Predicting Students' Academic Performances Using Machine Learning Algorithms in Educational Data Mining

Şenay KOCAKOYUN AYDOĞAN [1], Turgut PURA [2], Fatih BİNGÜL [3]

To Cite: Kocakoyun Aydoğan, Ş., Pura, T. & Bingül, F. (2024). Predicting students' academic performances using machine learning algorithms in educational data mining. *Malaysian Online Journal of Educational Technology, 12*(4), 131-153. http://dx.doi.org/10.52380/mojet.2024.12.4.557

[1] senay.aydogan@gedik.edu.tr, orcıd.org/0000-0002-3405-6497, İstanbul Gedik University,Türkiye

[2] turgut.pura@gedik.edu.tr, orcıd.org/0000-0002-4108-8518, İstanbul Gedik University, Türkiye

[3] fatihbingul@beykoz.edu.tr, orcıd.org/0000-0001-8246-2345, Beykoz University, Türkiye

ABSTRACT

In every culture and era, education is considered the most fundamental reality and rule that societies prioritize and deem essential. Throughout the process spanning thousands of years, from the emergence of writing to the present day, education has undergone various forms and formats of change. Education has been a continuous guide for shaping, influencing, sustaining societies, and maintaining its dynamics throughout these historical processes. The continuous evolution and growth of education systems and formats worldwide, with changes affecting the quality of education, have the potential to influence nations and societies in every field, ultimately leading to the emergence of an informed society, achievable only through quality education. In this study, the aim is to determine the factors affecting students' academic performance and predict students' end-of-term academic grades using machine learning algorithms within the scope of Earned Value Management (EVM). Such studies have great potential to increase efficiency in education, improve student achievement and improve education policies. With the use of machine learning algorithms, these goals can be achieved more quickly and efficiently. Five different machine learning algorithms, namely RF, KA, KNN, SVM, and NB, have been employed in the study. Binary and multiclass classification methods were used in prediction processes, and among these methods, the Random Forest (RF) algorithm achieved the highest success prediction rates of 0.97 and 0.93, respectively, in both classification methods.

Keywords: *EVK, random forest, education, machine learning*

Article History:
Received: 23 May 2024
Received in revised form: 3 July 2024
Accepted: 2 October 2024
Article type: Research Article

INTRODUCTION

In every culture and era, education has been acknowledged as one of the most fundamental realities and rules that all societies prioritize, emphasize, and deem essential. Over the thousands of years since the emergence of writing to the present day, education has undergone various forms and formats of change. Throughout these processes, education, which shapes and influences societies and cultures while simultaneously guiding continuous change, adapting to development, and preserving its dynamic structure, has been a phenomenon. The continuous evolution and growth of education systems and formats worldwide, undergoing constant changes, as highlighted by Abaidullah et al. (2015), contribute to the impact of education quality on nations and the societies within them in all fields. The emergence of an informed society is deemed possible only through quality education. In the current twenty-first century, characterized by the abundant emergence of knowledge, there is an observed increase in the need for education across all societies, as noted by Şahin and Tekdal (2005).

In today's rapidly advancing technological era, the volume of data is increasing rapidly in the field of education, as well as in all other areas, each passing day (Larose & Larose, 2014). The swift increase in data has led to recent applications of data mining techniques in the field of education, addressing a significant gap and contributing to the emergence of the concept of Educational Data Mining (EDM). These applications aim to enhance the impact, quality, and efficiency of educational systems (Peña-Ayala, 2014). Educational data mining is one of the applications of data mining, and its usage has become widespread in recent times, especially in student and school assessment processes. Data mining models used to extract relevant information from databases are generally categorized into two types: descriptive and predictive models (Romero & Ventura, 2007). The application of educational data mining in educational systems provides educational benefits in various fields for students, educators, academic officials, and administrators. When viewed by all participants in the education system, the products obtained through educational data mining applications offer the possibility of foresight that can benefit each participant in the system (Calders & Pechenizkiy, 2012). For example, aspects such as students' interest in the class, study durations, ages, family situations, genders, and socioeconomic statuses can be analyzed and predicted alongside end-of-term grades (Abdous et al., 2012). Additionally, the ability to predict potential glitches in the education and teaching process in advance is crucial. Rectifying and improving these glitches is achievable through educational data mining (Can, 2017). Due to these reasons, there is a proportional relationship between the quality of the education system and students' positive performance. As the quality of the education system increases, students' achievements also increase linearly (Andonie, 2010). In light of all these considerations, EVM is a technique that arises from the intersection of computer-aided sciences and educational sciences created by technology (Peña-Ayala, 2014).

In this study, examples and quotations will be taken from EVM applications related to student performance in educational and teaching processes. A dataset that analyzes student performance in terms of 33 attributes, available in the Irvine Machine Learning Repository, was used in this study (Cortez & Silva, 2008). The features of the dataset will be further detailed in the subsequent sections. Using this dataset as a foundation, various machine learning methods were applied to analyze and predict student performance, resulting in successful outcomes. Additionally, the aim is to contribute to future studies involving various applications of EVM. The logical progression of the study is as follows:

- Some of the past studies on student performance related to EVM will be examined.
- The characteristics and operations of the machine learning methods to be used in the thesis will be discussed.
- Performance predictions will be made on the selected attributes of the dataset to be used in this study using machine learning methods.
- The values of the prediction results will be analyzed.
- A comprehensive evaluation of the results of our study will be conducted, and conclusions will be drawn for future applications.

Education and Instruction

Education is considered an effort to intentionally and purposefully cultivate individuals with desired characteristics that align with the culture of society (Koçer, 1980). Instruction encompasses the planned, systematic education process delivered by specialized educators during an individual's school life. When an individual, or student, enters the education and instructional process, how instruction will be conducted, its duration, and all instructional processes are planned and presented (Eğiticinin Eğitimi, 2022). Instruction, like education, aims to instill positive and desired behaviors in the individual. Therefore, education encompasses instruction. If desired behaviors are instilled in the individual through instruction, then instruction turns into education (Akyüz, 1982).

Factors Affecting Academic Performance

All individuals are born, raised, and live within a specific culture. Their beliefs, languages, and adherence to the moral rules of the society they are in vary accordingly. All these rules also shape a person's education and learning life (Ergün, 1994). Studies on this subject address the effects of culture on learning from different perspectives. For example, a report by the Learning Forward organization emphasizes that the impact of culture on learning is not limited to students only, but teachers' teaching methods and educational policies are also affected by this interaction. This report states that cultural awareness is critical to ensuring justice and equity in education (Learning Forward, 2021).

Throughout the process of education and instruction, various factors influence students' success, starting from the family environment. Among these factors are geographical location, socio-economic status, physical and mental conditions, societal structure, and influential characteristics of the environment. These factors sometimes positively, sometimes negatively affect students' achievements. Therefore, students need to be shielded from all negative influences to perform at their highest level throughout their education and learning life (Alkan and Kurt, 1998).

In research on factors affecting student success, it has been found that various elements play an important role. Geographic location and socio-economic status stand out as critical factors in students' educational life. In addition, family environment, school resources and teacher quality also have decisive effects on success. For example, Shulruf, Hattie and Tumen (2009) comprehensively examined the effects of a student's family, school, teacher, and curriculum on success. Additionally, researchers such as Hughes and Pearce (2003) have analyzed how socio-economic status and geographic location affect students' participation and success in higher education. These findings show that environmental and individual factors must be carefully managed in order for students to be successful in education.

Academic success varies among students due to differences in their abilities, intelligence levels, research, and individual thinking characteristics. The learning coefficient, therefore, differs in each student. It is accepted that there is a positive correlation between intelligence and success (Yıldırım, 2000). Intelligence level is an important factor that directly affects students' academic success. Meta-analysis studies have shown that intelligence is a strong predictor of academic performance (Lozano-Blasco et al. 2022). Throughout the learning process, individuals, or students, should not be passive but demonstrate an active attitude, engage in success-oriented work, and be associated with the degree of self-regulation and self-efficacy. Students' ability to regulate their own learning processes and their motivation levels also play a decisive role in success. Salili, Chiu and Lai, (2001) studies examined the effects of motivational and self-regulated learning components on academic performance. The degree of motivation, which is related to the desire for learning, is also associated with factors such as the time allocated for learning, the attitude towards learning, the problems encountered in the learning process, and the ability to overcome these problems. Individuals with high motivation actively participate in the process of education and instruction (Alemdağ et al., 2014).

Literature Review

While reviewing the studies conducted in past periods, research utilizing Earned Value Management (EVM) and machine learning methods was surveyed. Studies that would contribute to this thesis were selected and evaluated based on the combined use of EVM and machine learning methods.

Upon evaluating these studies, it was observed that in 57 studies conducted between 2015 and 2020 (22.5%), decision tree algorithms were preferred. Naive Bayes classification algorithms were employed in 36 studies (14.2%), and Bayesian algorithms were frequently used in conjunction with decision tree structures. There were 45 studies (17.8%) involving Support Vector Machine algorithms, with regression models (linear-logistic) often present in these studies. Artificial Neural Networks were preferred in 20 studies and generally used as a single model (Tosunoğlu et al., 2021).

Studies on Educational Data Mining

In his study on computer usage and its implications, Petcu (2015) aimed to present quantitative research results on various aspects related to computer usage in the university learning process. The study covered areas such as the fields where computers are used, the impact and significance of computer usage in university education, faculty activities requiring computer usage, internet and website usage, as well as investments in information and communication technology and the accessibility of information technology resources. The research, conducted at Transylvania University, analyzed students' opinions regarding the computerization of the education system. Baradwaj and Pal (2012) opted for the categorization technique to analyze students' performance and the decision tree algorithm to separate data based on similarity. Through this applied algorithm, they revealed data explaining students' scores in the year-end exam. In this context, they arrived at the notion that instructors could serve as guiding mentors for students showing special needs or situations like leaving school. In his study on the emerging applications and trends in the field of computer science within the education system, Baran (2013) emphasized the close relationship between learning styles and educational data mining. Baker and Yacef (2009) conducted 45 studies on EVM, analyzing the reasons for the use of EVM based on their objectives. They classified these 45 studies into three main groups: improving student forms used in online education environments, enhancing domain forms, and providing pedagogical support to students along with presenting scientific studies related to their learning. Kumar et al. (2017) conducted a literature review to determine separate models for predicting the performance of students with different features, aiming to analyze the differences between predictive models using data mining in education. In a study by Erdoğan and Timor (2005), the correlation between students' university exam success scores and academic success scores during the university process was analyzed using the k-means algorithm, a machine learning method. Ayık, Özdemir and Yavuz (2007) analyzed the correlation values and relationships between the types of high schools attended by students at Ataturk University, their high school ranking scores, and the faculties where they were placed at the university. This analysis was carried out using educational data mining techniques.

Studies on Student Performances

In their study on predicting student performance, Ghorbani and Ghousi (2020) achieved the highest success rate using the RF algorithm among classification machine learning methods, including RF, KNN, ANN, XGBoost, SVM, DT, LR, and NB. Şengür and Tekin (2013) used machine learning algorithms, specifically YSA and KA methods, to predict students' graduation degrees. They found that the YSA machine learning method was more successful than the KA method. Ibrahim and Rusli (2007) utilized three different machine learning algorithms, namely YSA, KA, and linear regression, to predict students' academic success. In this study, the general GPA of the students was considered as a significant factor, while factors such as family economic status, the school's education system, and the extent to which the student used information technologies were deemed less influential. The YSA algorithm provided the best prediction results in this study. Bhardwaj and Pal (2012) employed the Naive Bayes classification algorithm in their study conducted in Faizabad, India, to identify the factors that most influenced students' success during the semester. Mayilvaganan and Kalpanadevi (2014) used the KNN machine learning classification algorithm to predict students' academic performance in their study. Oloruntoba and Akinode (2017) used the support vector machine (SVM) machine learning algorithm to predict students' academic performance. They compared SVM with other classification algorithms and found that SVM outperformed the others. Belachew and Gobena (2017) used SVM, NB, and YSA algorithms to predict student performances in their study, achieving the highest prediction rate of 95.7% with the YSA algorithm.

Almarabeh (2017) used five different machine learning methods, namely Bayesian Network, J48, Naive Bayes, ID3, and Neural Network, to predict and analyze students' performances. Bayesian Network achieved the best accuracy rate among these algorithms. Qasem et al. (2011) applied a decision tree algorithm to help students choose the best academic process after graduation, achieving an accuracy rate of 87.9%. Osmanbegovic and Suljic (2012) used the NB algorithm to predict students' academic success based on data from Tuzla University, obtaining an accuracy prediction of 76.65%. Yukselturk et al. (2014) employed three different machine learning algorithms—KNN, NB, and ANN—to predict students' likelihood of dropping out

of school. The NB algorithm achieved a success rate of 70%, KNN achieved 87%, and the KA method with 10-fold cross-validation achieved 79.7%. Ramesh, Ramesh and Ramar (2013) conducted a study in the Kancheepuram region of India, applying the NB, KA, Multilayer Perceptron, and Reduced Error Pruning methods to identify features affecting students' end-of-year academic performance and predict their final grades. Among these algorithms, the Multilayer Perceptron algorithm achieved the highest accuracy success rate of 72.38%. Ahmad et al. (2015) created a dataset pool for university students taking computer science courses to predict their academic success. They used the KA, NB, and Rule-Based algorithms for academic success prediction, with the Rule-Based classification algorithm providing the best result at 71.3%.

METHOD

Materials

The dataset to be used in this study was obtained from two secondary schools in Portugal by Cortez and Silva (2008). The attributes and features in the dataset were acquired through a survey. The dataset consists of 33 variables and 649 samples. The characteristics of the dataset will be examined in detail in a subsection.

The dataset is structured under classification and regression models. The dataset is designed to predict the academic performance of the students, which is also the focus of this thesis. The dependent variable in the dataset will be considered as the end-of-term grade. The features in the dataset are shown in Table 1.

Table 1. *Feature Names In The Data Set*

Serial No	Feature Name	Serial No	Feature Name
1	School		
2	Gender		
3	Age		
4	Address		
5	Number of Family Members		
6	Marital Status of Parents (Married or Separated)		
7	Mother's Educational Level		
8	Father's Educational Level		
9	Mother's Occupation		
10	Father's Occupation		
11	Reason for Choosing the School		
12	Student's Guardian - Mother or Father		
13	Commute Time to School		
14	Weekly Study Time		
15	Grade Repetition		
16	School Support		
17	Family Educational Support		
18	Private Tutoring		
19	Participation in Social Activities		
20	Preschool Education, presence or absence		
21	Desire to Pursue a Bachelor's Degree		
22	Availability of Internet at Home		
23	Relationship Status		
24	Level of Relationship with Family		
25	Non-School Leisure Time		
26	Going Out with Friends		
27	Alcohol Consumption During School Hours		
28	Alcohol Consumption During Weekends		
29	Current Health Status		
30	School Absenteeism		
31	First Semester Grade		
32	Second Semester Grade		
33	Final Grade		

Data Processing Tools

Python: Python, a programming language, was initiated by Guido Van Rossum, a Dutch national and a programmer, in the 1990s. Python, in terms of syntax, is considered to be easier than other languages and

does not require any compiler (Özgül, 2013). It is a free and open-source programming language. Python stands out from other languages due to the existence of numerous libraries. Among these libraries, prominent ones include TensorFlow, Keras, and Scikit-learn. In addition to these libraries, there are also libraries commonly used in data science and machine learning, such as pandas, numpy, and matplotlib, which serve data analysis and visualization purposes (Arslan, 2019).

Pandas Library: The pandas library, developed for Python, facilitates many operations to be performed more quickly and easily. The pandas library is a library developed for the Python programming language. With pandas, time series, numerical array operations, and data structures can be easily created and processed (Chen, 2017). To install the library, it is sufficient to use the "import pandas" command in our application (Pandas, 2024).

Numpy Library: It is a Python library developed for performing mathematical array operations quickly and easily in multidimensional arrays (Harris et al., 2020). To install the numpy library, simply type the "import numpy" command (NumPy, 2024).

Scikit-Learn Library: Like many other free libraries within Python, the Scikit-learn library is also open source. The Scikit-learn library is one of the most widely used Python libraries for machine learning and data science. Due to providing a variety of algorithms, it enables the execution of multiple tasks (Karabay, 2024).

Matplotlib Library: The Matplotlib library is a tool for visualizing data in static and interactive formats, with specific formats and various types of graphs. It allows for editable, zoomable graphics and enables obtaining high-quality outputs simultaneously (Matplotlib, 2024).

Anaconda: Anaconda is a platform developed for performing tasks in applications such as machine learning, data analysis, and data science. It is a unified Python ecosystem that encompasses many application software and libraries. The Anaconda platform includes the Jupyter Notebook tool, which we will use in this thesis (Medium, 2024). The Anaconda platform is illustrated in Figure 1 below.

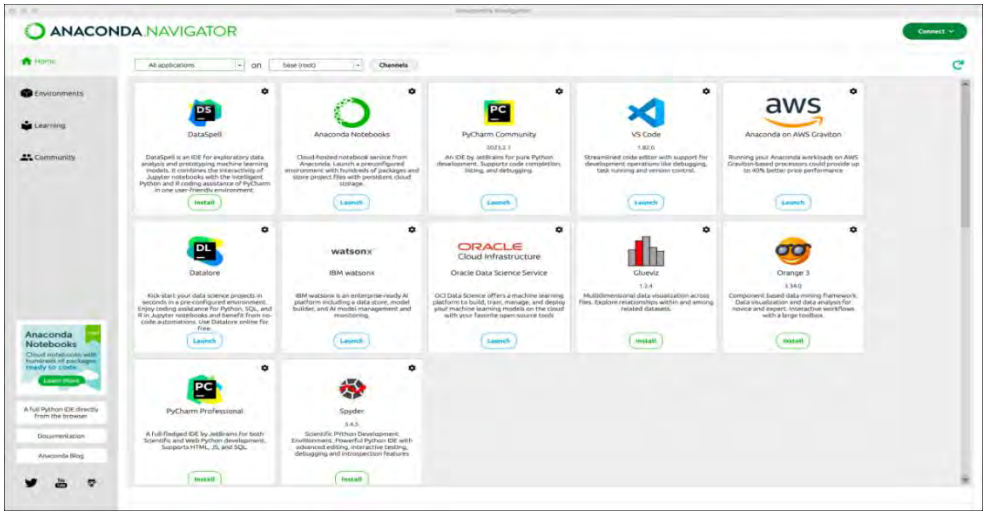

Figure 1. *The Anaconda Platform* (Anaconda, 2024)

Jupyter Notebook: Jupyter Notebook is software that provides support for programming in Python as well as other programming languages. The interface of Jupyter Notebook is very simple and straightforward. This simplicity allows the written code to be easily readable and compiled, making it particularly popular among those learning software development, especially the Python programming language (Tuzcu, 2020). The code screen and interface of the Jupyter Notebook application are illustrated in Figure 2 below.

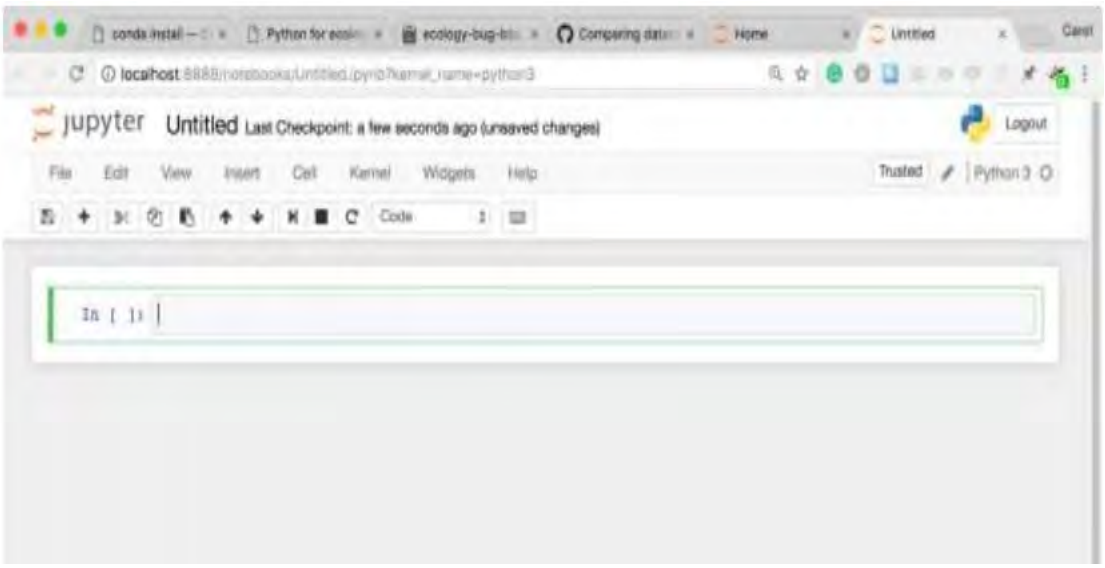

Figure 2. *The Jupyter Notebook Interface* (Data Carpentry, 2024)

Within the scope of this study, student academic performance will be predicted. In the prediction phase, an implementation will be carried out within Jupyter Notebook using the Python programming language and machine learning algorithms. Classification algorithms will be employed for prediction, and the following algorithms have been specified:

- Decision Trees (DT)
- Random Forest (RF)
- Naive Bayes (NB)
- K-Nearest Neighbors (KNN)
- Support Vector Machines (SVM) machine learning

The procedure to be followed using these algorithms is outlined as follows:

Data preprocessing will be performed in Jupyter Notebook using the Python programming language, along with the pandas and numpy libraries. Subsequently, models will be created for prediction using the sklearn library. The generated models will be divided into 40% for training and 60% for testing. Initially, binary classification will be conducted for each machine learning method, followed by multi-class classification for the same models, revealing the accuracy rates and prediction speeds of the models. Additionally, cross-validation will be conducted for each method and algorithm, allowing for the reevaluation of success rates.

CRISP-DM Methodology Processes

In data mining studies, the cross-industry standard process for data mining, known as the CRISP-DM cycle, is widely used in planning all steps from problem definition to reaching a solution (Şeker, 2018). The main method to be followed in this thesis will be shaped according to the CRISP-DM steps and flowchart. The CRISP-DM Steps and Flow are as follows:

- Business Understanding or Problem Definition
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

The flow steps of CRISP-DM are as depicted in Figure 3.

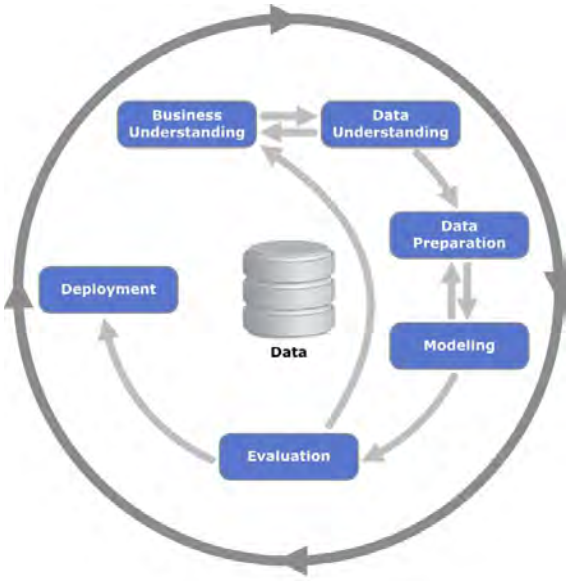

Figure 3. *The Steps and Flow of CRISP-DM* (Data Science Process Alliance, 2024)

Understanding the Business and Defining the Problem

In this stage, the definition of the problem is established. The aim of this study is to predict students' academic performance in the educational process using machine learning algorithms. Additionally, the intention is to reveal the impact levels of features, attributes, and variables affecting students' academic performance. The goal is to contribute to a faster, more efficient, and higher-quality progression of the educational process, obtain more qualified information about students, and provide meaningful, scientific, and useful feedback to students, educators, institutions, and stakeholders.

Modeling

During this stage, the process of designing and creating the actual model for practical application begins. In this step, the person responsible for the implementation analyzes and decides the proportion of data to be used for testing versus training in the model, prepares and establishes the test and training data pools for use in the model. Once the model is prepared, the application proceeds to the implementation of techniques to be used in practice. These techniques involve classifying the data allocated for training into classes, establishing relationships with training data, and applying clustering techniques. The training data and test data in the model are then comparatively tested, and the obtained results are analyzed.

Evaluation

In this stage, starting from understanding the business and defining the problem, the entire evaluation of data understanding, data preparation, and modeling stages is conducted. As described in the methodology section of this thesis, the evaluations of individual modeling for decision trees, random forest, naive Bayes, k-nearest neighbors, and support vector machines will be performed, including a comparative assessment of success rates.

Machine Learning

Due to the rapid evolution, development, and constant updates in technological revolutions in the current era, computers with high processing power have been produced. In this context, a vast amount of data has been generated, necessitating the need for processing and analysis. This emerging need has facilitated the widespread adoption of machine learning and its methods (Budak and Erpolat, 2012). Machine learning methods are developed to overcome the complexity of big data and enable meaningful analyses (Rashidi et al., 2019). Figure 4 illustrates the structure and steps of machine learning (Rashidi et al., 2019).

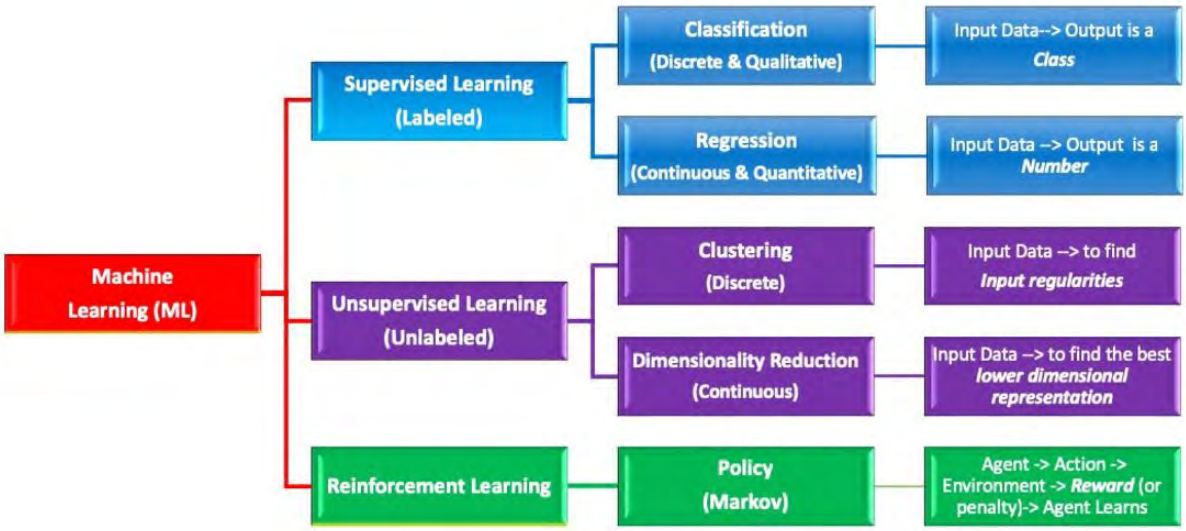

Figure 4. *The Structure and Steps of Machine Learning* (Rashidi et al., 2019)

Supervised Learning

Supervised learning, one of the stages of machine learning, is a method and stage in machine learning that aims to generate a comprehensive function based on previously known data and outcomes or observations derived from that data (Nizam and Akın, 2014).

Unsupervised Learning

The unsupervised learning method aims to facilitate the learning process of previously unobserved and unknown data during system training. In unsupervised learning, the outputs of the data are not known, making it impossible to perform classification or recognition processes. Therefore, the unsupervised learning method is employed for purposes such as identifying relationships between data features, exploring probabilities, and clustering based on co-occurrence, as the outputs of the data are not known. As evident from the definition of supervised learning, the results obtained from unsupervised learning can be utilized within the framework of supervised learning methods (Ozgur, 2004).

Reinforcement Learning

Reinforcement learning is a machine learning method fundamentally based on a trial-and-error process. It involves the development of algorithms to enable learning by interacting with the environment within which the trained system operates, aiming to find the optimal path towards a goal (Cruz et al., 2015). The structure of the reinforcement learning method is illustrated in Figure 5.

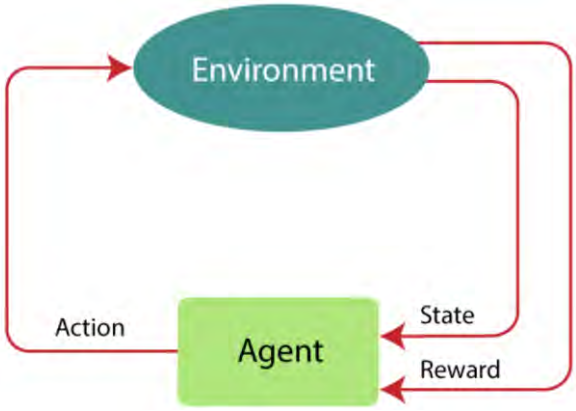

Figure 5. *Reinforcement Learning Methodology Structure* (Karagiannakos, 2018).

Machine Learning Algorithms: **Decision Tree**

A decision tree can be comprehensively utilized in both regression and classification machine learning methods in data science and mining. As the name suggests, decision trees develop suitable paths to reach predefined goals based on the anticipated objectives. Since decision trees can be used for both classification and regression, they are sometimes referred to as classification decision trees and regression decision trees (Maimon and Rokach, 2014). The decision tree algorithm consists of interconnected conditions, where branches from one condition to another correspond to all possible outcomes. For example, a decision tree diagram created at a simple level for two input conditions, X and Y, is illustrated in Figure 6.

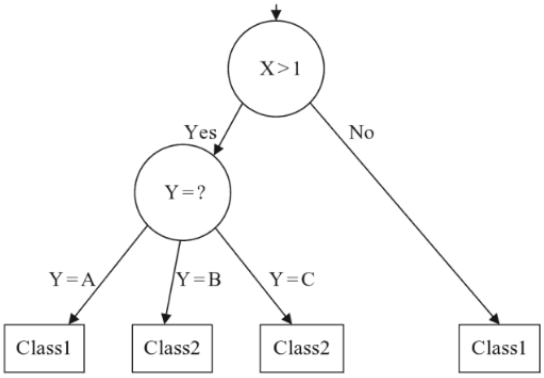

Figure 6. *Decision Tree Diagram* (Kantardzic, 2011)

Machine Learning Algorithms: **Random Forest**

Random Forest, a machine learning method/algorithm, is a decision tree algorithm, commonly known by its name "Random Forest." This algorithm creates multiple decision trees and utilizes a subset of randomly selected features from each condition of these generated decision trees as a sub-feature condition. The data within each tree is selected and used to minimize the relationship between the other trees. The outputs produced by the selected trees, determined at a specified rate, are analyzed, and classification is performed. After this classification, the most frequently occurring result is accepted as the valid outcome (Oshiro et al., 2012).

Machine Learning Algorithms: **Naive Bayes**

The Naive Bayes machine learning algorithm is commonly observed in scientific fields such as data science. It is a supervised learning method that falls under the category of machine learning. This method is used to predict the probability of a specific feature belonging to different predefined class sets. In this classification method, the Bayes decision formula is applicable (Alqaraleh, 2021). The Bayes theorem is a crucial aspect in numerical or mathematical statistical calculations. In the process of modeling any event within the scope of the Bayes theorem, it aims to obtain outputs using universally accepted assumptions by objective observers (Akar and Gündoğdu, 2014). The formula for Bayes' theorem is illustrated as shown in Figure 7.

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

Figure 7. *The formula for Bayes' theorem*

In the formula, P(A) represents the probability of the input data for the given problem scenario, P(B) denotes the probability of the output, P(A/B) signifies the probability of the occurrence of output B given input A has occurred previously, and P(B/A) expresses the probability of the occurrence of output B given that input A has happened before (Orhan and Adem, 2012).

Machine Learning Algorithms: **K-NN Algorithm**

The K-NN method, a machine learning technique, performs the learning process conditionally based on the data present in the dataset. In the learning process, a newly encountered data instance is considered

in the same category as the data instance in the dataset based on relational similarity (Mitchell, 1997). The KNN classification system relies on a logic that measures the distance between data instances from the same or different sets based on similarity conditions (Cömert, 2016). In the measurement of this distance, a pre-determined distance value is used to calculate which class the source, with an unpredictable class, is close to.

The most commonly used distance formula in calculating the similarity distance for the KNN machine learning algorithm is the Euclidean distance. This distance measurement formula is illustrated in Figure 8 (Hu et al., 2016). In the K-NN algorithm, the variable K is utilized. When a new data point is added to the system, its classification is determined by examining the distances to the nearest K neighbors. The decision is made based on which class the data point is closer to among its K nearest neighbors.

$$D(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2)}$$

Figure 8. *Euclidean Distance Formula*

Machine Learning Algorithms: **Support Vector Machines**

Support Vector Machines, belonging to the supervised learning class of machine learning, is a machine learning method with a statistical foundation. The mathematical algorithmic processes within the method have been developed for the classification of both binary and multiclass linear features, and over time, nonlinear data features have also been incorporated into the classification process. Support Vector Machines can develop a decision function to separate two classes most efficiently (Vapnik, 1999). The support vector machines algorithm has been used in the process of making predictions in various fields (Cristianini and Shawe-Taylor, 1999).

FINDINGS

Attribute Analysis of The Dataset

The dataset consists of 33 features, with 16 of them being numerical and the remaining 17 being categorical variables. The names and types of the features are presented in Table 2.

Table 2. *Attribute Names and Types of The Dataset*

Sequence No	Attribute Name	Type
1	School	Object
2	Gender	Object
3	Age	İnt64
4	Address	Object
5	Number of Family Members	Object
6	Marital Status	Object
7	Mother's Education Level	İnt64
8	Father's Education Level	İnt64
9	Mother's Occupation	Object
10	Father's Occupation	Object
11	Reason	Object
12	Parent	Objet
13	Commute Time to School	İnt64
14	Study Time	İnt64
15	Class Repetition	İnt64
16	School Aid	Object
17	Family Support	Object
18	Private Tutoring	Object
19	Participation in Social Activities	Object
20	Preschool Education	Object

21	Desire to Pursue Higher Education	Object
22	Availability of Internet at Home	Object
23	Relationship Status	Object
24	Level of Relationship with Family	İnt64
25	Non-School Leisure Time	İnt64
26	Hanging Out with Friends	İnt64
27	Alcohol Consumption during School Hours	İnt64
28	Alcohol Consumption on Weekends	İnt64
29	Current Health Status	İnt64
30	School Absenteeism	İnt64
31	Grade1	İnt64
32	Grade2	İnt64
33	Final Grade	İnt64

Binary Classification

In the data preparation phase, detailed under the Methodology section's CRISP-DM methods in the 3rd stage, we organize the feature "Final Grade" in the dataset through binary classification to prepare it for the modeling stage. The Final Grade consists of grades ranging from 1 to 20. Before modeling, the Final Grades are categorized as low for grades 1-10 and high for grades 10-20. The binary classification process is illustrated in Table 3.

Table 3. *Binary Classification Final Grade*

Binary Classification	Final Grade
High	20>=Final Grade=>11
Low	10>=Final Grade=>1

Multi-Class Classification

The first prediction model will be implemented using the binary classification method, while our second prediction model will utilize the multi-classification approach. Final grades ranging from 1 to 20 have been divided into five categories: very high, high, medium, low, and very low. The distribution of grades and corresponding levels is presented in Table 4.

Table 4. *Multiclass Classification Grade Distribution*

Final Grade	Level
0-9	Very Low
10-11	Low
12-13	Medium
14-15	Good
15-20	Very Good

Categorical Variable Transformation

There are 17 categorical features, i.e., object type, in the dataset. To apply these categorical variables to the prediction algorithm, it is necessary to convert them into numerical data types. In this context, categorical variables have been transformed into numerical data types, specifically uint8. After this type transformation, the feature count has increased from 33 to 41. This increase is due to the situation where one feature results in a sub-feature.

Results Obtained by Binary Classification

By employing binary classification, the K-NN, RF, GA, NB, and SVM machine learning algorithms provided prediction results, with the RF algorithm achieving the most successful outcome at a rate of 0.91. The RF algorithm is followed by SVM with a rate of 0.87, K-NN with a rate of 0.86, GA with a rate of 0.85, and finally, NB with a rate of 0.72. Success rates and the time elapsed in prediction are provided in Table 5 below.

Table 5. *Algorithm Success Rates and Elapsed Time*

Algorithm Name	RF	SVM	KNN	KA	NB
Success Rate	0,91	0,87	0,86	0,85	0,72
Elapsed Time (s)	0.12	0.095	0.17	0.05	0.06

Results of The Random Forest Algorithm: Analysis of Important Features with Random Forest Algorithm

The main theme of this study is the prediction of student performance, with the reference feature for measuring success being the students' final grade. The final grade is the dependent variable, meaning it is influenced by other features. The graph depicting the features from most influential to least influential on the final grade using the Random Forest algorithm is shown in Figure 9.

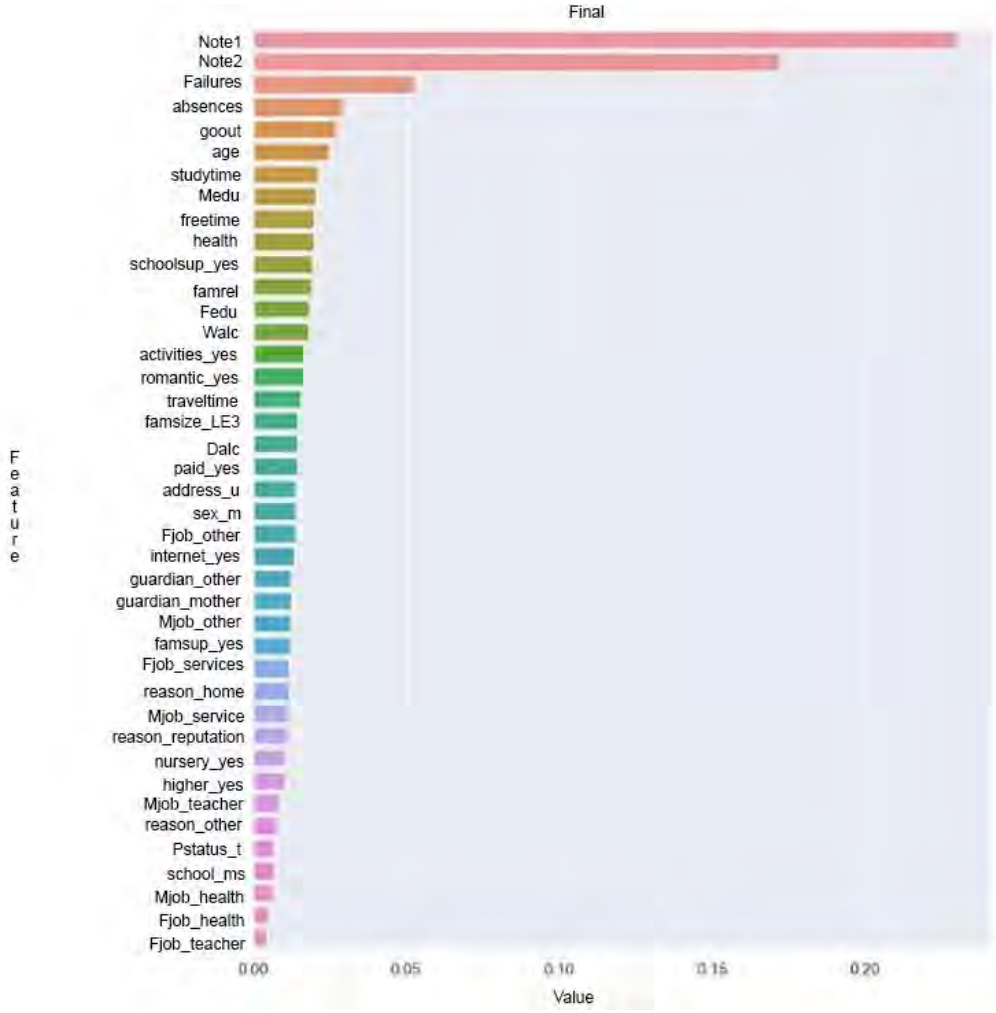

Figure 9. *Feature Ranking Impacting Final Grade*

Results of The Random Forest Algorithm: The Five Most Important Features and Coefficients with Random Forest

The final grade is a dependent variable that is influenced by all features with specific coefficients. The names of the influencing independent variables and their impact coefficients are shown in Table 6.

Table 6. *Influencing Features and Their Coefficients*

Impact Order	Feature Name	Impact Coefficient
1	Grade 2	0.24
2	Grade 1	0.18
3	School Absenteeism	0.05
4	Time Spent with Friends Outside	0.032
5	Age	0.031

Results of The Random Forest Algorithm: Cross-Validation for Random Forest Prediction

To assess the accuracy and prediction success of our model, a 5-fold cross-validation has been performed. According to the results obtained from cross-validation, a rate of 0.94 was achieved in the 5th cross-validation. Without cross-validation, the obtained result was 0.91. The cross-validation coefficients and their average are shown in Table 7.

Table 7. *Cross-Validation for Random Forest*

Order	Coefficient Value	Average Coefficient
1	0.89	
2	0.93	
3	0.88	0.91
4	0.88	
5	0.94	

The decision tree algorithms have achieved the 3rd most successful prediction outcome in the binary classification prediction model.

Table 8. *Prediction Accuracy of Decision Trees*

Decision	Decision Tree
Success Rate	0.85
Compilation Time	0.05

To assess the accuracy and prediction success of our model, a 5-fold cross-validation has been performed. According to the results obtained from cross-validation, a rate of 0.93 was achieved in the 2nd cross-validation. Without cross-validation, the obtained result was 0.91. The cross-validation coefficients and their average are shown in Table 9.

Table 9. *Cross-Validation Scores for Decision Trees*

Order	Coefficient Value	Average Coefficient
1	0.91	
2	0.86	
3	0.86	0.89
4	0.91	
5	0.91	

The data has been processed using the Naive Bayes algorithm and binary classification method. The results are presented in Table 10.

Table 10. Naive Bayes Prediction Accuracy

Decision	Decision Tree
Success Rate	0.72
Compilation Time	0.06

To assess the accuracy and prediction success of our model, a 5-fold cross-validation has been performed. According to the results obtained from cross-validation, a rate of 0.78 was achieved in the 2nd cross-validation. Without cross-validation, the obtained result was 0.72. The cross-validation coefficients and their average are shown in Table 11.

Table 11. *Cross-Validation Scores for Naive Bayes*

Order	Coefficient Value	Average Coefficient
1	0.64	
2	0.78	
3	0.63	0.69
4	0.72	
5	0.67	

The data has been processed using the KNN algorithm. The results are presented in Table 12.

Table 12. *KNN Prediction Accuracy*

Decision	Decision Tree
Success Rate	0.86
Compilation Time	0.17

To assess the accuracy and prediction success of our model, a 5-fold cross-validation has been performed. According to the results obtained from cross-validation, a rate of 0.92 was achieved in the 4th cross-validation. Without cross-validation, the obtained result was 0.86. The cross-validation coefficients and their average are shown in Table 13.

Table 13. *Cross-Validation Scores for KNN*

Order	Coefficient Value	Average Coefficient
1	0.84	
2	0.82	
3	0.87	0.86
4	0.92	
5	0.86	

The data has been processed using the SVM algorithm. The results are presented in Table 14.

Table 14. *SVM Prediction Accuracy*

Decision	Decision Tree
Success Rate	0.87
Compilation Time	0.009

To evaluate the accuracy and prediction success of our model, a 5-fold cross-validation has been conducted. According to the results obtained from cross-validation, a rate of 0.92 was achieved in the third and fourth cross-validation. Without cross-validation, the obtained result was 0.87. The cross-validation coefficients and their average are shown in Table 15.

Table 15. *Cross-Validation Scores for SVM*

Order	Coefficient Value	Average Coefficient
1	0.87	
2	0.88	
3	0.91	0.89
4	0.87	
5	0.91	

Results Obtained by Multi-Class Classification

In the first section, binary classification was performed using five different algorithms to predict success rates, and the success rates of the algorithms were compared. In this section, multiple classification was performed, and the success rates of these five algorithms were compared again. Cross-validation was conducted separately for each algorithm, and the success rates of each algorithm were evaluated. Finally, in the analysis conducted using the GridSearch library, it was concluded how much of the Final Grade data in the dataset was correctly predicted. Among the algorithms performed with multiple classification, the most successful prediction rate was achieved with the Random Forest algorithm with a 0.97 success rate. The decision tree algorithm followed with a success rate of 0.72. The ranking of other algorithms in terms of success rates is as follows: KNN with a success rate of 0.70, SVM with a success rate of 0.66, and Naive Bayes Algorithm with a prediction success rate of 0.34. The prediction success rates obtained from the algorithm are shown in Table 16.

Table 16. *Prediction Accuracy Rates and Prediction Durations*

Model	RF	KA	KNN	SVM	NB
Success Rate	0,93	0,72	0,70	0,66	0,34
Prediction Time (s)	0,14	0,07	0,007	0,09	0,04

The results of Multiple Classification with the Random Forest Algorithm are presented in Table 17.

Table 17. *Random Forest Multiclass Classification Accuracy*

Model	RF
Success Rate	0,93
Prediction Time (s)	0,14

The Cross-Validation results of Multiple Classification with the Random Forest Algorithm are presented in Table 18.

Table 18. *Results Obtained by Applying Cross Validation to Tandom Forest*

	1	2	3	4	5	Avg
Cross Validation Rates	0,97	0,92	0,94	1	0,97	0,96

The GridSearch results of Multiple Classification with the Random Forest Algorithm are presented in Table 19.

Table 19. *Random Forest Gridsearch Results*

Level	Sensitivity	Precision	Score Success
Very High	0,80	0,50	0,62
High	0,61	0,83	0,70
Medium	0,65	0,46	0,54
Low	0,67	0,74	0,70
Very Low	0,92	0,92	0,97
Average	0,76	0,75	0,75

In the framework of the decision tree algorithm, the importance order of the data in the dataset and the importance coefficients of these data within the dataset have been extracted. These features are shown in Table 20.

Table 20. *Dataset Importance Coefficients*

Order	Feature Index	Name Importance	Coefficient
1	15	Final	0.412082
2	14	Note 2	0.138535
3	13	Note 1	0.088713
4	5	Class Repetition	0.018968
5	12	School Absenteeism	0.015272
6	10	Weekend Alcohol	0.015202
7	1	Consumption	0.014803
8	8	Mother's Education Level	0.014607
9	0	Going Out with Friends Age	0.013219
10	6	Relationship Level with Family	0.012738
11	7	Non-School Leisure Time	0.012467
12	14	Study Time	0.012448
13	11	Current Health Status	0.012347
14	2	Father's Education Level	0.012088
15	9	School Time Alcohol Consumption	0.011334
16	3	School Transportation Time	0.010686
17	19	Number of Family Members	0.010256
18	36	Private Lessons	0.010056
19	17	Gender	0.009779
20	31	Reason	0.009684
21	38	Pre-school Education	0.009459
22	37	Participation in Social Activities	0.009334
23	35	Family Support	0.009295
24	22	Mother's Occupation_1	0.008544
25	23	Mother's Occupation_2	0.008314
26	34	School Assistance	0.008251
27	26	Father's Occupation_1	0.008137
28	32	Guardian	0.008129
29	18	Address	0.008060
30	29	Reason2	0.007588
31	40	Pre-school Education Existence	0.007499

32	27	Social Activities	0.007374
33	20	Family Support	0.006721
34	30	Mother's Occupation_3	0.005700
35	28	Mother's Occupation_4	0.005547
36	33	School Assistance	0.005223
37	16	Father's Occupation_2	0.005130
38	24	Guardian2	0.004888
39	21	School	0.004775
40	39	Mother's Occupation_5	0.004157
41	25	Mother's Occupation_6	0.002589

The first five features that influence the prediction result are illustrated in Figure 10 below. The first five important features, in order, are the final grade, Note 2, Note 1, class repetition, and school absenteeism. The success rates of Decision Trees in Multi-Class Classification are shown in Table 21.

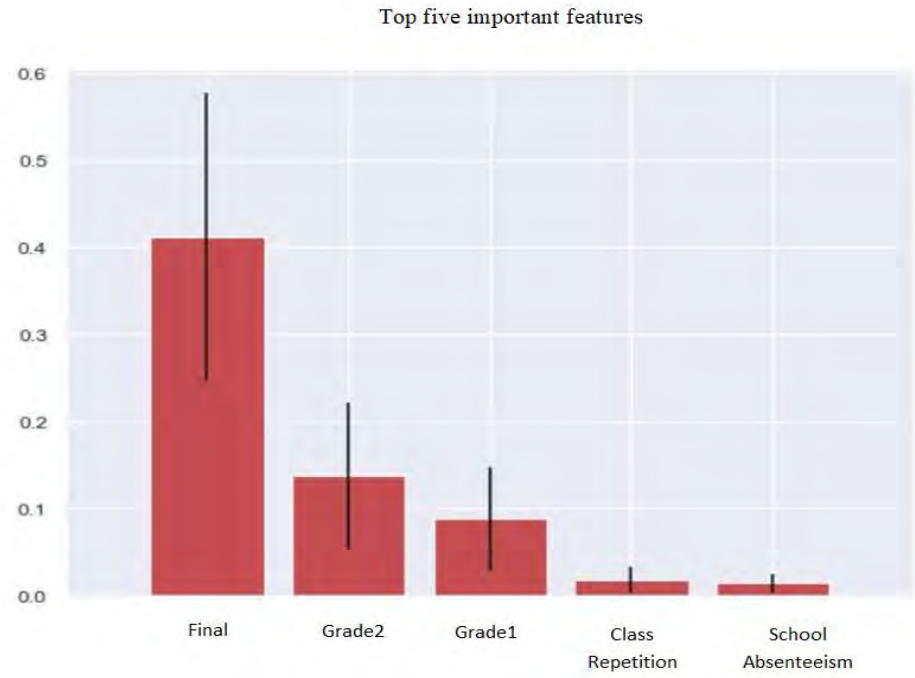

Figure 10. *Top Five Important Features*

Table 21. Decision Trees Multiclass Classification Accuracy

Model	**RF**
Success Rate	0,72
Prediction Time (s)	0,0049

Multi-Class Classification Decision Trees Cross Validation Results are presented in Table 22.

Table 22. *Results Obtained by Applying Cross Validation to KA*

	1	**2**	**3**	**4**	**5**	**Avg**
Cross Validation Rates	0,71	0,75	0,65	0,56	0,67	0,66

Multi-Classification Decision Trees GridSearch Results are given in Table 23.

Table 23. *Decision Trees Gridsearch Results*

Level	**Sensitivity**	**Precision**	**Score Success**
Very High	0,93	0,88	0,90
High	0,92	0,92	0,92
Medium	0,91	0,83	0,87
Low	0,85	0,94	0,89
Very Low	0,98	0,97	0,97
Average	0,92	0,91	0,91

The success rates of Decision Trees in Multi-Class Classification are shown in Table 24.

Table 24. *KNN Multiclass Classification Accuracy*

Model	**RF**
Success Rate	0,70
Prediction Time (s)	0,007

Multi-Classification KNN Cross Validation Results are presented in Table 25.

Table 25. KNN Cross Validation Results

	1	2	3	4	5	Avg
Cross Validation Rates	0,71	0,75	0,65	0,56	0,67	0,66

Multi-Classification SVM Algorithm Results are provided in Table 26.

Table 26. SVM Multiclass Classification Accuracy

Model	**SVM**
Success Rate	0,66
Prediction Time (s)	0,0099

Multi-Classification SVM Cross Validation Results are presented in Table 27.

Table 27. Cross Validation Results

	1	**2**	**3**	**4**	**5**	**Avg**
Cross Validation Rates	0,70	0,76	0,64	0,70	0,78	0,72

Multi-Classification NB Algorithm Results are provided in Table 28.

Table 28. *NB Multiclass Classification Accuracy*

Model	**SVM**
Success Rate	0,34
Prediction Time (s)	0,04

The results obtained by applying SVM Cross Validation are given in Table 29.

Table 29. NB Cross Validation Results

	1	**2**	**3**	**4**	**5**	**Avg**
Cross Validation Rates	0,31	0,34	0,25	0,32	0,37	0,32

DISCUSSİON, CONCLUSION AND RECOMMENDATIONS

The main aim of this study is to subject the features, attributes, educators, students, and all stakeholders involved in the educational and teaching process to analysis through modern evaluation methods introduced by contemporary technologies, in contrast to traditional assessment methods. In this context, the analysis of factors affecting students' academic achievements was conducted using educational data mining and machine learning techniques.

Within the scope of educational data mining, this study, centered around the CRISP-DM method, conducted analysis and prediction using machine learning algorithms. The algorithms used in these prediction and analysis processes are RF, KA, KNN, SVG, and NB, totaling five. For each algorithm, binary classification was performed first, dividing final grades from 1 to 20 into two categories through binary classification. The range from 1 to 10, including 10, was categorized as low grades, and the range from 11 to 20, including 20, was categorized as high grades. Among the 33 attributes in the dataset, 17 categorical attributes were transformed into numerical features by converting them to 1 and 0 using the Pandas and Numpy libraries in the Python programming language. After these transformations, the dataset was divided into 40% training data and 60% test data, and the final grade was treated as the dependent variable, establishing models separately for prediction with five different algorithms.

Out of these five algorithms, Random Forest algorithm provided the best prediction result with a success rate

of 0.91, followed by the SVM algorithm with a success rate of 0.87. The success rates of other algorithms are listed under the findings subheading in the subsection of binary classification. Within the Random Forest algorithm, the top five attributes that had the most impact on the dependent variable, the final grade, were listed in order. The five most influential features and their coefficients were, respectively, Not2 with a coefficient ratio of 0.24, Not1 with a coefficient ratio of 0.18, school absenteeism with a coefficient ratio of 0.05, the duration of going out with a friend with a coefficient ratio of 0.032, and finally, age with a coefficient ratio of 0.031.

Furthermore, the accuracies of the models established under each algorithm were re-evaluated with cross-validation and the GridSearch method. For Random Forest, the highest value obtained in five cross-validations was 94%, and the average of these five cross-validations was obtained as 0.91. After binary classification prediction processes, multiple classifications were performed for five algorithms, establishing modeling and prediction. In the multiple classification method, as in the binary classification, 40% training and 60% test data were separated. In the multiple classification method, final grades from 1 to 20 were divided into five categories: very low (0-9), low (10-11), medium (12-13), good (14-15), and very good (15-20). As a result of this multiple classification, as in binary classification, the Random Forest application achieved the best prediction success with a success rate of 0.93. The decision tree algorithm followed with a success rate of 0.72. Five cross-validations were performed for multiple classification predictions. The average of these five cross-validations was obtained as 0.96. Finally, using the GridSearch technique, sensitivity and precision values were determined. For the multiple classification method, which separated into five categories, the precision rate for the very good grade class was 0.50, the sensitivity rate was 0.80, and the total score rate was 0.62. Other category coefficients are detailed in Table 20 under the findings section.

When reviewing studies on predicting student academic performances, success rates vary between 0.85% and 95%, depending on the modeling techniques and algorithms used. The distinctiveness of this study is that five different machine learning algorithms were used for two different classification methods, predicting students' academic achievements, and predicting how much these features affect the result.

Education and teaching are crucial for all societies, and they will continue to be so. Therefore, improving and increasing the quality of education will proportionally increase the quality of the society. Other models can be added, such as gradient boosting machines (e.g. XGBoost, LightGBM), neural networks, or ensemble methods that combine multiple models for potentially better performance. Data augmentation techniques or synthetic data can be created to increase the dataset size and improve model training. Explainable AI techniques can be applied to understand and interpret the model's decisions. Tools such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Descriptions) can provide information about feature importance and model behavior. Studies can be conducted to ensure that the model generalizes well to different student populations and educational settings. Stress testing can be done by adding noise and distortions to the data to evaluate the stability and flexibility of models. The models developed can be piloted in real educational environments to collect practical feedback and evaluate their effectiveness in real-world scenarios. Educators and administrators can be worked with to generate actionable insights from model predictions and ensure they can be integrated into instructional and intervention strategies. Collaboration with other researchers and institutions can be encouraged to create larger, more diverse data sets that can provide richer information.

This thesis study was conducted to contribute to the improvement of the quality of education and teaching. Educational data mining and machine learning can be used not only to predict students' academic performances but also to predict the performance of educators. In addition, the benefit of educational data mining and machine learning methods for guidance services of schools, educational institutions, and other educational organizations in interpreting and evaluating students should not be overlooked. Thus, guidance services can provide more effective guidance for students or educators within the academic process. In the future, with more data collected, more comprehensive machine learnings can be performed.

REFERENCES

Abaidullah, A. M., Ahmed, N., & Ali, E. (2015). Identifying hidden patterns in students" feedback through Cluster Analysis. *International Journal of Computer Theory and Engineering*, *7*(1), 16. https://doi.org/10.7763/IJCTE.2015.V7.923

Abdous, M. H., Wu, H., & Yen, C. J. (2012). Using data mining for predicting relationships between online question theme and final grade. *Journal of Educational Technology & Society*, *15*(3), 77. https://www.jstor.org/stable/pdf/jeductechsoci.15.3.77.pdf

Ahmad, F., Ismail, N. H., & Aziz, A. A. (2015). The prediction of students' academic performance using classification data mining techniques. *Applied Mathematical Sciences*, *9*(129), 6415-6426. https://doi.org/10.12988/ams.2015.53289

Akar, M., & Gündoğdu, S. (2014). Bayes Teorisinin su ürünlerinde kullanım olanakları. *Journal of Fisheries Sciences.com*, *8*(1), 8-16. https://doi.org/10.3153/jfscom.2014002

Akyüz, Y. (1982). Türk eğitim tarihi: Başlangıçtan 1982'ye. *(No Title)*. https://cir.nii.ac.jp/crid/1130282271668763008

Alemdağ, C., Erman, Ö., & Yılmaz, A. K. (2014). Beden eğitimi öğretmeni adaylarının akademik motivasyon ve akademik öz-yeterlikleri. *Spor Bilimleri Dergisi*, *25*(1), 23-35. https://dergipark.org.tr/en/pub/sbd/issue/16369/171304

Alkan, C., & Kurt, M. (1998). *Özel öğretim yöntemleri.* Anı Yayıncılık.

Almarabeh, H. (2017). Analysis of students' performance by using different data mining classifiers. *International Journal of Modern Education and Computer Science, 9*(8), 9. https://doi.org/10.5815/ijmecs.2017.08.02

Alqaraleh, S. (2021). Efficient Turkish text classification approach for crisis management systems. *Gazi University Journal of Science,* 1-1. https://doi.org/10.35378/gujs.715296

Anaconda (2024, January, 5). *Anaconda navigator.* https://docs.anaconda.com/navigator/index.html

Andonie, R. (2010). Extreme data mining: Inference from small datasets. *International Journal of Computers, Communication & Control*. https://doi.org/10.15837/ijccc.2010.3.2481

Arslan, İ. (2019). *Python ile veri bilimi*. Pusula.

Ayık, Y. Z., Özdemir, A., & Yavuz, U. (2010). Lise türü ve lise mezuniyet başarısının, kazanılan fakülte ile ilişkisinin veri madenciliği tekniği ile analizi. *Atatürk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 10*(2), 441-454. https://dergipark.org.tr/en/pub/ataunisosbil/issue/2820/38029

Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, *1*(1), 3-17. https://jedm.educationaldatamining.org/index.php/JEDM/article/view/8

Baradwaj, B. K., & Pal, S. (2012). Mining educational data to analyze students' performance. *arXiv preprint arXiv:1201.3417*. https://arxiv.org/abs/1201.3417

Baran, E. (2013). Öğretim teknolojilerinde yeni eğilimler ve yaklaşımlar. K. Çağıltay & Y. Göktaş. *Öğretim teknolojilerinin temelleri: Teoriler, araştırmalar, eğilimler*bn (s. 567-581). Pegem.

Belachew, E. B., & Gobena, F. A. (2017). Student performance prediction model using machine learning approach: the case of Wolkite university. *International Journal of Advanced Research in Computer Science and Software Engineering*, *7*(2), 46-50. https://doi.org/10.23956/ijarcsse/V7I2/01219

Bhardwaj, B. K., & Pal, S. (2012). *Data Mining: A prediction for performance improvement using classification.* https://arxiv.org/abs/1201.3418

Budak, H., & Erpolat, S. (2012). Kredi riski tahmininde yapay sinir ağları ve lojistik regresyon analizi karşılaştırılması. *AJIT-e: Academic Journal of Information Technology*, *3*(9), 23-30. https://doi.org/10.5824/1309-1581.2012.4.002.x

Calders, T., & Pechenizkiy, M. (2012). Introduction to the special section on educational data mining. *Acm Sigkdd Explorations Newsletter*, *13*(2), 3-6. https://doi.org/10.1145/2207243.2207245

Can, E. (2017). *Temel eğitimden ortaöğretime geçiş sınavı kazanımlarının veri madenciliği yöntemleri ile değerlendirilmesi* [Master's thesis]. https://www.acikerisim.aku.edu.tr/xmlui/handle/11630/6273

Chen, D. Y. (2017). *Pandas for everyone: Python data analysis*. Addison-Wesley Professional. https://books.google.com.tr/books?hl=tr&lr=&id=7zhDDwAAQBAJ&oi=fnd&pg

Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance.

https://repositorium.sdum.uminho.pt/handle/1822/8024

Cömert, B. (2016). *Alın bölgesinden alınan elektrookülogram (EOG) işaretleri için ölçüm devresi tasarımı ve sınıflandırılması* [Master's thesis]. Balıkesir Üniversitesi.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press. https://doi.org/10.1017/CBO9780511801389

Cruz, F., Twiefel, J., Magg, S., Weber, C., & Wermter, S. (2015, July). Interactive reinforcement learning through speech guidance in a domestic scenario. In *2015 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE. https://doi.org/10.1109/IJCNN.2015.7280477

Data Carpentry (2024). *What is Data Carpentry? https://datacarpentry.org/*

Data Science Process Alliance (2024, January, 23). *What is CRISP DM?.* https://www.datascience-pm.com/crisp-dm-2/

Eğiticinin Eğitimi (2022, December, 2). *Öğretim Nedir?* https://www.egiticininegitimi.gen.tr/ogretim_nedir.php

Erdoğan, Ş. Z., & Timor, M. (2005). *A data mining appliaction in a student database.*

Ergün, M. (1994). *Eğitim sosyolojisi*. Ocak Yayınları, 5. https://d1wqtxts1xzle7.cloudfront.net/36219308/1987Egitim_sosyolojisi

Ghorbani, R., & Ghousi, R. (2020). Comparing different resampling methods in predicting students' performance using machine learning techniques. *IEEE Access, 8,* 67899-67911. https://doi.org/10.1109/ACCESS.2020.2986809

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357-362. https://doi.org/10.1038/s41586-020-2649-2

Hu, L. Y., Huang, M. W., Ke, S. W., & Tsai, C. F. (2016). The distance function effect on k-nearest neighbor classification for medical datasets. *SpringerPlus*, *5*(1), 1-9. https://doi.org/10.1186/s40064-016-2941-7

Hughes, D., & Pearce, D. (2003). Secondary school decile ratings and participation in tertiary education. *New Zealand Journal of Educational Studies*, 193-206.

Ibrahim, Z., & Rusli, D. (2007, September). Predicting students' academic performance: comparing artificial neural network, decision tree and linear regression. In *21st Annual SAS Malaysia Forum, 5th September*.

Kantardzic**,** M**.** (2011). Data mining: concepts, models, methods and algorithms, John Wiley & Sons https://doi.org/10.1002/9781118029145

Karabay (2024, January, 5). *Scikit-learn Nedir?* https://www.karabayyazilim.com/blog/python/scikit-learn-nedir-2020-02-12-062241

Karagiannakos, S. (2018). *The secrets behind reinforcement learning*. https://theaisummer.com/Reinforcement_learning/

Koçer, H. A. (1980). *Eğitim tarihi*. Ankara Üniversitesi Eğitim Fakültesi Yayınları, (89).

Kumar, M., Singh, A. J., & Handa, D. (2017). Literature survey on student's performance prediction in education using data mining techniques. *International Journal of Education and Management Engineering*, *7*(6), 40-49. https://doi.org/10.5815/ijeme.2017.06.05

Larose, D. T., & Larose, C. D. (2014). *Discovering knowledge in data: an introduction to data mining* (Vol. 4). John Wiley & Sons. https://doi.org/10.1002/9781118874059

Learning Forward, 2021. Culture impacts learning — and not just for students . https://learningforward.org/wp-content/uploads/2021/02

Lozano-Blasco, R.; Quílez-Robres, A.; Usán, P.; Salavera, C.; Casanovas-López, R. Types of Intelligence and Academic Performance: A Systematic Review and Meta-Analysis. *J. Intell, 10,* 123. https://doi.org/10.3390/jintelligence10040123

Maimon, O. Z., & Rokach, L. (2014). *Data mining with decision trees: Theory and applications* (Vol. 81). World Scientific. https://books.google.com.tr/books?hl=tr&lr=&id=OVYCCwAAQBAJ&oi=fnd&pg

Matplotlib (2024, January, 5). *Matplotlib: Visualization with Python.* https://matplotlib.org/

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

Medium (2024, January, 18). *Python ve Anaconda kurulumu.* https://medium.com/kodcular/python-ve-anaconda-kurulumu-b8931bd80e64

Mitchell, T. M. (1997). *Machine learning*. https://thuvienso.hoasen.edu.vn/handle/123456789/9610

Nizam, H., & Akın, S. S. (2014). Sosyal medyada makine öğrenmesi ile duygu analizinde dengeli ve dengesiz veri setlerinin performanslarının karşılaştırılması. *XIX. Türkiye'de İnternet Konferansı*, *1*(6), 873-883.

NumPy (2024, January, 10). *NumPy: the absolute basics for beginners.* https://numpy.org/doc/stable/user/absolute_beginners.html

Oloruntoba, S. A., & Akinode, J. L. (2017). Student academıc performance predıctıon usıng support vector machıne. *Internatıonal Journal of Engıneerıng Scıences & Research Technology, 6(*12), 588-598. https://d1wqtxts1xzle7.cloudfront.net/55391078/74-libre.pdf?1514458425

Orhan, U., & Adem, K. (2012). aive Bayes yönteminde olasılık çarpanlarının etkileri (The effects of probability factors in aive Bayes method). *Ionosphere*, *351*, 34. https://www.emo.org.tr/ekler/3896071e2f0ee60_ek.pdf

Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012). How many trees in a random forest?. In *Machine learning and data mining in pattern recognition: 8th International Conference, MLDM 2012, Berlin, Germany, July 13-20, 2012. Proceedings 8* (pp. 154-168). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-31537-4_13

Osmanbegovic, E., & Suljic, M. (2012). Data mining approach for predicting student performance. *Economic Review: Journal of Economics and Business*, *10*(1), 3-12. https://www.econstor.eu/handle/10419/193806

Ozgur, A. (2004). *Supervised and unsupervised machine learning techniques for text document categorization* [Master's thesis]. Boğaziçi University, İstanbul.

Özgül, F. (2013). *Python kılavuzu.* https://d1wqtxts1xzle7.cloudfront.net/32859502/Python3x-libre.pdf

Pandas (2024, January, 18). *Getting started Installation instructions.* https://pandas.pydata.org/getting_started.html

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert systems with applications*, *41*(4), 1432-1462. https://doi.org/10.1016/j.eswa.2013.08.042

Petcu, N. (2015). Data mining techniques used to analyze students' opinions about computization in the educational system. *Bulletin of the Transilvania University of Brasov. Series V: Economic Sciences*, 289-298. https://webbut.unitbv.ro/index.php/Series_V/article/view/4424

Ramesh, V. A. M. A. N. A. N., Parkavi, P., & Ramar, K. (2013). Predicting student performance: a statistical and data mining approach. *International journal of computer applications*, *63*(8). https://doi.org/10.5120/10489-5242

Rashidi, H. H., Tran, N. K., Betts, E. V., Howell, L. P., & Green, R. (2019). Artificial intelligence and machine learning in pathology: the present landscape of supervised methods. *Academic pathology*, *6*, 2374289519873088. https://doi.org/10.1177/2374289519873088

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, *33*(1), 135-146. https://doi.org/10.1016/j.eswa.2006.04.005

Salili, F., Chiu, C. Y., & Lai, S. (2001). The influence of culture and context on students' motivational orientation and performance. *Student motivation: The culture and context of learning*, 221-247.

Shulruf, B., Hattie, J., & Tumen, S. (2008). Individual and school factors affecting students' participation and success in higher education. *Higher Education*, *56*, 613-632.

Şahin, M. C., & Tekdal, M. (2005). İnternet tabanlı uzaktan eğitimin etkililiği: Bir meta-analiz çalışması. *Akademik Bilişim*, *2*(4), 1-11. https://www.researchgate.net/profile/Mehmet-Tekdal/publication/237803864

Şeker, Ş. E. (2018). CRISP-DM: Endüstriler arası standart işleme–veri madenciliği için (Cross Industry Standard Processing–Data Mining). *YBS Ansiklopedi*, *5*(2). https://ybsansiklopedi.com/wp-content/uploads/2018/08/crispdm.pdf

Şengür, D., & Tekin, A. (2013). Öğrencilerin mezuniyet notlarının veri madenciliği metotları ile tahmini. *Bilişim Teknolojileri Dergisi*, *6*(3), 7-16. https://dergipark.org.tr/en/pub/gazibtd/issue/6629/88010

Tosunoğlu, E., Yılmaz, R., Özeren, E., & Sağlam, Z. (2021). Eğitimde makine öğrenmesi: Araştırmalardaki güncel eğilimler üzerine inceleme. *Ahmet Keleşoğlu Eğitim Fakültesi Dergisi*, *3*(2), 178-199. https://dergipark.org.tr/en/pub/akef/issue/65129/969332

Tuzcu, S. (2020). Çevrimiçi kullanıcı yorumlarının duygu analizi ile sınıflandırılması. *Eskişehir Türk Dünyası Uygulama ve*

Araştırma Merkezi Bilişim Dergisi, *1*(2), 1-5. https://dergipark.org.tr/en/pub/estudambilisim/issue/53654/676052

Vapnik, V. (1999). *The nature of statistical learning theory*. Springer Science & Business Media. https://doi.org/10.1007/978-1-4757-3264-1

Yıldırım, İ. (2000). Akademik başarını yordayıcısı olarak yalnızlık sınav kaygısı ve sosyal destek. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, *18*(18). http://www.efdergi.hacettepe.edu.tr/yonetim/icerik/makaleler/1093-published.pdf

Yukselturk, E., Ozekes, S., & Türel, Y. K. (2014). Predicting dropout student: An application of data mining methods in an online education program. *European Journal of Open, Distance and e-learning*, *17*(1), 118-133. https://doi.org/10.2478/eurodl-2014-0008