [13]

The working principle of genetic algorithms is summarized in the flow chart in Figure 3.[14]

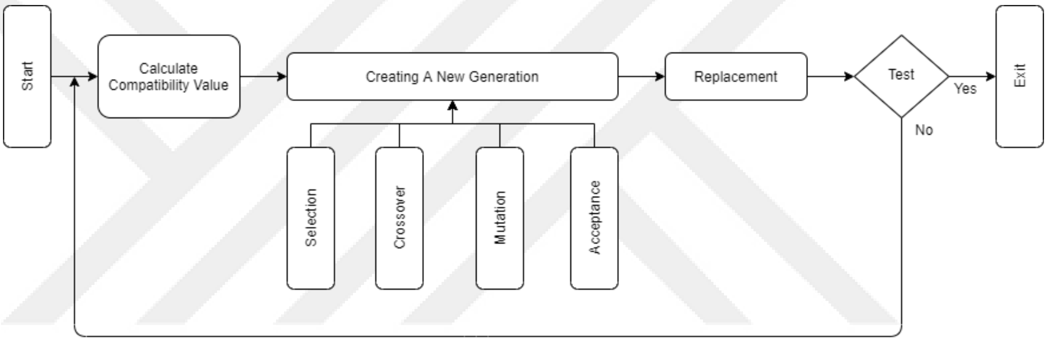


Figure 3 General flow chart of the genetic algorithm

We can explain how genetic algorithms work in four steps. Firstly, a solution group is formed in which the possible solutions of a problem are coded, and this group is called the population, and the codes of the solutions are called chromosomes. These names are given because of the similarity in biology. That is, a population with n chromosomes randomly is generated for a problem. The second step is that the compatibility value of each chromosome is calculated. This calculator indicates how good the chromosomes are. In the next step, selection, crossover and mutation processes are done until a new population is formed. In the fourth and last step, when the new population is accepted, it is replaced with the old population. If the targeted fitness value is reached, the program is stopped. The result is the best solution for the population. [14]

2.2. MACHINE LEARNING

Machine learning is an artificial intelligence methodology and hence a subset of artificial intelligence. It is necessary to use machine learning to achieve artificial intelligence. Artificial intelligence refers to machines that have the ability to think, learn, and move in a way similar to humans. Artificial intelligence systems can be created, programmed and developed by humans. All of this is related to the concept of machine learning.

Machine learning was first used by computer scientists at the Dartmouth Conference in 1956 and has been the subject of many studies until today. [15] Machine learning is done by designing structure algorithms that can be trained on data. The data set to be used in the selection of these algorithms is of great importance.

Machine learning is an area at the intersection of statistics, artificial intelligence, computing, predictive analysis and statistical learning about obtaining information from available data. The outputs of the inputs that were not learned before can be estimated from the input and output sets given by machine learning. Python has been the common language for many data science applications. Python includes libraries in data processing, statistics, image processing, and many more. [16]

Machine learning can be examined around three main research areas. These main research areas are task-oriented studies, cognitive simulation and theoretical analysis.

- **Task-Oriented Studies:** The performance of previously defined tasks may degrade over time. The performance of predefined tasks can degrade over time. The development of learning systems to increase this performance is called task-oriented studies.
- **Cognitive Simulation:** Cognitive Simulation, also known as cognitive modeling approach, is the study and computer simulation of human learning processes.
- **Theoretical Analysis:** Theoretical Analysis is the theoretical examination of existing learning methods and algorithms, regardless of practical areas. [17]

The machine's ability to learn by using the data set given as an experience is provided with the help of various algorithms developed for this purpose and suitable for the problem type. Whether the machine learns or not is determined by the latest performance. If the performance is high, it will be possible for the machine to produce the correct output whenever a new output is requested. On the other hand, if the performance is low, changes are made in the factors affecting the learning of the machine and the performance is optimized. Four factors can be mentioned that affect machine learning. The first of these factors is the data set given to the machine as an experience. If the data set given as experience includes more and different possible situations, the more it will contribute positively to learning. The second factor is the use of more than one data that has no effect on the result or has the same effect in the data set. Attributes that are found to be highly related to each other or have little contribution to the solution should be eliminated. The third factor is the chosen learning strategy. The learning strategy is related to both the task the machine is desired to learn and the required data set. Machine learning needs an output value (target attribute) in the data set to solve some problems, but it does not need this value for some problems. Prediction and classification problems are the problems that need target attribute. Therefore, it is important to determine the learning strategy according to the problem to be addressed in machine learning. The fourth and last factor is the algorithm used for learning. The path that the machine must follow step by step in order to learn from its data set is provided by algorithms.

For machine learning, model / models are established on the basis of data set and algorithm, and the model with the highest performance is preferred. For this reason, many machine learning methods have been developed, some of which are; k-nearest neighbor algorithm, simple (naive) Bayes classifier, decision trees, logistic regression analysis, k-means algorithm, support vector machines and artificial neural networks. Some of these algorithms are capable of prediction and estimation, some of them are clustering and some of them are capable of classification. These algorithms are mentioned in section 2.3.

Machine learning algorithms are examined under two main headings in terms of their learning methods; supervised learning and unsupervised learning. However, with the advancement and development of technology and algorithms, additional methods have been developed besides these two main methods. These methods can be listed as semi-supervised, reinforcement, transductive inference, online, active learning. [18]

2.2.1. Machine Learning Methods

In this section, information is given about the concepts of supervised, unsupervised and reinforcement learning, which are among the learning methods. Figure 4 can be examined to get summary information on the subject.

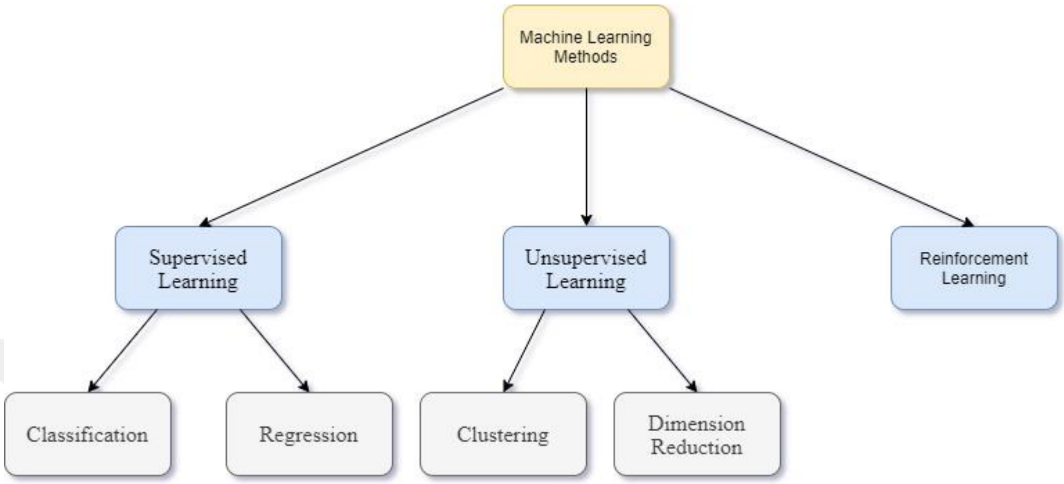


Figure 4 Machine Learning Methods

2.2.1.1. Supervised Learning

The main purpose of supervised learning is to provide a meaningful result from this information by giving the input data and output data given as a training set to the machine under the supervision of an inspector. In order to achieve this goal, there must be a training set with inputs and outputs. Each data in this training set is taught to the machine by a supervisor. After the machine is taught the relationship between data, the machine can make assumptions for a sample that has never been introduced using the relationship it has learned.

The supervised learning method tries to minimize the error difference between the taught and the real by using the training set. The error difference can be defined as the difference between the algorithm-generated output and the actual value of each input. At the stage of teaching the training set to the system, the error difference is determined for each data and compared with the error threshold value previously determined. If the difference is higher than the predetermined error value, the system continues its training. The training is completed when the difference reaches the desired range. After the model reaches the desired level, a new observation that was not in the training set before is processed by the model and the closest estimation to the truth is made.

As an example, a person who wants to classify comments on social media should show every comment posted on the system as input. Then, this person shows the system the intended output (in which subject it can be classified) of each input (comment). After learning the connection between the inputs and outputs in this large training set, the system can classify each new comment according to its subject by using the training set from which it has learned. Besides, classification of mails as spam can be given as an example of supervised learning practice. [18]

Supervised learning algorithms are used for two main purposes: regression and classification. Support vector machines, decision trees, artificial neural networks (ANN), k-nearest neighbor algorithm, logistic regression analysis and naïve bayes classifiers are examples of commonly used supervised learning algorithms.

2.2.1.2. Unsupervised Learning

The purpose of unsupervised learning is to group samples belonging to close characteristics and variables among the observations by gathering them under the same cluster. Only input variables are given to the system as a training set. There is no output or labeling information in the training set. In addition, there is no supervisor that gives data set to the system to learn. The system tries to perform operations such as classification and clustering by examining the relationship between input variables. [18]

For instance, in order to obtain the highest efficiency from students in a school, it is desired to divide students into classes, taking into account their demographic characteristics and abilities. Unsupervised learning algorithm can group students who are compatible with each other by determining various patterns among students from the information of students at school. In this way, classes can be organized. In this problem, it is assumed that students with the same characteristics work more efficiently.

Unsupervised learning is a learning method used in clustering problems and dimension reduction. Correlation analysis, factor analysis, cluster analysis, self-regulating maps, k-mean unsupervised learning can be given as examples.

2.2.1.3. Reinforcement Learning

Thanks to the reinforcement learning method, the system tries to maximize the good evaluation results from the environment by adapting itself to the actions it encounters. [19] The desired output for each input is not available in the training set. The auditor tries to teach the model to the system by evaluating the outputs produced by the system as good or bad according to the inputs each time. The system improves itself according to the evaluation results of the outputs it receives. Reinforcement learning is a type of learning applied in places such as robot navigation, skill acquisition, real-time decision making, artificial intelligence, computer games. The best example of reinforcement learning method is chess game.

2.3. ALGORITHMS

Machine Learning is the modeling of systems with computers that make inferences and make predictions using certain algorithms on a data set by performing mathematical operations. Machine learning algorithms are a kind of engine of machine learning, that is, they are algorithms that transform a data set into a model. Linear Regression, Logistic Regression, Decision Tree, SVM, Naive Bayes, kNN, K-Means and Random Forest are commonly used algorithms for machine learning. In our thesis study, three algorithms were used. These are Decision Tree, Random Forest and Logistic Regression. Detailed explanations about the algorithms are given in the headings 2.3.1, 2.3.2 and 2.3.3.

2.3.1. Decision Tree Algorithms

Decision trees are frequently used among classification techniques. It is a supervised learning algorithm used in both classification and regression models. The case where the dependent variable is categorical is a classification problem. A tree is drawn in a similar way to a flowchart and this tree is called a classification tree. The main purpose of decision trees is to divide the data set into small subgroups by applying certain rules using the divide-conquer method. [1] The name of the decision tree comes from the schema that the output of the algorithm looks like a tree. There are four basic structures on a decision tree. These are root node, decision nodes, branches and leaf nodes. Decision nodes are attributes used to make decisions, classify or predict in the data set, and they can be divided into two or more branches. Leaf nodes hold decisions. The node at the top of the tree is called the root node. The classification process starts from the root node. In order to reach a decision, a certain path is followed from the root of the tree to the leaf nodes. The last unbranched node of the tree is the leaf node. Figure 5 illustrates the decision node, leaf node and main rood more clearly.

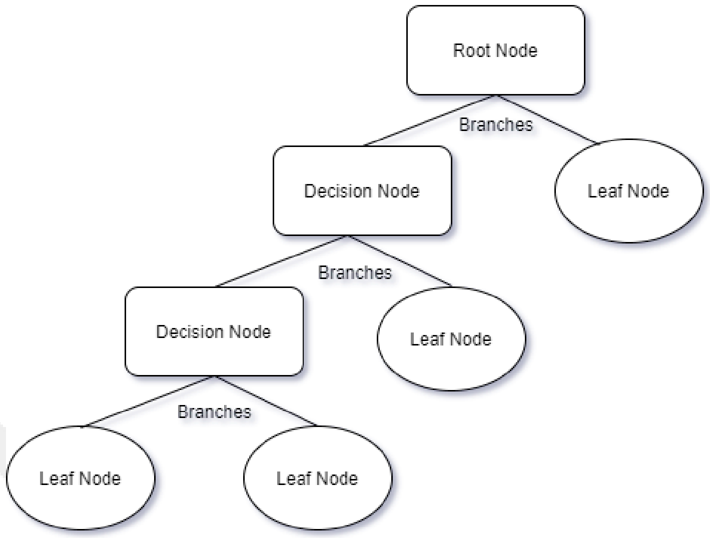


Figure 5 Example of Decision Tree

While classification is made with decision trees, a decision tree is created with the data set. This means that the rule base required for prediction is subtracted. With this rule, code improvements become easier for developers. The general algorithm steps used while creating the decision tree are listed below [20].

- The tree is created using the divide-and-conquer method iteratively from top to bottom.
- The tree starts with a single node of all training data.
- Attributes are continuously separated and included in the analysis. Predictive attributes and target attributes are categorical.
- If the attributes all belong to the same class, the node terminates as a leaf and gets the class label.
- If not all attributes belong to the same class, the attribute that best divides the samples into classes is chosen.
- Attributes are selected on the basis of intuitive or statistical criteria.
- If all instances in a node belong to the same class, there are no attributes left to divide the instances, or there is no instance with the value of the remaining attributes, the decision tree generation process ends. [21] When working with the same data set, more than one different decision tree can be created.

2.3.2. Random Forest Algorithms

Collective classification methods are learning algorithms that produce multiple classifiers instead of one classifier and then classify new data with the votes taken from their predictions. One of the most common classifications is the random forest algorithm. The random forest algorithm does not branch out each node, using the best branch among all variables. It branches each node by using the best among the randomly selected variables at each node. Each dataset is generated by displacement from the original dataset. Then trees are developed using random feature selection. [22] Random Forest is also very fast, resistant to overfitting, and can be worked with as many trees as desired.[23] Figure 6 can be examined to see the definition of random forest visually.

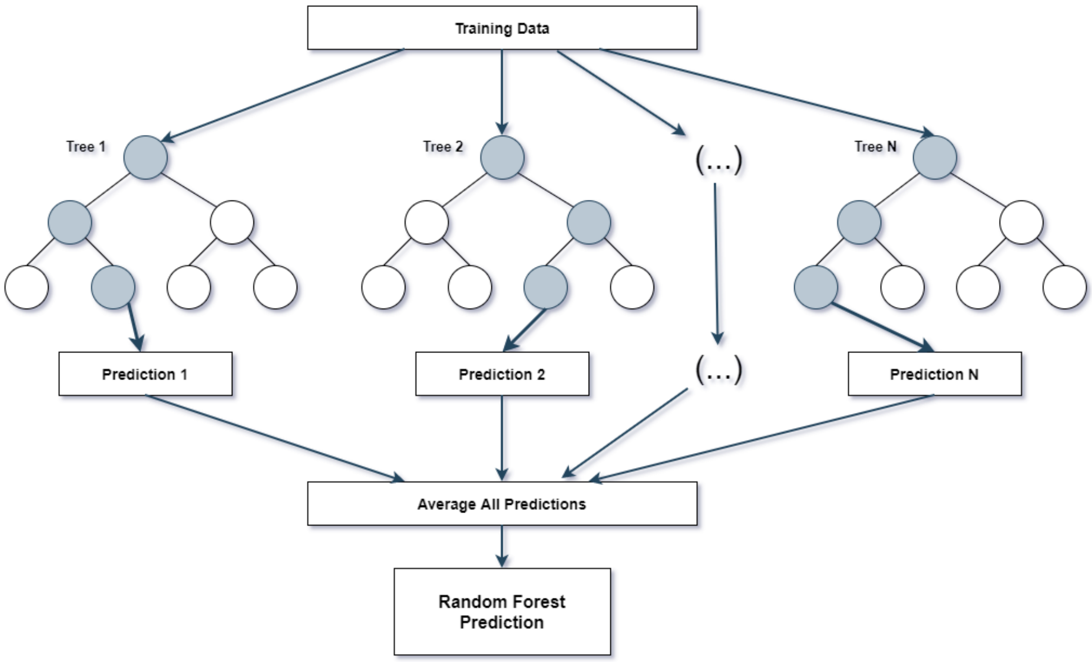


Figure 6 Random Forest Example

Random Forest algorithm is used in both categorical and continuous data sets and both; at the same time, it can be easily used in large or small sized data sets.

In this method, as in the Classification and Regression trees, the Gini index is used as the division criterion. A decrease in the Gini index is desirable because it indicates an increase in purity, and if this index is ultimately equal to zero, it means maximum purity.[24] Gini index for a given node t;

$$\boldsymbol{GINI(t)} = \sum_{j}[\boldsymbol{p(j \backslash t)}]^{2} \qquad \textbf{(2.1)}$$

In Equation 13, $\boldsymbol{p(j \backslash t)}$ represents the relative probability of the class at node t. As the Gini index increases, class heterogeneity increases, while as the Gini Index decreases, class homogeneity increases. A branch is successful when the Gini index of a child node is less than the Gini index of a parent node. When the Gini index reaches zero, that is, one class remains at each leaf node, the tree branching process ends. [24]

Random forest method is explained step by step.

- To start the algorithm, 2 parameters must be defined by the user.
- These parameters are the number of variables “m” used at each node to determine the best split and the number of trees to be developed “N”.
- First, bootstrapping samples are created from 2/3 of the training dataset.
- The remaining 1/3 of the training dataset, also called Out Of-bag (OBB) data, is used to test for errors.
- The tree is developed without pruning from each bootstrapping sample.
- At each node, m variables are randomly selected among all variables and the best branch is determined among these variables.
- It is extremely important to select the number of variables that provide sufficient predictive power and sufficient low correlation. [25]
- The number of m variables taken equal to the square root of the total number of M variables generally gives the closest result to the optimum result.[26]

2.3.3. Logistic Regression

Regression analysis is used to measure the relationships between two or more variables and provides descriptive and inferential analysis. Quantitative variables are generally used in these analyzes. However, in some cases, the variables used for analysis may be qualitative variables. The main purpose of logistic regression is to establish an acceptable model that can describe the relationship between dependent and independent variables in a way that has the best fit with the least variable.[27]

In logistic regression, none of the valid assumptions are sought in linear regression. Therefore, it is more preferred by researchers. The most important difference between logistic regression and linear regression is that the dependent variable in logistic regression is categorical.

Logistic regression analysis can be used for three different situations. These are binary logistic regression, ordinal logistic regression and multinomial logistic regression. Binary logistic regression is when the dependent variable has two categories. Ordered logistic regression is when the dependent variable has more than two categories and is sortable. Multi-category logistic regression is the situation where the dependent variable has more than two categories and is unordered. [28]

Logistic regression is a distinctive model based on the logistic function. This function is also called sigmoid function. The logistic regression (sigmoid) function is shown in Equation 2.1.

$$\boldsymbol{s(z) = \frac{1}{1+e^{-z}}} = \boldsymbol{\frac{e^{z}}{e^{z}+1}}\ , \mathbb{R} \rightarrow \boldsymbol{[0,1]} \qquad \textbf{(2.2)}$$

The sigmoid function helps the logistic regression model compress values between (-k,k) and (0.1). The graph of the logistic function showing this definition is shown in Figure 7.[29]

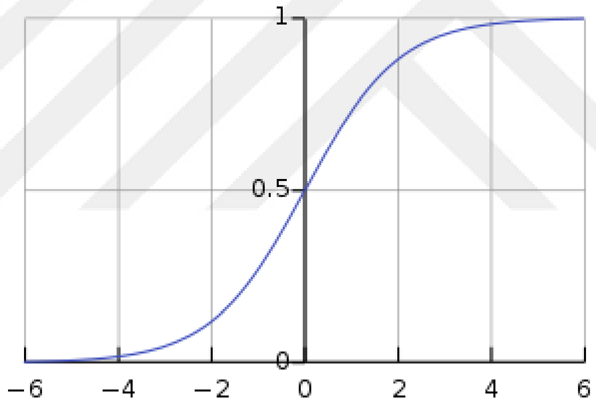


Figure 7 The Logistic Curve

The figure shows that if the logistic function value, which takes a value between 0 and 1, is 0.5 or more, y is considered to be 1. Otherwise, y is considered 0 if the logistic function value is below 0.5. It is also clearly explained with the Equation 2.2.

$$\text{If } s(z) \geq 0.5\ ,\ \text{predict y=1}$$
$$\text{If } s(z) < 0.5\ ,\ \text{predict y=0}$$

(2.3)

CHAPTER 3
LITERATURE REVIEW

As the importance of machine learning, data mining, artificial intelligence with student performance estimation increases day by day, it is observed that the studies on these subjects are increasing. In this section, the National Thesis Center, libraries, journals, articles, local and foreign sources, and some studies on the subject are mentioned.

Some of the machine learning algorithms used in these studies are Artificial Neural Network (ANN), Naïve Bayes (NB), Decision Tree (DT), C 4.5 Decision Tree, Logistic Regression, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest, Multilayer Perceptron.

In a study conducted in Malaysia, to predict and classify students' academic performance, artificial neural network (ANN) technique and combination of decision tree classification and clustering are compared and used. The student data used in this study were taken from the National Defence University of Malaysia. 85 students' data are used in this study. These data contain demographics information such as gender, race, secondary school, educational information such as computer skill, name of courses taken and personality information such as interest, reading level and motivation. [30]

In 2013, a study was conducted to predict students' graduation grades. Artificial neural network (ANN) and decision tree are used to make this prediction. Year-end grades of 49 cultural and vocational courses in total that 127 students took over 4 years were used. Two different scenarios have been tried to predict the graduation grade. In the first scenario, first two university years grades are used for prediction. In the second and last scenarios, first three university years grades are used for prediction. As a result, it has been observed that ANN provides better prediction performance compared to DT and the second scenario makes better predictions compared to the first scenario. [31]

In another study, factors affecting the academic performance of first year students are investigated and their success based on these factors is predicted. Traditional information such as age, gender, high school GPA and non-traditional information such as general self-efficacy, academic self-efficacy were obtained for the performance prediction study. CRISP-DM is adopted. Multilayer Perceptron and Random Forests were used as machine learning techniques. As a result of the research, the Random Forests method showed the best performance. The factor that had the greatest effect on students' year-end performance was found to be "First semester GPA". [32]

In a study conducted in Istanbul, the course success of high school students was predicted by using some classification techniques. Data includes anxiety, exhaustion, academic motivation and socio-demographic variables (age, gender, educational status of mother and father, etc.). In addition to these factors, data also include the student's year-end success average and absenteeism information. Using K-Nearest Neighborhood, Naive Bayes Classifier, C 4.5 Decision Tree, Logistic Regression and Support Vector Machine techniques, some different models were created. The C4.5 Decision Tree Algorithm produced more successful results than others in predicting academic success. [33]

In 2006, a study was conducted to predict the test scores of the students based on individual skills using Bayesian networks. Moreover, the best result was a predictive error of 15%. Information was collected on the USA's 8th-grade mathematics tests from the online tutoring system. [34]

In 2008, a study was conducted to predict student performance in two secondary schools in Portugal. Student grades, demographic, social and school-related features, etc. information was collected by using the questionnaire and school reports. Decision Trees, Random Forest, Neural Networks and Support Vector Machines were tested. The best solution was obtained by a Naive Bayes method. [35]

In another study, using decision tree techniques, students' results were predicted. While applying this technique, five attributes of 120 students were analyzed such as lab work, class performance, attendance, assignment, sessional performance. [36]

In 2014, aptitude, personality, motivation learning strategies keys were reviewed to predict learner progression in tertiary education in Ireland. 914 students were analyzed with these attributes. Naïve Bayes (NB), Decision Tree (DT), Logistic Regression, Support Vector Machines (SVM), ANN (Artificial Neural Network), k-Nearest Neighbor (KNN) techniques are applied. [37]

To predict the student's performance at the final examination, a study was conducted in 2014 in Greece. In this study, ANN, DT, NB, Rule-Learning, SVM techniques are applied and students' grades are used for prediction. The best solution belongs to ANN and SVM for this study. [38]

In the UK, a study was conducted about prediction 'at-risk' students. For this study, Personal and demographic information, student satisfaction and integration information of 149 students are used. ANN and Logistic Regression were used as machine learning techniques. As a result of the research, the Logistic Regression method showed the best performance. [39]

Table 1 Comparison of Literature Search

Paper	Dataset	Dataset Size	Machine Learning Algorithms	Best Algorithm
[30] MuslihahWook et al, 2009	Demographics information such as gender, race, secondary school, educational information such as computer skill, name of courses taken and personality information such as interest, reading level and motivation	85 students	• Artificial neural network (ANN) • Decision tree	Artificial neural network (ANN)

Table 1 Continuation

[31] Şengür, 2013	The first three university years grades	127 students	• Artificial neural network (ANN) • Decision tree	Artificial neural network (ANN)
[32] Hakyemez, 2015	Traditional information such as age, gender, high school GPA and non-traditional information such as general self-efficacy, academic self-efficacy	Business faculty undergraduate first year students	• Multilayer Perceptron • Random Forests	Random Forests
[33] Özdemir, 2016	Anxiety, exhaustion, academic motivation, socio-demographic variables (age, gender, educational status of mother and father, etc.), the student's year-end success average and absenteeism information	1706 students	• K-Nearest Neighborhood Naive Bayes Classifier • C 4.5 Decision Tree • Logistic Regression Support Vector Machine	C4.5 Decision Tree
[34] Pardos et al, 2009-6	8th grade mathematics tests	600 students	• Bayesian networks	Bayesian networks (error rate %15)

Table 1 Continuation

[35] Cortez et al, 2008	Student grades, demographic, social and school related features	Secondary education students	• Decision Trees • Random Forest • Neural Networks • Support Vector Machine	Naive Bayes (error rate %15)
[36] Guleria et al, 2014	Lab work, class performance, attendance, assignment, sessional performance.	120 students	• Decision tree	Decision tree
[37] Gray et al, 2014	Aptitude, personality, motivation learning strategies	914 students	• Naïve Bayes (NB) • Decision Tree (DT) • Logistic Regression • Support Vector Machines (SVM) • ANN (Artificial Neural Network) • k-Nearest Neighbor (KNN)	Support Vector Machines
[38] Livieris et al, 2012	Student's grades	Students in the course of Mathematics	• ANN • DT • NB • Rule-Learning • SVM	ANN SVM
[39] Sarker et al, 2014	Personal and demographics information, student satisfaction and integration information	149 students	• ANN • Logistic Regression	Logistic Regression (error rate %15.33)

CHAPTER 4
PREDICTION PROCESSES

In this section, the dataset used to make predictions is explained. The IDE, programming language and libraries used are presented. Algorithms used to determine the academic performance of students are mentioned. The graphs and tables that emerged as a result of the algorithms are shown.

4.1. DATASET

We have used dataset from Kaggle website[1] in our study. Data were obtained from students' knowledge from two schools in Portugal. These data were collected through surveys and school reports. The data includes a total of 33 attributes for each student, including demographic information, school grades, school information, and social information [35]. Of the 33 attributes, 16 attributes hold integers, 9 attributes strings, and 8 attributes boolean data. The names and descriptions of the attributes are given in Table 2.

[1] https://www.kaggle.com/datasets

Table 2 Attribute List of Dataset

Attribute Name	Attribute Description
school	student's school (GP or MS)
sex	student's sex (F or M)
age	student's age (15-22)
address	student's home address type (U or R)
famsize	family size (LE3 or GT3)
Pstatus	parent's cohabitation status (T or A)
Medu	mother's education (0,1,2,3 or 4)
Fedu	father's education (0,1,2,3 or 4)
Mjob	mother's job (teacher,health,services,at_home)
Fjob	father's job (teacher,health,services,at_home)
reason	reason to choose this school (home,reputation,course,other)
guardian	student's guardian (mother,father or other)
traveltime	home to school travel time (1,2,3 or 4)
studytime	weekly study time (1,2,3 or 4)
failures	number of past class failures
schoolsup	extra educational support (yes or no)
famsup	family educational support (yes or no)
paid	extra paid classes within the course subject (yes or no)
activities	extra paid classes within the course subject (yes or no)
nursey	attended nursery school (yes or no)
higher	wants to take higher education (yes or no)
internet	internet access at home (yes or no)
romantic	with a romantic relationship (yes or no)
famrel	quality of family relationships (1-5)
freetime	free time after school (1-5)
gout	going out with friends (1-5)
Dalc	workday alcohol consumption (1-5)
Walc	weekend alcohol consumption (1-5)
healty	current health status (1-5)
absences	number of school absences (0-93)
G1	first period grade of Math or Portuguese (0-20)
G2	second period grade of Math or Portuguese (0-20)
G3	final grade of Math or Portuguese (0-20)

4.2. CONFUSION MATRIX

The confusion matrix allows to compare the predicted values with test data and measure the performance of the model made.

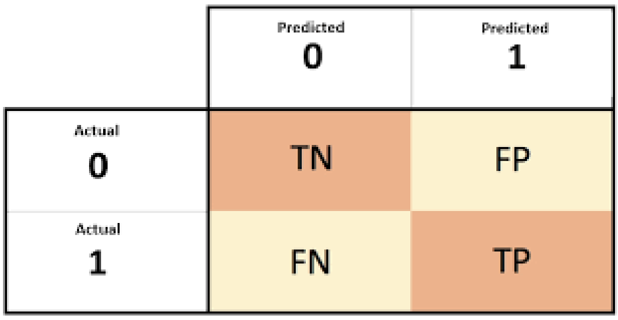


Figure 8 Confusion Matrix

- True Negative (TN): This value gives the number of correct values. For instance, this value is the number of students that the model predicts as having low academic performance from students who actually have low academic performance.
- False Positive (FP): This value gives the number of incorrect values. For instance, this value is the number of students that the model predicts as having low academic performance from students who actually have high academic performance.
- False Negative (FN): This value gives the number of incorrect values. For instance, this value is the number of students that the model predicts as having high academic performance from students who actually have low academic performance.
- True Positive (TP): This value gives the number of correct values. For instance, this value is the number of students that the model predicts as having high academic performance from students who actually have high academic performance.

Accuracy, Recall, Precision, F1-Score, ROC/AUC values, which measure model success by using TP, TN, FP and FN values obtained from the confusion matrix, are mentioned in sections 4.3 and 4.4.

4.3. CLASSIFICATION REPORT

With the Classification report, the number of classes used, the precision, recall, f1-score, accuracy, macro avg, weighted avg and support values of the classes are displayed. Figure 9 is an example image for the report. Values will be explained on this figure.

	precision	recall	f1-score	support
Class 1	0.94	0.65	0.77	26
Class 2	0.95	0.99	0.97	169
accuracy			0.95	195
macro avg	0.95	0.82	0.87	195
weighted avg	0.95	0.95	0.94	195

Figure 9 Classification Report Example

- **Accuracy**

Accuracy value is calculated with the following formula. Accuracy value is a value between 0 and 1. The model is considered successful when the Accuracy value approaches 1.

$$Accuracy = \frac{TP + TN}{\mathrm{TP + FP + TN + FN}} \tag{4.1}$$

- **Precision**

Precision is a metric that shows how many of the values we predicted as positive are actually positive.

$$Precision = \frac{TP}{\mathrm{TP + FP}} \tag{4.2}$$

- **Recall**

Recall is a metric that shows how many of the transactions we need to positively predict, we predict positively.

$$Recall = \frac{TP}{\mathrm{TP + FN}} \tag{4.3}$$

- **F1-Score**

In order not to ignore the extreme cases, it is necessary to use a harmonic mean instead of a simple mean. F1 Score value shows the harmonic average of Precision and Recall values.

$$F1\ Score = 2 * \frac{Precision * \ Recall}{\text{Precision} \ + \ \text{Recall}}$$

(4.4)

The main reason for using the F1 Score value instead of Accuracy is not to make an incorrect model selection in unevenly distributed datasets.

- **Macro Average**

The macro average is the usual average we all know. It is calculated by adding all the values and dividing by the number of values.

$$Macro\ Average\ Precision = \frac{\text{class1 precision} + .. + \text{classN precision}}{\text{number of classes}}$$

(4.5)

$$Macro\ Average\ Recall = \frac{\text{class1 r}ecall + .. + \text{classN r}ecall}{\text{number of classes}}$$

(4.6)

$$Macro\ Average\ F1Score = \frac{\text{class1 f1score} + .. + \text{classN f1score}}{\text{number of classes}}$$

(4.7)

- **Weighted Average**

After multiplying the relevant metrics and their weights, these values are added together and the value formed by dividing by the sum of the weights is called Weighted Average.

$$Weighted\ Average\ Precision = \frac{(\text{class1 precision} \ * \ \text{class1 support}) + .. + (\text{classN precision} \ * \ \text{classN support})}{\text{number of supports}}$$

(4.8)

$$\begin{aligned}&Weighted\ Average\ Recall\\&= \frac{(\text{class1 recall } * \text{ class1 support}) + .. + (\text{classN recall } * \text{ classN support})}{\text{number of supports}}\end{aligned} \quad \textbf{(4.9)}$$

$$\begin{aligned}&Weighted\ Average\ F1score\\&= \frac{(\text{class1 f1score } * \text{ class1 support}) + .. + (\text{classN f1score } * \text{ classN support})}{\text{number of supports}}\end{aligned} \quad \textbf{(4.10)}$$

- **Support**

Support value shows the number of observations for each class.

4.4. ROC CURVE

The ROC curve is a graph drawn to summarize the performance of all possible values. A graph is drawn by showing true positive rates on the Y axis and false negative rates on the X axis. The ROC curve example is shown in the Figure 10.

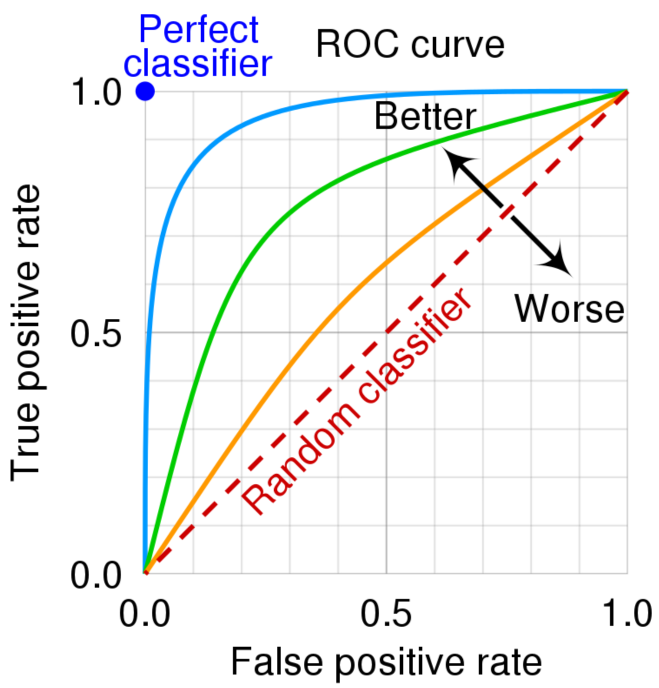


Figure 10 ROC Curve Example

The purpose of the ROC curve is to increase the true positive area and decrease the false positive rate. From the graph, it is understood that a positive classification can never be given at the point (0,0), absolute positive classification will be obtained at the point (1,1), and perfect classification at the point (0,1) [40]. If the ROC curve is close to the upper left corner, it means that a good classifier with high performance is used. AUC is the area under the ROC curve. The higher the AUC, the better the model predicts.

CHAPTER 5
STUDENT ACADEMIC PERFORMANCE PREDICTION

In our study, we used three different algorithms to make performance prediction. Detailed explanations about these algorithms can be found in section 2.3. Using the dataset mentioned in section 4.2, the G3 data (the final grade) was predicted.

Before the three algorithms work, we need to reach clean data. There may be missing values in the data and instead of these missing parts, 0, NaN, space, NULL, undefined can be found. These missing data should be deleted or the missing parts should be filled with predicted values. On the other hand. there may be some deviations in the value ranges that the values can take depending on their type. Clean data is achieved by getting rid of such missing or incorrect data. Prediction made with clean data gives more accurate results.

Moreover, we prepared three datasets with different student numbers. In this way, we observed which algorithm made more accurate predictions as the number of data increased or decreased. The first of our datasets contains 245 student data, the second contains 395 student data, and the last contains 649 student data.

30% of the dataset was used as test data. The remaining 70% data was used as training data. This ratio is valid for all three algorithms.

By using three algorithms, the confusion matrix, accuracy, recall, and precision values are shown in Section 5.1, 5.2 and 5.3.

5.1. PREDICTION WITH DECISION TREE ALGORITHM

Using the decision tree algorithm mentioned in Section 2.3.1, the end-of-year academic performance of the students was predicted with the data of students. 70% of the dataset was used as training data. 30% dataset was used for testing. Some graphs, tables and confusion matrix created. First, the prediction results with 649 student data, then the prediction results with 395 student data, and the last prediction results with 245 student data are explained.

- **For 649 students**

Firstly, the confusion matrix is mentioned. With the decision tree algorithm, the accuracy rate for 649 students is 0.949. The confusion matrix with decision tree algorithm for 649 students is shown with Figure 11.

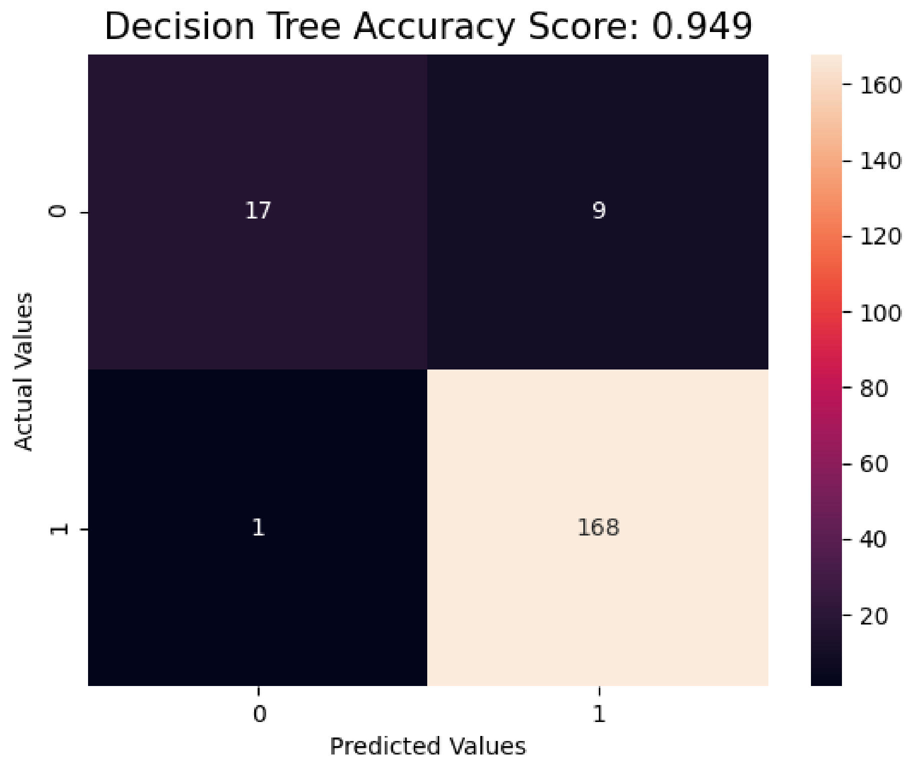


Figure 11 Confusion matrix with Decision Tree for 649 students

True Negative (TN), False Positive (FP), False Negative (FN), True Positive (TP) values with decision tree for 649 students are clearly shown in the Table 3.

Table 3 Details of Confusion matrix with Decision Tree for 649 students

True Negative (TN)	17
False Positive (FP)	9
False Negative (FN)	1
True Positive (TP)	168

The sum of these values is 30% of the total dataset. Because we reserved 30% of our data for testing.

	precision	recall	f1-score	support
Final Grade >= 10	0.94	0.65	0.77	26
Final Grade < 10	0.95	0.99	0.97	169
accuracy			0.95	195
macro avg	0.95	0.82	0.87	195
weighted avg	0.95	0.95	0.94	195

Figure 12 Classification Report of Decision Tree for 649 students

When decision tree classification report is generated for 649 students, as you can see in Figure 12, accuracy, precision, recall, f1-score, support, macro average, weighted average values are reached.

- **For 395 students**

Firstly, the confusion matrix is mentioned. With the decision tree algorithm, the accuracy rate for 395 students is 0.924. The confusion matrix with decision tree algorithm for 395 students is shown with Figure 13.

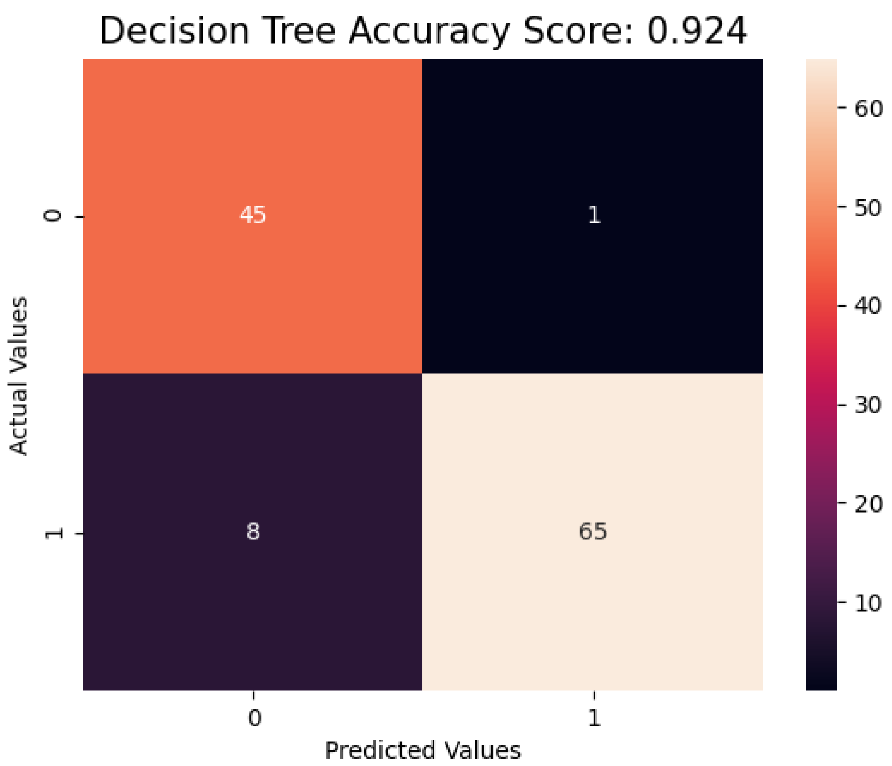


Figure 13 Confusion matrix with Decision Tree for 395 students

True Negative (TN), False Positive (FP), False Negative (FN), True Positive (TP) values with decision tree for 395 students are clearly shown in the Table 4.

Table 4 Details of Confusion matrix with Decision Tree for 395 students

True Negative (TN)	45
False Positive (FP)	1
False Negative (FN)	8
True Positive (TP)	65

The sum of these values is 30% of the total dataset. Because we reserved 30% of our data for testing.

	precision	recall	f1-score	support
Final Grade >= 10	0.85	0.98	0.91	46
Final Grade < 10	0.98	0.89	0.94	73
accuracy			0.92	119
macro avg	0.92	0.93	0.92	119
weighted avg	0.93	0.92	0.93	119

Figure 14 Classification Report of Decision Tree for 395 students

When decision tree classification report is generated for 395 students, as you can see in Figure 14, accuracy, precision, recall, f1-score, support, macro average, weighted average values are reached.

- **For 245 students**

Firstly, the confusion matrix is mentioned. With the decision tree algorithm, the accuracy rate for 245 students is 0.905. The confusion matrix with decision tree algorithm for 245 students is shown with Figure 15.

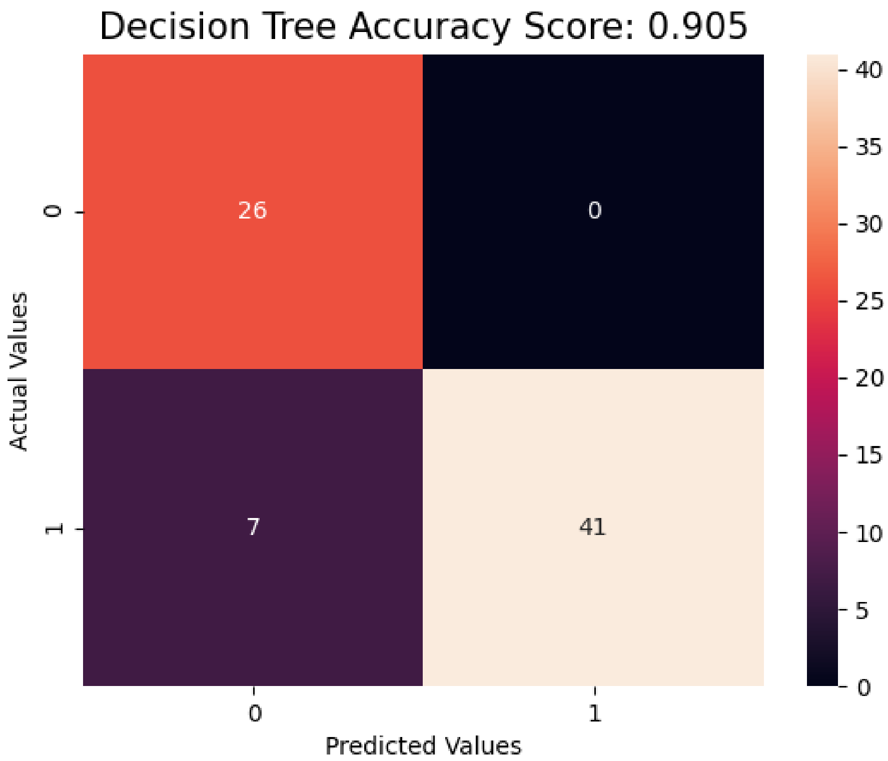


Figure 15 Confusion matrix with Decision Tree for 245 students

True Negative (TN), False Positive (FP), False Negative (FN), True Positive (TP) values with decision tree for 245 students are clearly shown in the Table 5.

Table 5 Details of Confusion matrix with Decision Tree for 245 students

True Negative (TN)	26
False Positive (FP)	0
False Negative (FN)	7
True Positive (TP)	41

The sum of these values is 30% of the total dataset. Because we reserved 30% of our data for testing.

	precision	recall	f1-score	support
Final Grade >= 10	0.79	1.00	0.88	26
Final Grade < 10	1.00	0.85	0.92	48
accuracy			0.91	74
macro avg	0.89	0.93	0.90	74
weighted avg	0.93	0.91	0.91	74

Figure 16 Classification Report of Decision Tree for 245 students

When decision tree classification report is generated for 245 students, as you can see in Figure 16, accuracy, precision, recall, f1-score, support, macro average, weighted average values are reached.

5.2. PREDICTION WITH RANDOM FOREST ALGORITHM

Using the random forest algorithm mentioned in Section 2.3.2, the end-of-year academic performance of the students was predicted with the data of students. 70% of the dataset was used as training data. 30% dataset was used for testing. Some graphs, tables and confusion matrix created. First, the prediction results with 649 student data, then the prediction results with 395 student data, and the last prediction results with 245 student data are explained.

- **For 649 students**

Firstly, the confusion matrix is mentioned. With the random forest algorithm, the accuracy rate for 649 students is 0.933. The confusion matrix with decision tree algorithm for 646 students is shown with Figure 17.

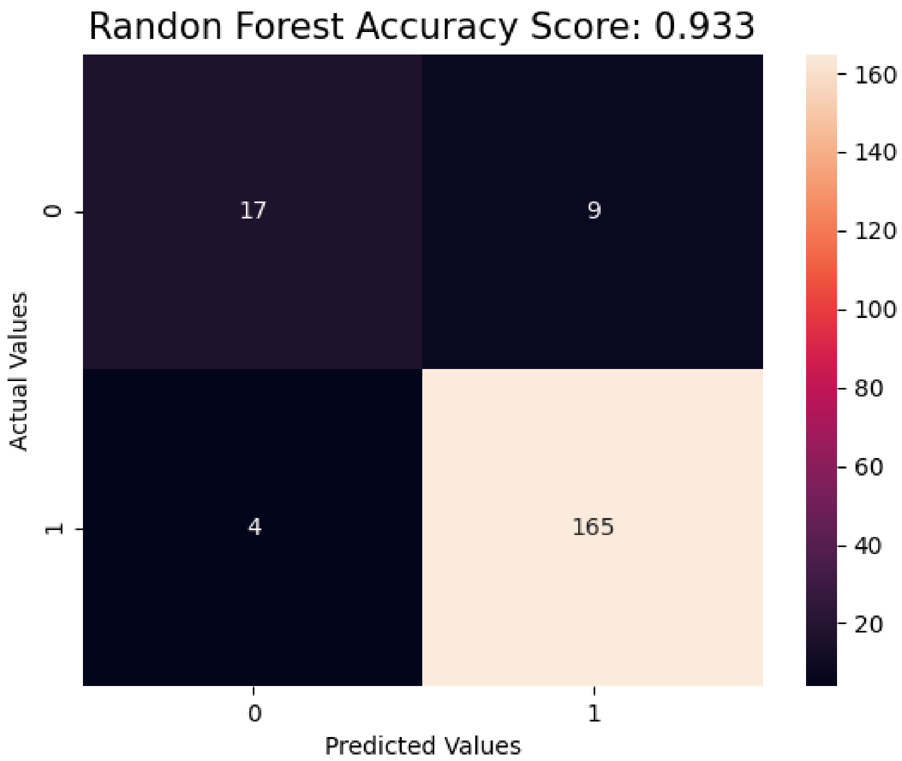


Figure 17 Confusion matrix with Random Forest for 649 students

True Negative (TN), False Positive (FP), False Negative (FN), True Positive (TP) values with random forest for 649 students are clearly shown in the Table 6.

Table 6 Details of Confusion matrix with Random Forest for 649 students

True Negative (TN)	17
False Positive (FP)	9
False Negative (FN)	4
True Positive (TP)	165

The sum of these values is 30% of the total dataset. Because we reserved 30% of our data for testing.

	precision	recall	f1-score	support
Final Grade >= 10	0.81	0.65	0.72	26
Final Grade < 10	0.95	0.98	0.96	169
accuracy			0.93	195
macro avg	0.88	0.82	0.84	195
weighted avg	0.93	0.93	0.93	195

Figure 18 Classification Report of Random Forest for 649 students

When random forest classification report is generated for 649 students, as you can see in Figure 18, accuracy, precision, recall, f1-score, support, macro average, weighted average values are reached.

- **For 395 students**

Firstly, the confusion matrix is mentioned. With the random forest algorithm, the accuracy rate for 395 students is 0.899. The confusion matrix with decision tree algorithm for 395 students is shown with Figure 19.

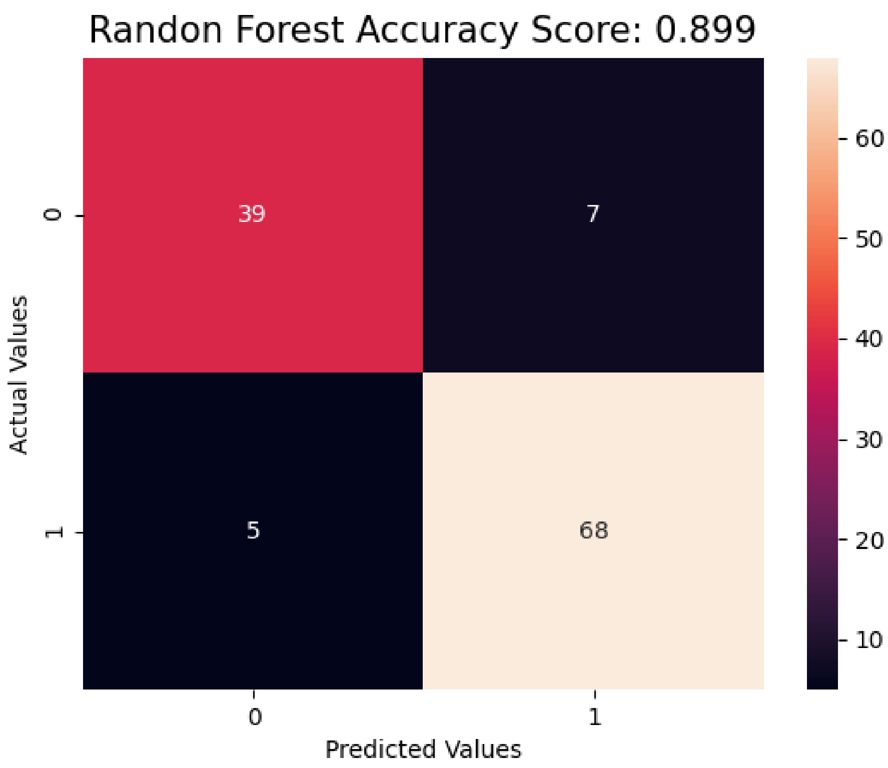


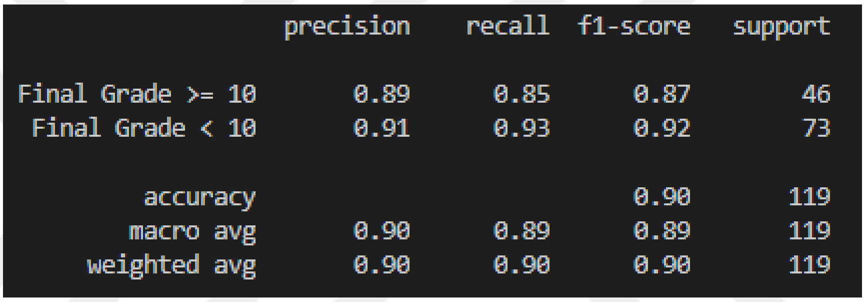
Figure 19 Confusion matrix with Random Forest for 395 students

True Negative (TN), False Positive (FP), False Negative (FN), True Positive (TP) values with random forest for 395 students are clearly shown in the Table 7.

Table 7 Details of Confusion matrix with Random Forest for 395 students

True Negative (TN)	39
False Positive (FP)	7
False Negative (FN)	5
True Positive (TP)	68

The sum of these values is 30% of the total dataset. Because we reserved 30% of our data for testing.



	precision	recall	f1-score	support
Final Grade >= 10	0.89	0.85	0.87	46
Final Grade < 10	0.91	0.93	0.92	73
accuracy			0.90	119
macro avg	0.90	0.89	0.89	119
weighted avg	0.90	0.90	0.90	119

Figure 20 Classification Report of Random Forest for 395 students

When random forest classification report is generated for 395 students, as you can see in Figure 20, accuracy, precision, recall, f1-score, support, macro average, weighted average values are reached.

- **For 245 students**

Firstly, the confusion matrix is mentioned. With the random forest algorithm, the accuracy rate for 245 students is 0.851. The confusion matrix with decision tree algorithm for 245 students is shown with Figure 21.

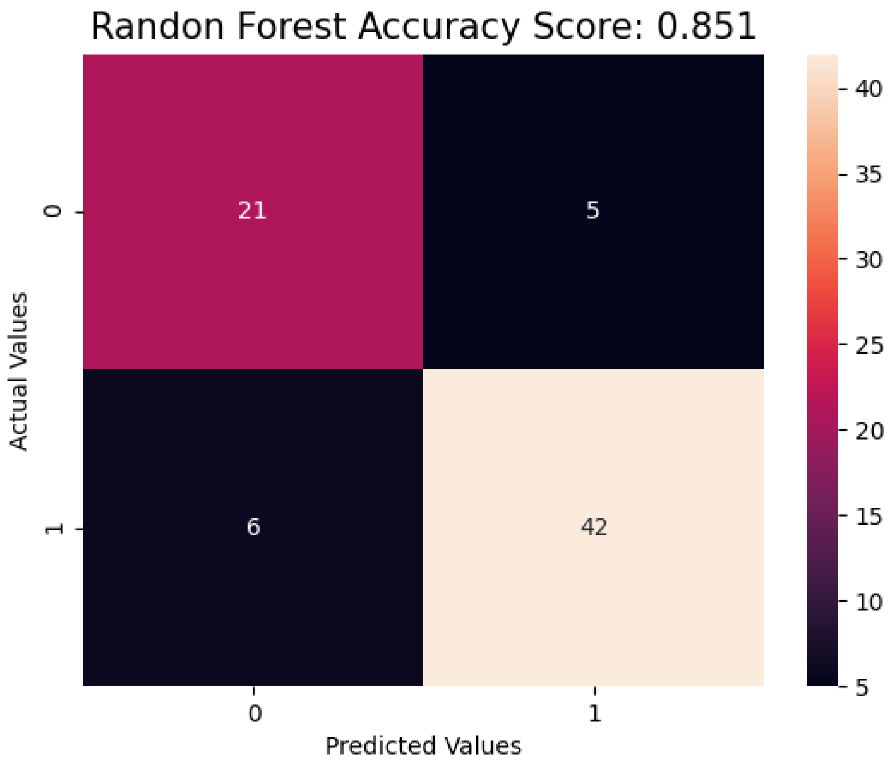


Figure 21 Confusion matrix with Random Forest for 245 students

True Negative (TN), False Positive (FP), False Negative (FN), True Positive (TP) values with random forest for 245 students are clearly shown in the Table 8.

Table 8 Details of Confusion matrix with Random Forest for 245 students

True Negative (TN)	21
False Positive (FP)	5
False Negative (FN)	6
True Positive (TP)	42

The sum of these values is 30% of the total dataset. Because we reserved 30% of our data for testing.

	precision	recall	f1-score	support
Final Grade >= 10	0.78	0.81	0.79	26
Final Grade < 10	0.89	0.88	0.88	48
accuracy			0.85	74
macro avg	0.84	0.84	0.84	74
weighted avg	0.85	0.85	0.85	74

Figure 22 Classification Report of Random Forest for 245 students

When random forest classification report is generated for 245 students, as you can see in Figure 22, accuracy, precision, recall, f1-score, support, macro average, weighted average values are reached.

5.3. PREDICTION WITH LOGISTIC REGRESSION ALGORITHM

Using the logistic regression algorithm mentioned in Section 2.3.3, the end-of-year academic performance of the students was predicted with the data of students. 70% of the dataset was used as training data. 30% dataset was used for testing. Some graphs, tables and confusion matrix created. First, the prediction results with 649 student data, then the prediction results with 395 student data, and the last prediction results with 245 student data are explained.

- **For 649 students**

Firstly, the confusion matrix is mentioned. With the logistic regression algorithm, the accuracy rate for 649 students is 0.908. The confusion matrix with logistic regression algorithm for 649 students is shown with Figure 23.

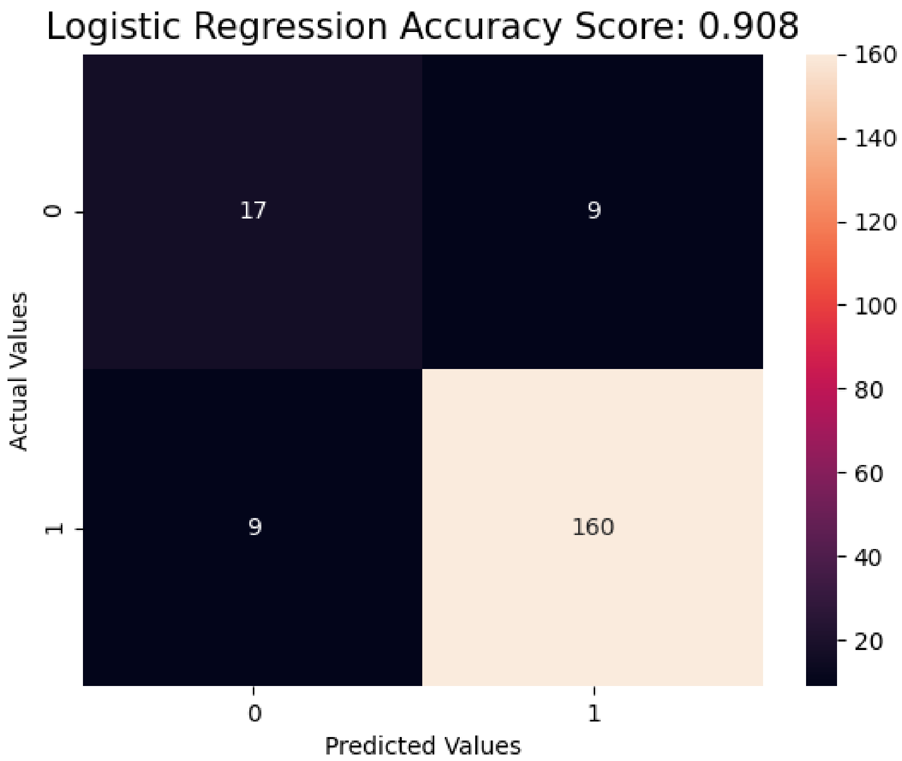


Figure 23 Confusion matrix with Logistic Regression for 649 students

True Negative (TN), False Positive (FP), False Negative (FN), True Positive (TP) values with logistic regression for 649 students are clearly shown in the Table 9.

Table 9 Details of Confusion matrix with Logistic Regression for 649 students

True Negative (TN)	17
False Positive (FP)	9
False Negative (FN)	9
True Positive (TP)	160

The sum of these values is 30% of the total dataset. Because we reserved 30% of our data for testing.

	precision	recall	f1-score	support
Grade >= 10	0.65	0.65	0.65	26
Grade < 10	0.95	0.95	0.95	169
accuracy			0.91	195
macro avg	0.80	0.80	0.80	195
weighted avg	0.91	0.91	0.91	195

Figure 24 Classification Report of Logistic Regression for 649 students

When logistic regression classification report is generated for 649 students, as you can see in Figure 24, accuracy, precision, recall, f1-score, support, macro average, weighted average values are reached.

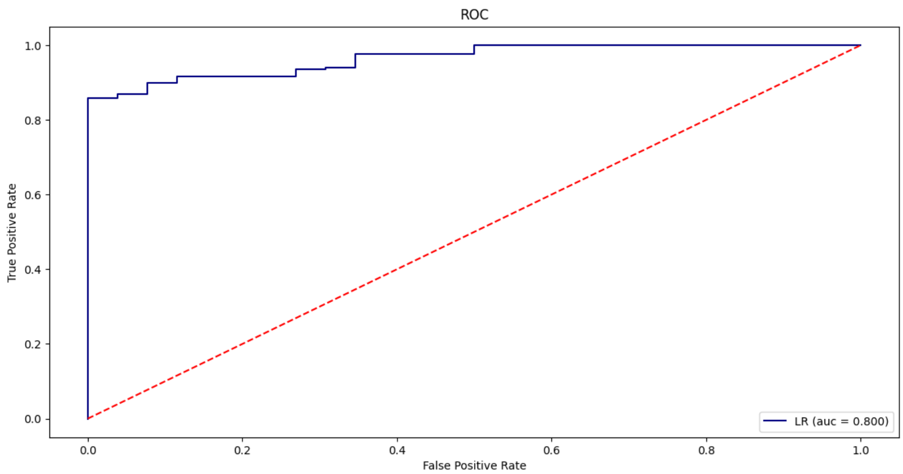


Figure 25 ROC Curve with Logistic Regression for 649 students

ROC Curve in Figure 25 shows that the AUC value is 0.800 for 649 students. We know that the higher the AUC value, the better the prediction. The ideal value for AUC is 1. 0.800 is a good result compared to 1.

- **For 395 students**

Firstly, the confusion matrix is mentioned. With the logistic regression algorithm, the accuracy rate for 395 students is 0.916. The confusion matrix with logistic regression algorithm for 395 students is shown with Figure 26.

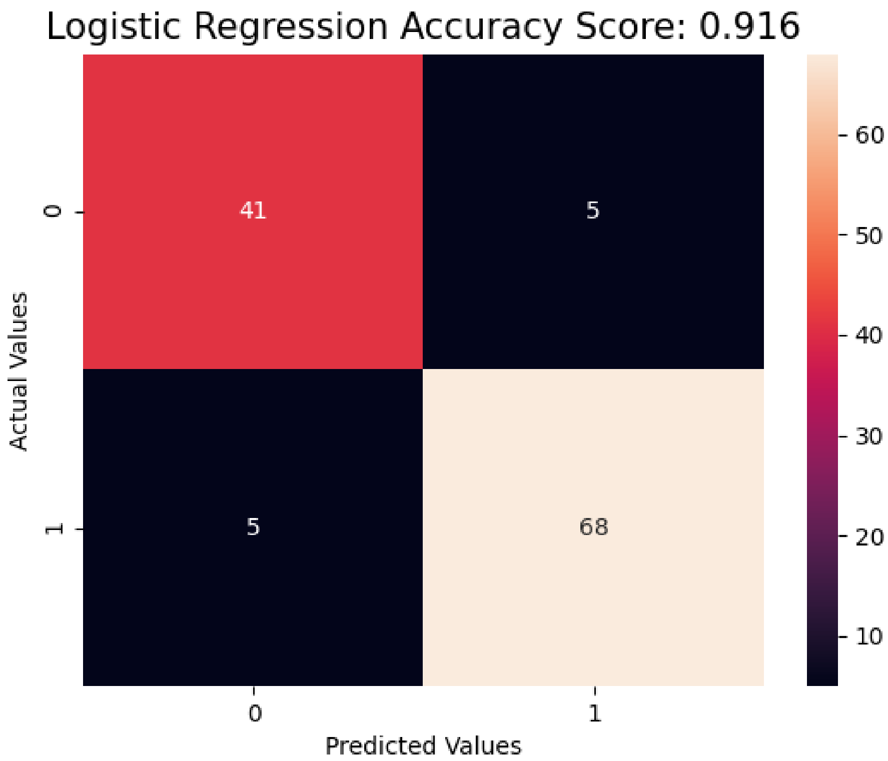


Figure 26 Confusion matrix with Logistic Regression for 395 students

True Negative (TN), False Positive (FP), False Negative (FN), True Positive (TP) values with logistic regression for 395 students are clearly shown in the Table 10.

Table 10 Details of Confusion matrix with Logistic Regression for 395 students

True Negative (TN)	41
False Positive (FP)	5
False Negative (FN)	5
True Positive (TP)	68

The sum of these values is 30% of the total dataset. Because we reserved 30% of our data for testing.

	precision	recall	f1-score	support
Grade >= 10	0.89	0.89	0.89	46
Grade < 10	0.93	0.93	0.93	73
accuracy			0.92	119
macro avg	0.91	0.91	0.91	119
weighted avg	0.92	0.92	0.92	119

Figure 27 Classification Report of Logistic Regression for 395 students

When logistic regression classification report is generated for 395 students, as you can see in Figure 27, accuracy, precision, recall, f1-score, support, macro average, weighted average values are reached.

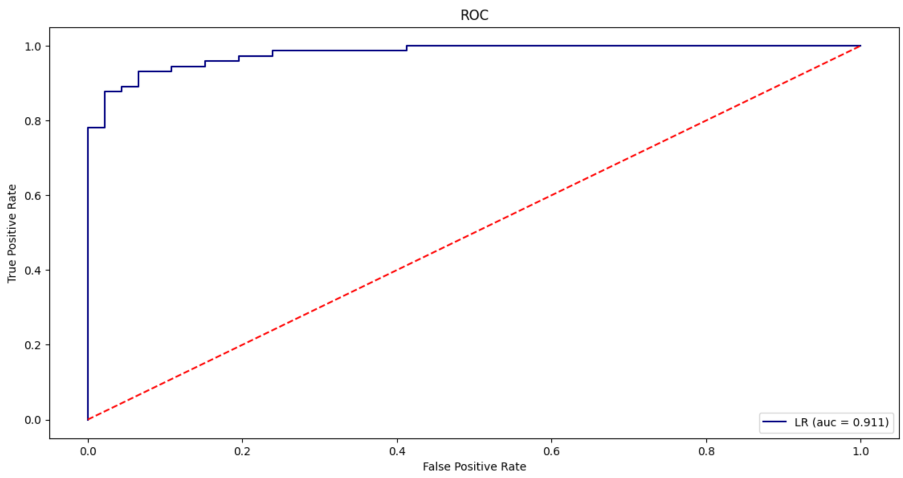


Figure 28 ROC Curve with Logistic Regression for 395 students

ROC Curve in Figure 28 shows that the AUC value is 0.911 for 395 students. We know that the higher the AUC value, the better the prediction. The ideal value for AUC is 1. 0.911 is a good result compared to 1.

- **For 245 students**

Firstly, the confusion matrix is mentioned. With the logistic regression algorithm, the accuracy rate for 245 students is 0.878. The confusion matrix with logistic regression algorithm for 245 students is shown with Figure 29.

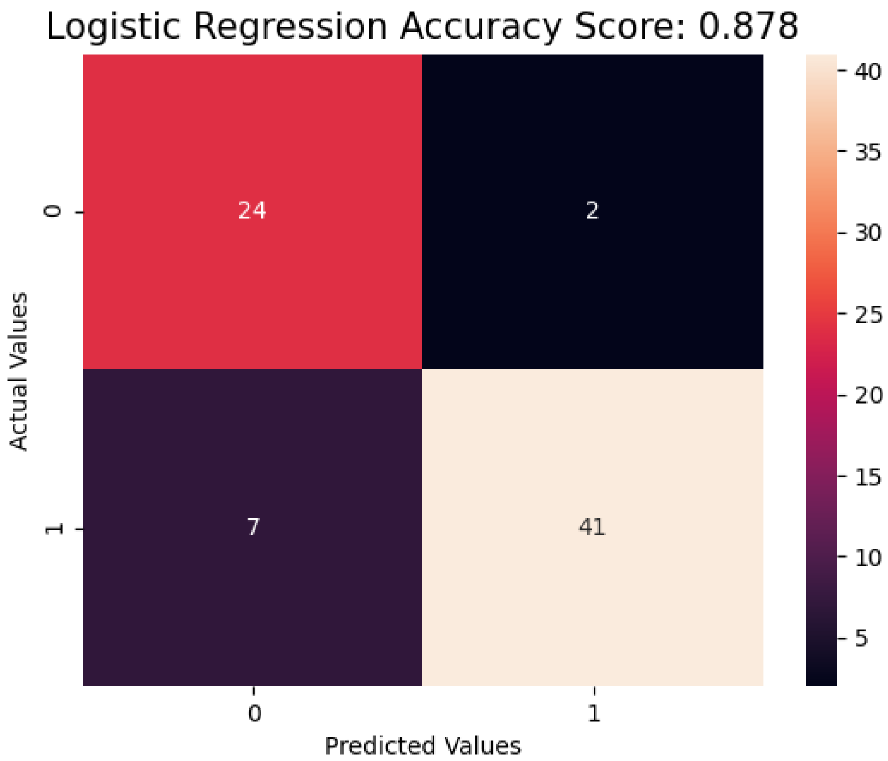


Figure 29 Confusion matrix with Logistic Regression for 245 students

True Negative (TN), False Positive (FP), False Negative (FN), True Positive (TP) values with logistic regression for 245 students are clearly shown in the Table 11.

Table 11 Details of Confusion matrix with Logistic Regression for 245 students

True Negative (TN)	24
False Positive (FP)	2
False Negative (FN)	7
True Positive (TP)	41

The sum of these values is 30% of the total dataset. Because we reserved 30% of our data for testing.

	precision	recall	f1-score	support
Grade >= 10	0.77	0.92	0.84	26
Grade < 10	0.95	0.85	0.90	48
accuracy			0.88	74
macro avg	0.86	0.89	0.87	74
weighted avg	0.89	0.88	0.88	74

Figure 30 Classification Report of Logistic Regression for 245 students

When logistic regression classification report is generated for 245 students, as you can see in Figure 30, accuracy, precision, recall, f1-score, support, macro average, weighted average values are reached.

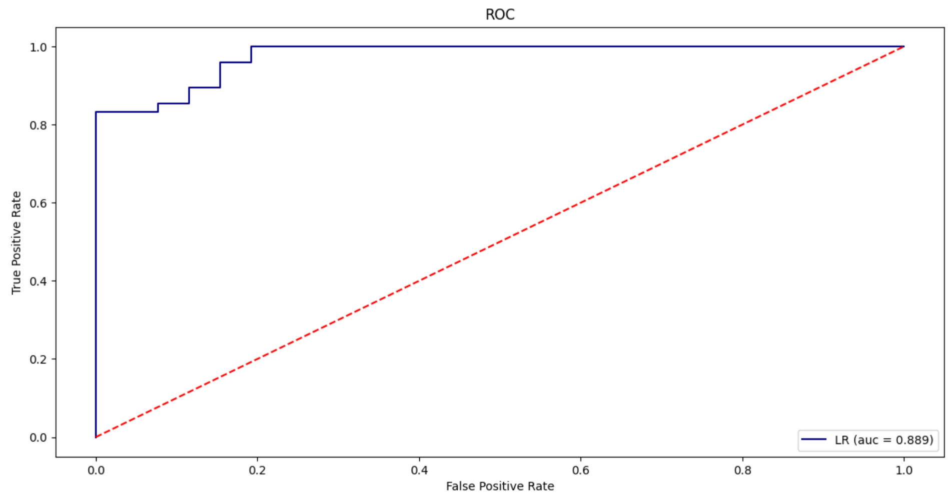


Figure 31 ROC Curve with Logistic Regression for 245 students

ROC Curve in Figure 25 shows that the AUC value is 0.889 for 245 students. We know that the higher the AUC value, the better the prediction. The ideal value for AUC is 1. 0.899 is a good result compared to 1.

CHAPTER 6
EXPERIMENTS

6.1. EXPERIMENTAL SETUP

We done the coding via VSCode IDE to make a prediction. PYTHON 3 programming language is used for coding. The dataset is kept as a csv file. Pandas, numpy, matplotlib.pyplot and seaborn libraries are used.

The machine used for our study has an Intel Core i5 processor. This machine has NVIDIA GeForce GT 940M graphics card. The operating system used is Windows.

6.2. EXPERIMENTAL RESULTS

We worked with three different algorithms to predict students' academic performance. These algorithms are decision tree, random forest and logistic regression. We changed the data size by increasing or decreasing the number of students.

For test data, 30 percent of all data was used, and for training, 70% of the collected data was used. Since there is a tree structure in the decision tree and random forest algorithms, we calculated the accuracy value according to the max depth values in these algorithms. We calculated a total of 20 accuracy values for the three algorithms. These values are shown in detail in the Table 12 below.

Table 12 Accuracy rate results of three algorithms

Algorithm Name	Number of Students	Accuracy Rate	Max-Depth
Decision Tree	649	0.949	2
		0.933	8
		0.933	-
	395	0.924	2
		0.849	8
		0.849	-
	245	0.905	2
		0.878	8
		0.878	-
Random Forest	649	0.897	2
		0.923	8
		0.933	-
	395	0.714	2
		0.908	8
		0.899	-
	245	0.770	2
		0.865	8
		0.851	-
Logistic Regression	649	0.908	-
	395	0.916	-
	245	0.878	-

Table 13 shows the comparison of the AUC values, which is the area under the ROC curves that we extracted using the logistic regression algorithm, according to the number of students.

Table 13 AUC values with logistic regression

Algorithm Name	Number of Students	AUC Value
Logistic Regression	649	0.800
	395	0.911
	245	0.889

When both tables are evaluated, decision tree algorithm gives the best accuracy rate with max depth 2 value with 649 student data. This ratio is 0.949. The random forest algorithm gives the best accuracy with 649 student data. This ratio is 0.933. The logistic regression algorithm gives the best accuracy with 395 student data. This ratio is 0.916. The AUC value also supports the best accuracy rate given by the logistic regression algorithm. The best AUC value is 0.911 with 395 students. Comparison of ROC curves is shown in the Figure 32. Also, the best AUC value is displayed as 0.911 in this figure.

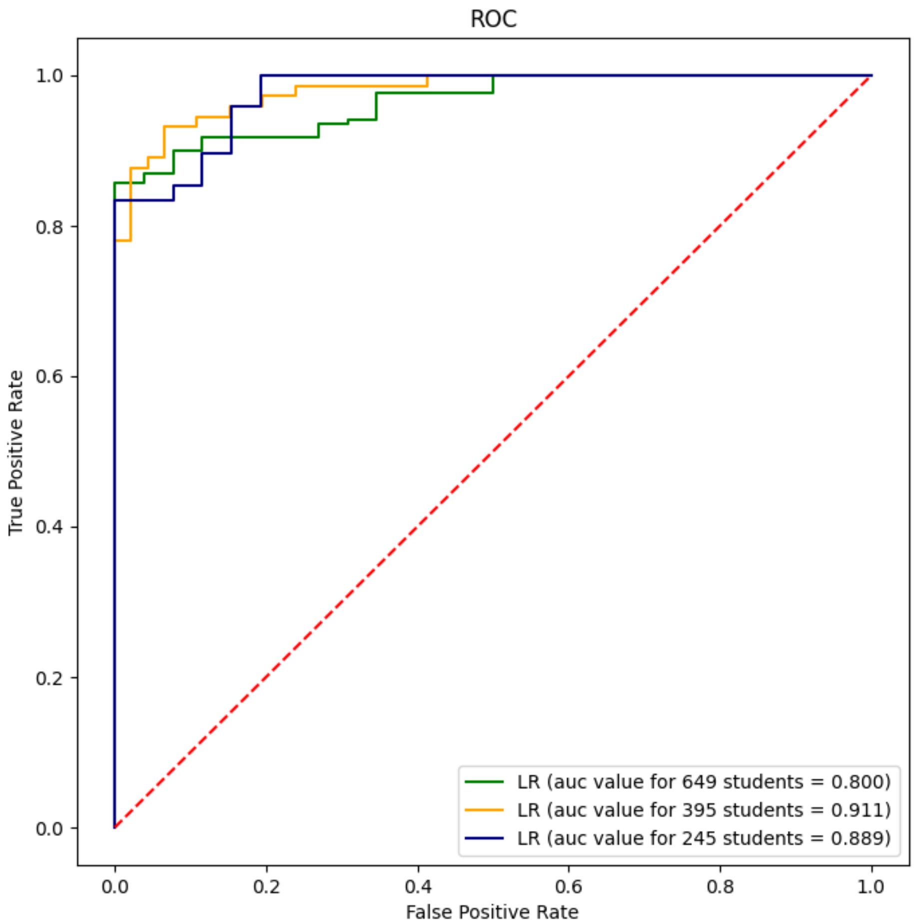


Figure 32 Comparison of ROC Curves

The tree structure extracted for the random forest and decision tree algorithm is shown in Figure 33 and Figure 34.

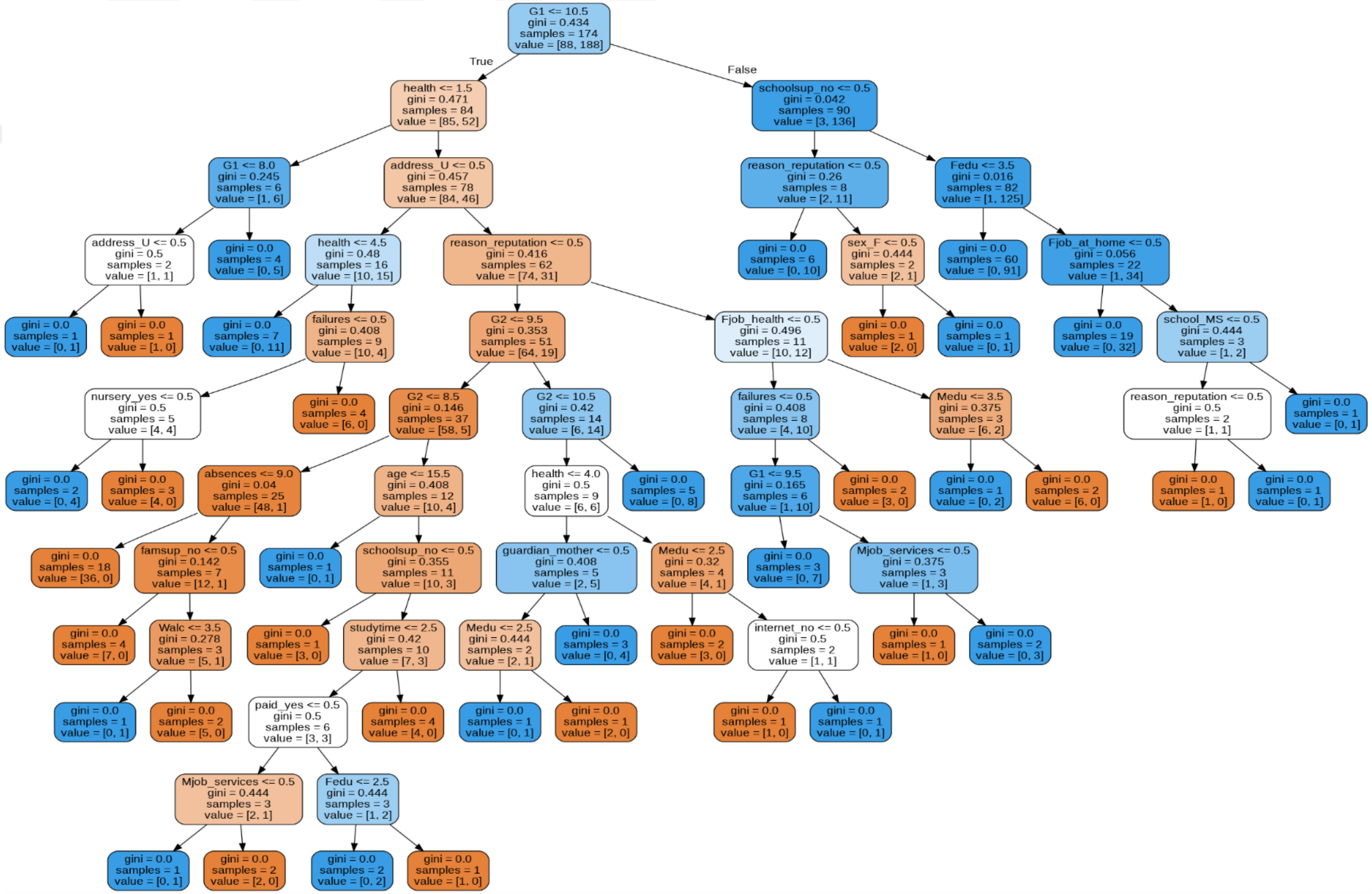


Figure 33 Tree structure with random forest algorithm

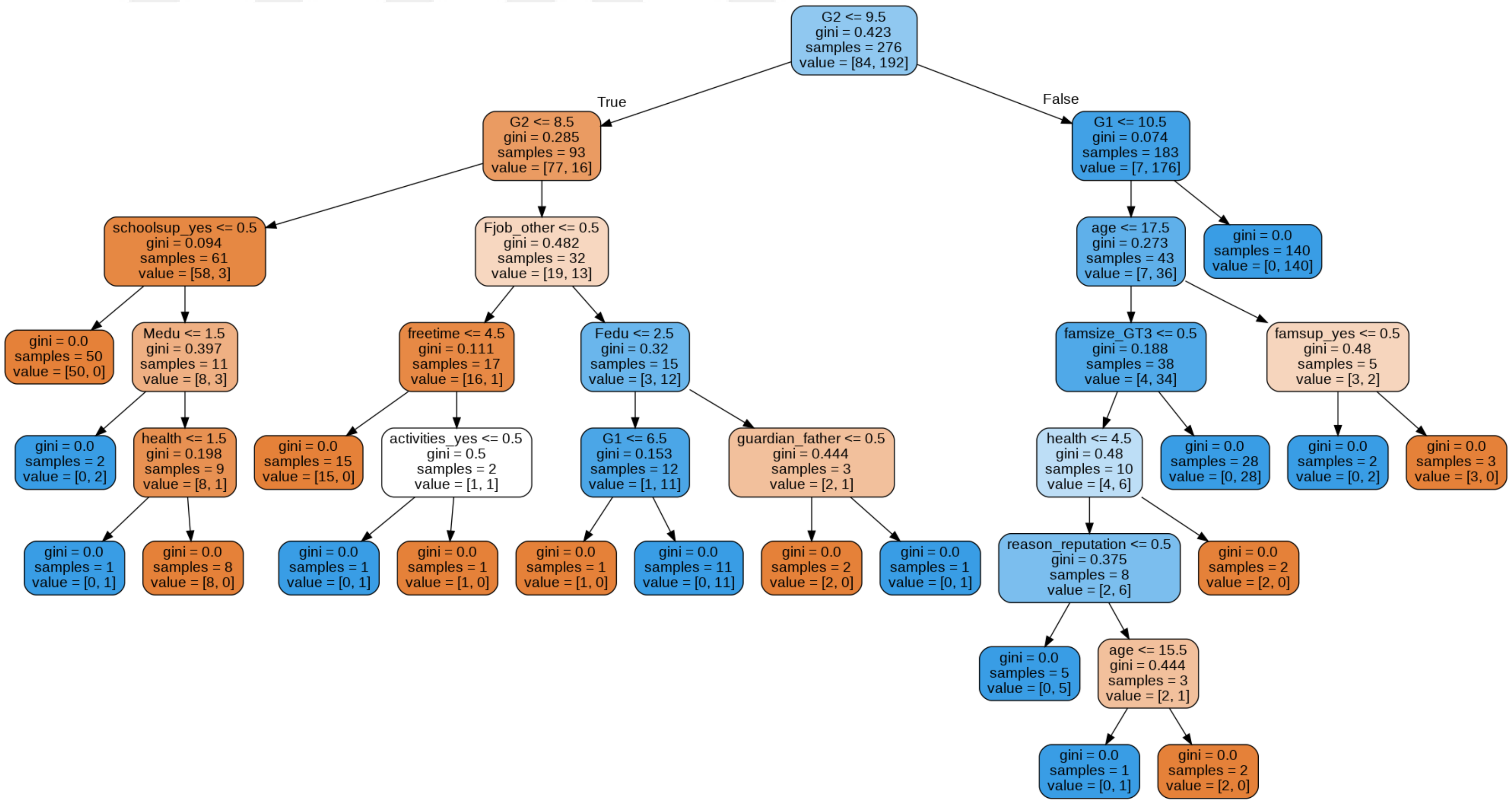


Figure 34 Tree structure with decision tree algorithm

CHAPTER 7
CONCLUSIONS AND FUTURE

With this study, we have shown that academic performance is predicted by machine learning algorithms and the accuracy of this prediction is numerical values. Moreover, a comparison of performance prediction accuracy with three different algorithms was made. This study shows that performance prediction can be made with students' demographic, grades information, social and school related features.

This study shows that when working with the highest number of students, the best accuracy rate belongs to the Decision Tree algorithm. In addition, the best accuracy rate for the lowest number of students belongs to the Decision Tree algorithm.

In future studies, the number of students can be increased and the rates can be compared. On the other hand, predictions can be made with support vector machine (SVM) algorithm to achieve higher accuracy or several algorithms can be combined and a new algorithm can be written and predictions can be made for the same number of students.

REFERENCES

[1] M. Balaban and E. Kartal, *Veri Madenciliği ve Makine Öğrenmesi Temel Algoritmaları ve R Dili ile Uygulamaları*, 2nd ed. İstanbul: Çağlayan Kitabevi, 2018, p. 4.

[2] R. Kurzweil, The Age of Intelligent Machines. Cambridge, Massachusetts: MIT “Press, 1990.

[3] G. Luger and W. Stubblefield, Artificial Intelligence: Structures and Strategies for Complex Problem Solving, 2nd ed. Harlow, England: Addison-Wesley, 1993.

[4] S. Russell and P. Norvig, *Artificial intelligence: A Modern Approach*. Englewood Cliffs, New Jersey: Prentice Hall, 1995, pp. 4-8.

[5] N.J. Nilsson, *The Quest for Artificial Intelligence: A Hıstory of Ideas and Achıevements*. California, United States: Cambridge University Press, 2009.

[6] E. Öztemel, *Yapay sinir ağları*, 2nd ed. İstanbul: Papatya Yayıncılık, 2006, pp. 15-28.

[7] D. Graupe, *Principles of artificial neural networks*, 2nd ed. Singapore: World Scientific, 2007, pp. 1-3.

[8] M. Balaban and E. Kartal, *Veri Madenciliği ve Makine Öğrenmesi Temel Algoritmaları ve R Dili ile Uygulamaları,* 2nd ed. İstanbul: Çağlayan Kitabevi, 2018, p. 10.

[9] M. Rychener, *Expert Systems for Engineering Design*, 1st ed. United States of America: Academic Press, 1998, pp. 4-6.

[10] L. Zadeh, "Fuzzy sets", Information and Control, vol. 8, no. 3, pp. 338-353, 1965.

[11] M. Sarı, Y. Murat and M. Kırabalı, "Fuzzy Modelling Approach And Applications", Journal of Pamukkale University, vol. 9, pp. 77-92, 2005.

[12] D. Goldberg, Genetic algorithms in search, optimization, and machine learning, 1st ed. Boston: Addison-Wesley, 1989, pp. 1-6.

[13] Ç. Taşkın and G. Emel, "Genetik Algoritmalar ve Uygulama Alanları", Uludağ Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, vol. 21, no. 1, pp. 129-152, 2002.

[14] B. Bolat, K. Erol and C. İmrak, "Genetıc Algorıthms In Engıneerıng Applıcatıons And The Functıon Of Operators", Mühendislik ve Fen Bilimleri Dergisi, vol. 4, pp. 264-271, 2004.

[15] E. Karamanlı, Makine öğrenmesi algoritmaları kullanarak, metin madenciliği ve duygu analizi ile müşteri deneyiminin geliştirilmesi. Yüksek Lisans Tezi. İstanbul: İstanbul Üniversitesi, 2019, p. 6.

[16] A. Müller and S. Guido, Introduction to Machine Learning with Python: A Guide for Data Scientists, 1st ed. United States of America: O'Reilly Media, 2016, pp. 1-6.

[17] F. Camastra and A. Vinciarelli, Machine Learning for Audio, Image and Video Analysis: Theory and Applications, 2nd ed. Springer, 2015.

[18] M. Mohri, A. Rostamizadeh and A. Talwalkar, Foundations of machine learning, 2nd ed. Cambridge, UNITED STATES: MIT Press, 2018.

[19] R. Sutton and A. Barto, Reinforcement Learning: an Introduction, 1st ed. Cambridge: MIT Press, 1998.

[20] M. Dunham, Data mining introductory and advanced topics. Upper Saddle River (New Jersey): Pearson Education, 2003, pp. 58-61.

[21] T. Özcan, Veri Madenciliği ve Bilgi Keşfi: Karar Ağaçları ile Sınıflandırma. 2015.

[22] K. Archer and R. Kimes, "Empirical characterization of random forest variable importance measures", Computational Statistics & Data Analysis, vol. 52, no. 4, pp. 2249-2260, 2008. Available: 10.1016/j.csda.2007.08.015

[23] L. Breiman and A. Cutler, "Random forests - classification description", Stat.berkeley.edu, 2005. [Online]. Available: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm.

[24] J. Watts, S. Powell, R. Lawrence and T. Hilker, "Improved classification of conservation tillage adoption using high temporal and synthetic satellite imagery", Remote Sensing of Environment, vol. 115, no. 1, pp. 66-75, 2011. Available: 10.1016/j.rse.2010.08.005.

[25] N. Horning, "RandomForests: An algorithm for image classification and generation of continuous fields data sets", International Conference on Geoinformatics

for Spatial Infrastructure Development in Earth and Allied Sciences (GISIDEAS) 2010, 2010.

[26] L. Breiman, "Manual on setting up, using, and understanding random forests V3.1", Stat.berkeley.edu, 2002. [Online]. Available: https://www.stat.berkeley.edu/~breiman/Using_random_forests_V3.1.pdf. Accessed: 27- Oct- 2021].

[27] Ö. Çokluk, "Lojistik regresyon analizi: kavram ve uygulama", Kuram ve Uygulamada Eğitim Bilimleri, vol. 10, no. 3, pp. 1357 - 1407, 2010.

[28] R. Alpar, Uygulamalı çok değişkenli istatistiksel yöntemler. Ankara: Detay Yayıncılık, 2013, pp. 591-660.

[29] I. Schuster and P. Jähnichen, "Classification using Logistic Regression", Asv.informatik.uni-leipzig.de, 2013. [Online]. Available: http://asv.informatik.uni-leipzig.de/uploads/document/file_link/530/TMI05.2_logistic_regression.pdf.

[30] M. Wook, Y. Yahaya, N. Wahab, M. Isa, N. Awang and H. Seong, "Predicting NDUM Student's Academic Performance Using Data Mining Techniques", 2009 Second International Conference on Computer and Electrical Engineering, 2009. Available: 10.1109/iccee.2009.168

[31] D. Şengür, "Öğrencilerin Akademik Başarılarının Veri Madenciliği Metotları ile Tahmini", Fırat Üniversitesi, Eğitim Bilimleri Enstitüsü, Doktora Tezi, 2013.

[32] T.C. Hakyemez, "İlk Yıl Öğrencilerinin Akademik Performansına Etki Eden Faktörlerin Araştırılması ve Bu Faktörlere Bağlı Olarak Başarılarının Tahminine Yönelik Bir Karar Destek Sistemi Tasarım", Sakarya Üniversitesi, Sosyal Bilimler Enstitüsü, Doktor Tezi, 2015.

[33] Ş. Özdemir, "Eğitimde Veri Madenciliği ve Öğrenci Akademik Başarı Öngörüsüne İlişkin Bir Uygulama", İstanbul Üniversitesi, Fen Bilimleri Enstitüsü, Doktora Tezi, 2016.

[34] Z. Pardos, N. Heffernan, B. Anderson and C. Heffernan, "Using Fine-Grained Skill Models to Fit Student Performance with Bayesian Networks", 2006. Available: 10.1201/b10274-32

[35] P. Cortez and A. Silva, "Using Data Mining To Predict Secondary School Student Performance", pp. 5-12, 2008.

[36] P. Guleria, N. Thakur, and M. Sood, "Predicting student performance using decision tree classifiers and information gain," Proc. 2014 3rd Int. Conf. Parallel, Distrib. Grid Comput. PDGC 2014, pp. 126–129, 2015.

[37] G. Gray, C. McGuinness, and P. Owende, “An application of classification models to predict learner progression in tertiary education,” in Souvenir of the 2014 IEEE International Advance Computing Conference, IACC 2014, 2014.

[38] I. E. Livieris, K. Drakopoulou, and P. Pintelas, “Predicting students’ performance using artificial neural networks,” Proc. 8th Pan-Hellenic Conf. “Information Commun. Technol. Educ., pp. 28–30, 2012.

[39] F. Sarker, T. Tiropanis, and H. C. Davis, “Linked data, data mining and external open data for better prediction of at-risk students,” in Proceedings - 2014 International Conference on Control, Decision and Information Technologies, CoDIT 2014, 2014.

[40] T. Fawcett, "An introduction to ROC analysis", Pattern Recognition Letters, vol. 27, no. 8, pp. 861-874, 2006