PREDICTING STUDENTS' PERFORMANCE USING CLASSIFICATION ALGORITHMS AND GENERATIVE ADVERSARIAL NETWORK

2023
MASTER THESIS
COMPUTER ENGINEERING

Aws Mohammed KHUDHUR

Thesis Advisor
Assist. Prof. Dr. Nehad T.A RAMAHA

PREDICTING STUDENTS' PERFORMANCE USING CLASSIFICATION ALGORITHMS AND GENERATIVE ADVERSARIAL NETWORK

Aws Mohammed KHUDHUR

Thesis Advisor

Assist. Prof. Dr. Nehad T.A RAMAHA

T.C.

Karabuk University

Institute of Graduate Programs

Department of Computer Engineering

Prepared as

Master Thesis

KARABUK

July 2023

I certify that in my opinion the thesis submitted by Aws Mohammed KHUDHUR titled “PREDICTING STUDENTS' PERFORMANCE USING CLASSIFICATION ALGORITHMS AND GENERATIVE ADVERSARIAL NETWORK” is fully adequate in scope and in quality as a thesis for the degree of Master of Science.

Assist. Prof. Dr. Nehad T.A RAMAHA .....
Thesis Advisor, Department of Computer Engineering

This thesis is accepted by the examining committee with a unanimous vote in the Department of Computer Engineering as a Master of Science thesis. July 25, 2023

Examining Committee Members (Institutions)	Signature
Chairman : Assist. Prof. Dr. Kürşat Mustafa KARAOĞLAN (KBU)	.....
Member : Assist. Prof. Dr. Nehad T.A RAMAHA (KBU)	.....
Member : Assist. Prof. Dr. Ali HAMITOĞLU (ISU)	.....

The degree of Master of Science by the thesis submitted is approved by the Administrative Board of the Institute of Graduate Programs, Karabuk University.

Prof. Dr. Müslüm KUZU .....
Director of the Institute of Graduate Programs

"I declare that all the information within this thesis has been gathered and presented in accordance with academic regulations and ethical principles and I have according to the requirements of these regulations and principles cited all those which do not originate in this work as well."

Aws Mohammed KHUDHUR

ABSTRACT

M. Sc. Thesis

PREDICTING STUDENTS' PERFORMANCE USING CLASSIFICATION ALGORITHMS AND GENERATIVE ADVERSARIAL NETWORK

Aws Mohammed KHUDHUR

Karabuk University
Institute of Graduate Programs
The Department of Computer Engineering

Thesis Advisor:
Assist. Prof. Dr. Nehad T.A RAMAHA
July 2023, 58 pages

Predicting students' academic performance has played a vital role in measuring educational success. Academic performance prediction is useful for mitigating academic challenges like delayed graduation, dropping out of university, etc. Other than predicting students' performance helps in examining student learning behavior, improving the learning environment, addressing student problems, and enabling data-driven decision-making. To identify the strengths and limitations of current performance prediction techniques, this research critically analyzed the latest student performance prediction models. Predicting student performance is an important area of education research. Machine learning (ML) techniques are often used to improve the accuracy and reliability of predicting student performance. In this study, we propose a new approach to predict student performance using five machine learning techniques, including data analysis, preprocessing techniques, and data augmentation using GANs. We realistically evaluate the proposed method.

A dataset of students' academic performance and compare the results with those obtained without data enrichment. Our results show that data augmentation significantly improves the accuracy and reliability of predicting student performance. Algorithm results (RF 99.63%, DT 99.63%, ANN 99.52%, Liner SVM 99.65%, RBF SVM 99.68%). This research contributes to the field of education by providing more comprehensive and accurate predictive models of student performance that can support informed decision-making and improve educational outcomes.

Key Words : Machine Learning, GAN, classification, Student’s performance.
Science Code : 92431

ÖZET

Yüksek Lisans Tezi

SINIFLANDIRMA ALGORİTMALARI VE ÜRETKEN ÇATIŞMA AĞLARI KULLANARAK ÖĞRENCİ PERFORMANSINI TAHMİN ETMEK

Aws Mohammed KHUDHUR

Karabük Üniversitesi
Lisansüstü Eğitim Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı:
Dr. Öğr. Üyesi Nehad T.A RAMAHA
Temmuz 2023, 58 sayfa

Öğrencilerin akademik performansını tahmin etmek, eğitim başarısını ölçmede hayati bir rol oynamıştır. Eğitilmiş performans tahmini, mezuniyetin gecikmesi, üniversiteden ayrılma gibi akademik zorlukların hafifletilmesine yardımcı olur. Öğrencilerin performansını tahmin etmenin dışında, öğrencinin öğrenme davranışını incelemeye, öğrenme ortamını iyileştirmeye, öğrenci sorunlarını ele almaya ve veriye dayalı karar vermeyi etkinleştirmeye yardımcı olur. Mevcut performans tahmin tekniklerinin güçlü yanlarını ve sınırlamalarını belirlemek için bu araştırma, en son öğrenci performans tahmin modellerini eleştirel bir şekilde analiz etti. Öğrenci performansını tahmin etmek, eğitim araştırmasının önemli bir alanıdır. Öğrenci performansı tahmininin doğruluğunu ve güvenilirliğini artırmak için makine öğrenimi (ML) teknikleri yaygın olarak kullanılmaktadır. Bu çalışma, veri analizi, ön işleme teknikleri ve GAN veri artırma dahil olmak üzere beş makine öğrenimi tekniği kullanarak öğrenci performansı tahminine yönelik yeni bir yaklaşım

önermektedir. Önerilen yaklaşımı, öğrencilerin akademik kayıtlarından oluşan gerçekçi bir veri seti kullanarak değerlendirdik ve sonuçları veri artırmadan elde edilen sonuçlarla karşılaştırdık. Bulgularımız, veri artırmanın öğrenci performansı tahmininin doğruluğunu ve güvenilirliğini önemli ölçüde geliştirdiğini göstermektedir. Algoritmaların sonuçlarıydı (RF %99,63, DT %99,63, KNN %99,52, Liner SVM %99,65, RBF SVM %99,68). Bu araştırma, bilinçli karar vermeyi destekleyebilen ve eğitim sonuçlarını iyileştirebilen, öğrenci performansını tahmin etmek için daha kapsamlı ve doğru bir model sağlayarak eğitim alanına katkıda bulunur.

Anahtar Kelimeler **:** Öğrenci performansı, sınıflandırma, Makine Öğrenimi, GAN.
Bilim Kodu **:** 92431

ACKNOWLEDGMENT

First and most I would like to show my thanks and love to Allah almighty for giving me the power and courage to get to this stage of my life. Second, I want to extend my gratitude and appreciation to my thesis advisor, Assist. Prof. Dr. Nehad T.A RAMAHA for his guidance, support, and help throughout the writing of my thesis I will always be in his debt for the rest of my life.

My thanks and love are to my family who never let me down, and who have always been there to show their support and love.

Finally, I would like to thank my friends and colleagues for their help and support and I wish them all the best.

CONTENTS

Page

Page

Page

LIST OF FIGURES

Page

Page

LIST OF TABLES

Page

ABBREVITIONS

AI	: Artificial Intelligence
ML	: Machine learning
DT	: Decision Tree
RF	: Random Forest
KNN	: K- Nearest Neighbor
LR	: Logistic Regression
GAN	: Generative Adversarial Network
SVM	: Support Vector Machine
OOB	: Out OF – Bag
MLPs	: Multi-Layer Perceptron
RBF	: Radial Bases Function
CGAN	: Conditional Generative Adversarial Network
ROC	: Receiver Operating Characteristics
NB	: Naïve Bayes
GBM	: Gradient Boosting Machine
ANN	: Artificial Neural Network

PART 1

INTRODUCTION

1.1. OVERVIEW

Recent advances in multiple disciplines have resulted in a large amount of data being gathered [1]. Processing large amounts of data and then extracting important information is a slow process for humans. As a result, data mining techniques can be employed to harvest accurate and significant information from a large volume of data [2]. It is common knowledge that academic organizations have a highly competitive and evolving environment [3,4]. The greatest obstacle to academic institutions is understanding their full capabilities, recognizing their singularity, and developing strategies for the future [5].

Educational organizations are now recognized that machine learning has a significant impact on their performance. Machine learning (ML) computers are employed to analyze educational information and participate in educational decision-making [6]. Highly accurate prediction of student performance is advantageous because it helps in the early identification of low-achieving learners [7–8]. Educational institutions help expand the educational performance of their students by analyzing education-related data [9-10]. Predicting students' academic performance is critical for their educational career, but it is also challenging because performance depends on various dynamic factors [11-12].

This education is more commonly accessible than ever before. However, the performance of students is markedly different in every educational foundation. [13]. Student performance is paramount in the school to all parties, including the direct participants, such as students, teachers, and educational institutions, as well as the indirect participants, such as student caretakers, sponsors, the Department of

Education, and the entire state. All of these events facilitate the students to pass at school. The greatest difficulties in higher education are poor student performance and low student retention [14]. The issue can be overcome by predicting student success and taking immediate action [15]. Teachers can predict which students will receive honors visually. However, this method is primarily based on the teacher's perception of the student's previous accomplishments and performance during the semester. It is typically only possible after the teacher has become familiar with the student. [16, 17].

In this research, we investigate the effectiveness of machine learning and generative adversarial networks (GANs) in predicting student success. Traditional approaches to evaluating student performance, such as attendance records,provide information about students' abilities and areas for improvement but are not foolhardy. These approaches may lack the precision and bias that would adversely affect the accuracy of student assessments. There must be more than these traditional methods and techniques to understand student behavior.

In this study, we propose five different algorithms of machine learning that are intended to assess student performance using a dataset of student performance before and after data augmentation using a GAN.

1.2. PROBLEM STATEMENT

This research aims to predict a student's academic performance and explore the causes of a student's failure by analyzing educational data. Various frameworks for predicting academic performance have been proposed. However, more features and training samples are needed to achieve high accuracy [18]. Collecting large data samples for training is difficult and time-consuming [19]. The problem we aim to address is the challenge of accurately predicting students' academic performance. Traditionally, educators have relied on a student's historical academic data and other demographic factors to make predictions about their future performance. However, these predictions may not always be accurate, and there is a risk of overlooking students' failure.

Furthermore, there is a need to enhance the accuracy of the model predictive by incorporating data from multiple sources. This requires overcoming the challenge of dealing with missing or incomplete data, as well as the challenge of ensuring that the model is robust and reliable. To address these challenges, this research proposes utilizing a Generative Adversarial Network (GAN) to create synthetic data that can be incorporated into the existing data and enhance the accuracy of the predicted mode.

1.3. RESEARCH OBJECTİVE

The main goal of this research is to provide a model that teachers can use to identify students at risk of poor academic performance and take proactive action to help those students. This work ultimately leads to an improvement in educational outcomes.
To achieve this goal, we explore the following objectives.

- To predict students' performance. This objective was achieved by implementing five machine learning models based on generative adversarial network algorithms.
- To evaluate the performance of predictive models. This research used performance measures such as (accuracy, recall, and F1 score).

1.4. CONTRIBUTION

The main contributions of the search are set as follows.

- Developing a generative adversarial network model to increase and balance data which will enhance the accuracy of predictive model.
- Proposing a new model using machine learning algorithms based on generative adversarial network to predict students' performance.

1.5. ORGNAZATION OF THESIS

Six chapters make up the thesis work. The first chapter describes the background and research problem and purpose of the study. The Literature Review presented in Chapter2 is the subject of Chapter2, the most recent and pertinent research is also incorporated. The ML methods are described in Chapter 3 and discussed in great detail the various types of machine learning and algorithms that were employed. The approach is discussed in Chapter4. Chapter5 contains discussion and results of our study. chapter6 concluded the study with a summary of the findings and conclusion.

PART 2

LITERATURE REVIEW

Several researchers have observed the benefits and increasing popularity of machine learning (ML) as a form of predictions in the academic field. The main goal of assisting in the development of more accurate and dependable methods of predicting students' academic success, this part attempts to provide a comprehensive description of the most recent ML methods.

Chen, S et al showed that machine learning has the capacity to forecast academic achievement in Pennsylvania. This study proved that ML models are reliable and practical. The authors revealed that people with high education levels in tiny rural counties had lower crime rates and might have a greater impact on the academic performance of Pennsylvania schools using ML models as the foundation. The accuracy of the support vector machine, decision tree, random forest, logistic regression, and neural network was 48%, 54%, 50%, 51, and 60%, respectively [20].

Pallathadka, H., et al In this study, they relied on predicting students' performance in some training courses using their previous success in similar courses. Some previously unknown patterns that can be useful in analytics and predictions have been found through data mining. A dataset from the UCI library was used to test a number of machine learning algorithms, and the results (Nave Bayes.85%, ID3.70%, C4.5 80%, and SVM 96%) were calculated based on performance metrics. Including accuracy and error rate [21].

Xu, K, et al. Focused on physical fitness factors to predict ' students' performance using machine learning algorithms, such as muscle strength, aerobic endurance, and body mass of primary school pupils. This study proved the existence of a relationship between student's academic success and physical fitness. The prediction results of the

algorithms were RF 66.67, SVM 62.3, and KNN 68.85. This study can be used as the foundation for an approach that uses physical activity to improve academic achievement and physical fitness [22].

Holicza, B et al. Used a variety of machine learning algorithms, including support vector machines with different cores, decision trees, random forests, and k-nearest neighbor algorithms, to predict student performance depending on several elements, including sleep, study time, and screen time, affect academic success. The algorithms relied on two sets of databases. one contains data for online learning, and the other contains data for associated offline learning features. The algorithm with the best accuracy rate, 98%, is the kernel support vector algorithm. [23].

Sarwat, S et al. Suggested method would improve accuracy of predictions by creating new data samples using an augmented Conditional Generative Adversarial Network (CGAN). The investigation utilized multiple traits associated with school, home, and private instruction that were intended to predict student success. They also proposed an enhanced version of Deep Support Vector Machine (SVM) that was based on augmented CGAN. The performance of the Deep SVM was increased through the use of multiple cores. experimental investigations were conducted from different perspectives, and the results demonstrated that multi-core learning was more successful than single-core learning. The Deep SVM had the greatest success with few layers. The analysis of performance using ML demonstrated that SVM combined with learning methods had a higher accuracy of 97.2%, a higher specificity of 97.1%, and a higher AUC of 96.2%, [24].

Nabil et al. Collected a dataset consisting of 4,266 anonymized student records, with 12 academic performance-related attributes during their first two years of college. These attributes solely relate to academic characteristics and refer to the student's freshman grades. The collected academic features are typically utilized by universities to create predictive models that reduce error rates. Typically, deep neural networks are used to assess students' academic abilities. However, in this study, the researchers utilized students' previous course grades to predict their future course performance. Researchers utilized ML algorithms like Logistic Regression (LR),

(RF), (SVC), decision trees (DT), K-Nearest Neighbors (KNN), and Gradient Boosting Machine (GBM). The issue of having uneven data is resolved using dawn sampling methods. The study achieved accuracy 89%. [25].

Gajwani, J et al. Introduced a new approach to measuring student's academic success by combining behavior and academic metrics. To achieve this, the proposed method utilizes methods of attribute selection and various machine learning algorithms that are supervised, including decision trees, (LR), (NB), and cluster technologies such as reinforcement, packaging, voting, and random forests. Researchers employed several visualizations to identify the most significant attributes. They found that machine learning models performed the most effectively, with 75% accuracy on their dataset. The method of utilizing this approach is not limited to academic evaluation alone. It has the potential to aid students in improving their academic performance [26].

Ghorbani, R et al. Explored the practice of using resampling methods to address uneven data in the prediction of student performance using machine learning (ML). The survey studies the influence of class and attribute structure on the quality of predictions, utilizing methods like random stratifying and shuffling repeated. The findings demonstrate classifiers of the ML type have a greater advantage with a smaller number of classes and less complex nominal attributes, and that balancing data can lead to increased performance. The investigation employs the Friedman test to assess the efficacy of different sampling methods, it finds that SVM-SMOTE is the most influential approach. They also conclude that combining SVM-SMOTE with the RF classifier leads to the greatest success, with an accuracy of 76.83% [27].

Waheed, H et al. Proposed a new method of recognizing students at risk of distraction and providing early intervention in virtual learning settings. using deep artificial neural networks. They instruct the network in the extraction of hand-crafted features from clickstream data, this increases the performance of the network relative to standard (SVM), (LR). The accuracy of classification is between 79.82% and 85.60% for LR, and between 79.95% and 89.14% for SVM. The investigation concerns the value of including information regarding the assessment process and historical information regarding the effectiveness of the model. Additionally, results

demonstrate that students who have access to the material early on are more likely to have a higher performance. The ultimate goal of the study is to assist institutions in developing a pedagogical framework and guiding higher education decision-making toward sustainable education [28].

Hashim, S et al. Conducted a research study that contrasted the capabilities of virous supervised machine learning algorithms, including DT, NB, LR, SVM, KNN, Minimalization, Neural Network, in predicting the students' final exam grades. They attained datasets from the bachelor's degree programs from the University of Basra department of Computer Science. Research revealed that the LR classifier had the greatest capacity to accurately predict the final grades of students, having an accuracy of 68.7% for students that successfully completed the test and 88.8% for students that failed. These findings can facilitate educational institutions in predicting student performance and providing effective treatments to enhance their academic success [29].

Chui, K et al. In this study, the researchers proposed a new algorithm for creating a vector machine that would support the prediction of student performance in both homes and schools. Because of the limited size of the academic dataset, the proposed method benefits from the combination of two techniques. Incremental Conditional Generative Adversarial Networks (ICGAN) and Deep Support Vector Machines (DSVM). ICGAN increases the amount of data, while DSVM increases the fidelity of predictions made by deep learning-based technology. The results demonstrated that the proposed ICGAN-DSVM approach had a high degree of performance, with a specific, sensitive, and AUC of 0.968, 0.971, and 0.954, respectively. The findings also showed including the two school and the home education in the model increases the effectiveness of the model over the sole influence of school or home attributes of instruction. [30].

Altabrawee, H, et al. Presented a research study on enhancing the learning abilities of students. in Al-Muthanna University's computer science program by using (ML) algorithms for predicting student performance. The researchers employed four different methods of ML, including the (ANN), Naive Bayes Decision Tree, and LR,

for predicting students' performance based on their utilization of the internet and the amount of time spent on social media. Performance of each ML method was evaluated by the classification accuracy and (ROC) index, the ANN model had best accuracy of 77.04%. The DT algorithm can also identify the factors that contribute most to student success, this information can be utilized to enhance the educational experience [31].

Al-Shehri, H et al. Employed a dataset from the University of Minho in Portugal. For predicting the students' performance on final exams, the dataset contains 395 examples of math subject performance. While previous work on the dataset mostly relied on the KNN technique, which resulted in poor accuracy, the researchers decided to also employ the SVM algorithm, which is known for its powerful prediction abilities. The two methods were evaluated using the dataset to determine which one provided more accurate results. The results showed that there is little difference between the two algorithms SVM and KNN with a correlation coefficient of 0.96 for SVM and 0.95 for KNN. These findings indicate that using more accurate models, such as the SVM algorithm, can lead to more accurate predictions of student performance, which can then help to improve academic results [32].
For readers' reference, we also provide a brief description of the relevant studies introduced in this section.

Table 2.1. The latest research on predicting student achievement.

Reference	Year	Method	Description	Limitations	Accuracy
Chen, S et al [20]	2023	DT, LR, RF, SVM	Focused on the factors of educational levels in rural areas and crime rates to predict academic performance	Low accuracy	48%, 54%, 50%, 60%
Pallathadka, H., et al[21]	2023	Nave Bayes. ID3. C4.5 and SVM	This study forecasts students' achievement in a course based on their prior performance in related courses.	Limited data sample and no feature selection	85%, .70%, 80%, 96%
Xu, K., et al [22]	2023	RF, SVM, KNN	study mainly focuses on physical fitness factors such as muscle strength, aerobic endurance, and body mass of primary school students for prediction	poor accuracy and limited comparing methods	66.67%, 62.3%, 68.85%.
Holicza, B et al [23]	2023	DT, RF, KNN	relied on certain factors to predict student performance such sleep, study screen, time.	limited sampling and features	98%

Sarwat, S et al. [24]	2022	CGAN, DSVM	The study suggested using an augmented CGAN to create synthetic data samples,	focus on home and school tutoring only to predict students' performance	98 %
Nabil et al [25]	2021	RF, DT, KNN, SVM	The researchers used past course grades of students as input to predict their future academic performance	. Only academic feature used for predictions	89 %
Gajwani, J et al [26]	2021	Naive Bayes, DT, LR	A new method has been proposed that integrates academic and behavioral factors to assess a student's academic achievement	. Using only behavioral data to predict ability to pass the exam	75 %
Ghorbani, R et al [27]	2020	SVM-SMOTE, RF	the utilization of resampling methods to address uneven data in predicting student performance	Result needs to improve	76.83 %
Waheed, H et al [28]	2020	SVM, LR	The research highlights the significance of incorporating assessment-related and legacy data to enhance the accuracy of the model.	limited comparing methods	89.14 %
Hashim, S et al [29]	2020	DT, NB, LR, SVM, KNN	contrasted the capabilities of various supervised machine learning algorithms to predict the students' final exam grades.	low data sample	88.8 %
Chui, K et al [30]	2020	ICGAN, DSVM	Researchers developed an enhanced support vector machine algorithm to predict student performance in home and school contexts	Limited methods for comparison	96 %
Altabrawee, H,et al [31]	2019	ANN, NB, DT, LR	The researchers relied on two important factors to predict student performance. a student's Internet use and the amount of time spent on the social media	They relied on only two factors for prediction, as well as results need improvement	77.04%
Al-Shehri, H et al [32]	2017	KNN, SVM	This study developed two models for student performance prediction on the same datasets and the results were contrasted.	Limited comparison methods were applied	96 %

PART 3

THEORITICAL BACKGROUND

Machine learning is considered an important branch of AI that involves developing computers that can learn on their own. Instead of programming every rule for decision-making or pattern extraction, machine-learning techniques rely on training on huge data sets to understand concepts and structures. This means that the algorithm can be learned independently [33]. Machine learning is characterized by computers being able to make accurate predictions based on past experience.

ML offers numerous benefits, such as identifying relationships within large datasets, easily processing image-based data, aiding experts in complex decision-making, and rapidly processing vast amounts of data that would be unattainable for humans within a short time frame [34].

3.1. TYPES OF MACHINE LEARNING ALGORITHMS

In this section, we'll discuss four varieties of learning. supervised, unsupervised, Semi- Supervised, and Reinforcement, as shown in Figure 3.1. Subsequently, we will highlight the focus of this research on the implementation of five supervised learning algorithms aimed at multi-class classification.

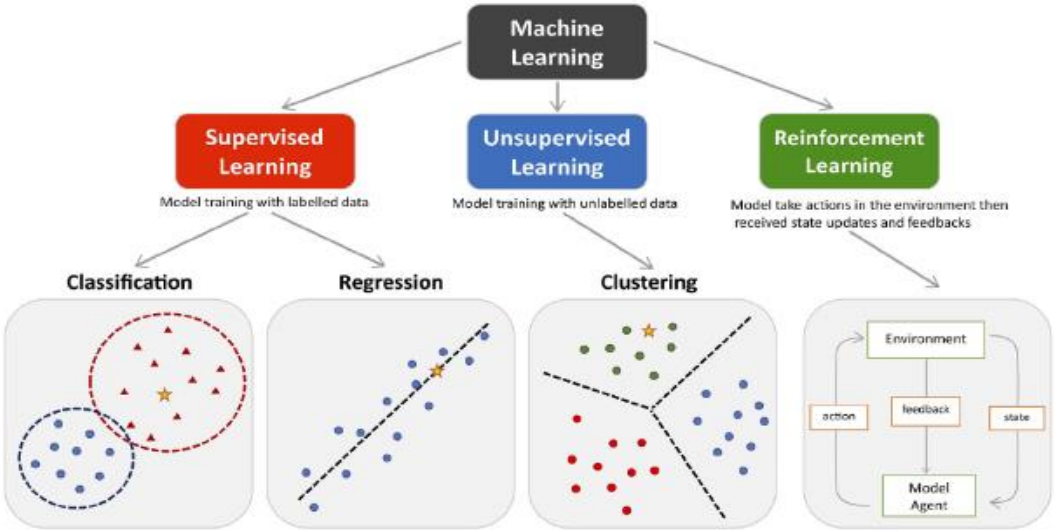

Figure 3.1. Three main types of ML techniques [35].

3.1.1. Supervised Learning

It is an initial form of learning that involves training an algorithm on a dataset with labeled features and associated outputs. The algorithm analyzes the labeled data to identify patterns and connections between inputs and outputs. Once a trained algorithm has been employed, it can utilize the knowledge it has gained to preface outcomes on new, untrained data [36]. Supervised learning is typically employed to address two main issues. classification problems, and regression problems.

3.1.2. Unsupervised Learning

The second type involves training algorithms on an unlabeled dataset without any output data (label). The algorithm examines this dataset to identify connections and patterns within the data. Following this, the data is grouped according to these relationships, allowing the algorithm to make predictions on new data [33].

Apart from supervised and unsupervised learning, another form of machine learning is called semi supervised, this is particularly useful for learning complex patterns [36]. This method combines elements of both guided and random learning processes. Employs datasets that are labeled only partially and larger unlabeled datasets to train computers.

3.1.3. Reinforcement Learning

Unlike the previous two learning types, Reinforcement learning involves an active participation from the agent in order to achieve a particular objective. This type of learning is characterized by exploration in order to reach a goal. If the agent chooses something environmental, it will be compensated based on the decision it makes. A positive reward is given if the decision brings the agent closer to its goal, and a negative reward otherwise [36].

3.2. CLASSIFICATION TECHNIQUES

Classification is a crucial and widely used technique in supervised machine learning. Basically, it can be classified into two categories. The first type, known as binary classification, deals with classifying data into two distinct groups. For example, predicting whether a student is good or not would result in two possible outcomes. good or poor. The second type, called multi-class classification, involves predicting more than two outcomes. In this case, the possible predictions could include poor, failed, and good. This versatile method allows for a more nuanced understanding of the data and enables more detailed predictions across various scenarios.

3.2.1. Decision Tree (DT) Approach

Decision trees are significant for machine learning that is supervised in nature and used to address classification and regression issues [37]. It's designed to repeatedly partition a dataset based on specific attributes, as demonstrated in Figure 3.2. The tree's components are primarily. decision nodes that assess individual attributes, and leaves that show the assessment results., with the highest decision node called the "root" [38, 39]. Several different decision tree approaches are available, including ID3, C4.5, C5, and CART which use unique mathematical methods to separate the training data. These algorithms cater to different scenarios and provide flexibility in solving various real-world problems, making decision trees a versatile and powerful tool in this field.

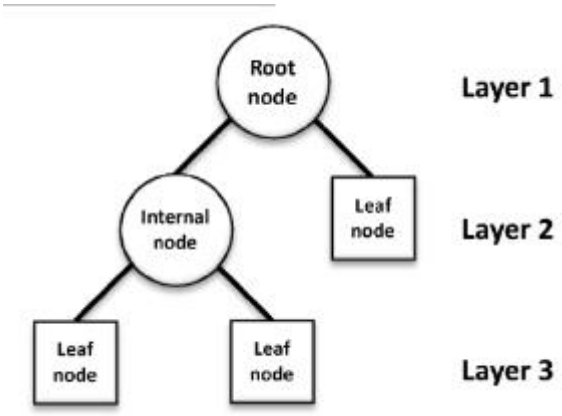

Figure 3.2. Decision Tree [40].

In this investigation, the utilized algorithm is a more sophisticated version of the CART algorithm [41], implemented using the scikit-Learn library. The CART algorithm tackles both classification and regression issues, with the nature of the dependent variable determining the tree type created by the algorithm. A classification tree is generated for categorical variables, while a regression tree is constructed for numerical variables. In our proposed research, the target variable is the student's performance (poor, fail, or good), which is categorical [42].

The CART algorithm uses Gini pollution (symbol G) to estimate the probability of mislabeling in a subset. This process takes into account the distribution of labels and is random. This approach ensures that the resulting decision tree effectively separates the data into meaningful categories, leading to more accurate predictions and improved model performance. can be calculated using the following equation.

$$-\sum\_(j = 1)\text{^}a((Pi))\text{^}2 \tag{3.1}$$

In this context, the term G is associated with the Gini impurity measure. The variable j has a range of values, which represents different classes. Pi is the percentage section that is assigned to j.

3.2.1.1. DT Approach Benefits

- Clear and comprehensible structure.
- Minimal data preparation is required for the DT algorithm.
- Low cost for tree construction.
- Applicable to both numerical and categorical data.
- They can be used to classify instances into multiple classes or make binary decisions.
- The decision tree approach is considered a white-box technique. This means that the operations and decisions made by the algorithm can be easily interpreted and understood by humans.
- Decision tree effectiveness can be evaluated using evaluation matrix.

3.2.1.2. Limitations Of the DT Approach

- Decision tree approaches are prone to overfitting. Solutions include reducing the number of branches, stating the small number of samples in a leaf node.
- Unbalanced datasets can cause biased tree generation.

3.2.2. Support Vector Machine (SVM) Approach

This technique, which was developed in the middle of the 1990 by Vapnik and colleagues [43], Despite SVM's capacity to serve both classification and predictions, it's primarily utilized for classification because of its superior performance as a machine learning method. [44]. The algorithm separates data either linearly or non-linearly by defining a hyperplane in an N-dimensional vector space to differentiate between two classes [45], as depicted in Figure 3.3. The hyperplane (H) is determined by this equation.

$$H). * Xi + b = 0 \quad (3.2)$$

W is a set of weights, where (H) denotes the hyperplane, xi denotes the input example features, and b is the bias. The kernel function is a key parameter for the success of the classifier, and it consists of three primary types.

- Linear kernel. linear kernel function uses the dot product of two vectors, as in this equation.

$$K(A, B) = \text{sum}(A * B) \quad (3.3)$$

- Polynomial Kernel. This kernel is employed to differentiate between complex and curvilinear input spaces. it can be calculated equation.

$$K(A, B) = 1 + \text{sum}(A * B) \text{ ^X} \quad (3.4)$$

- Radial basis function (RBF). The RBF used in classification, and it maps the input space to an infinite-dimensional space. Also, can be calculated by.

$$K(A, B) = \exp(-B||A - B||\text{^2}) \quad (3.5)$$

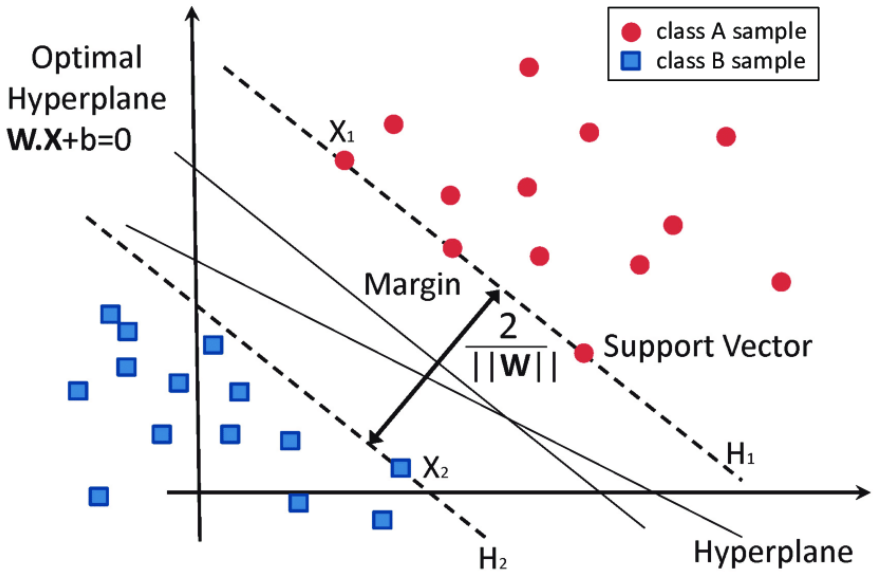

Figure 3.3. Data classification using SVM [46].

3.2.2.1. Advantages Of the SVM Approach.

- High performance effectiveness.
- Practical even if the number of dimensions exceeds the sample size.
- Memory consumption is very economical due to the subset of training points.
- Variability in the way various functions of the decision process is defined.

3.2.2.2. Limitations Of the SVM Approach.

- The choice of a kernel function and a regularization term that will avoid overfitting becomes important, during the number of features is significantly greater than the number of samples.
- SVM can be very time consuming due to the calculation of probability estimates using five-fold cross-validation.

3.2.3. Random Forest (RF) Approach

This algorithm is a supervised learning, proposed by Leo (2001), uses multiple decision trees instead of one as the ensemble learning method, resulting in a more powerful classifier [37, 47]. As shown in Figure 3.4, Random Forest generates a set of classifiers and combines the results to classify new instances [48]. CART is a decision tree that is built by successively partitioning the data into nodes, starting from the root node, this covers the entire training set [49].

An important aspect of Random Forest is the out-of-bag (OOB) error. For certain trees, each data point acts as an OOB observation. Consequently, these trees can consider these observations as internal validation data, as they were not used in building the trees. The OOB error of Random Forest is the average error rate achieved when predicting the observations of the dataset using the OOB trees. This internal validation process results in a more conservative error estimation, which is typically believed to be a trustworthy indicator of predictive errors for unseen data [50]. The following steps outline the creation of an RF.

- Figure out how many trees are there.
- Calculate number of trees (T) using the following criteria.
- Bootstrap size n and draw random samples.
- A tree's maximum potential is reached without additional maintenance.
- The majority vote of the tree determines the output of the classifier.

3.2.3.1. Benefits Of the RF Method

- Ability to handle large datasets with high dimensions.
- provides a successful way to balance errors in datasets with imbalanced classes.
- Works with both kinds of features numeric and categorical.
- The value of all of the categorical traits can be determined.

3.2.3.2. Limitations Of the RF Method

- The complexity of RF is increased by combining multiple decision trees.
- When the number of decision trees in the forest increases the expense increased as.
- Training takes longer due to the complexity of the data.

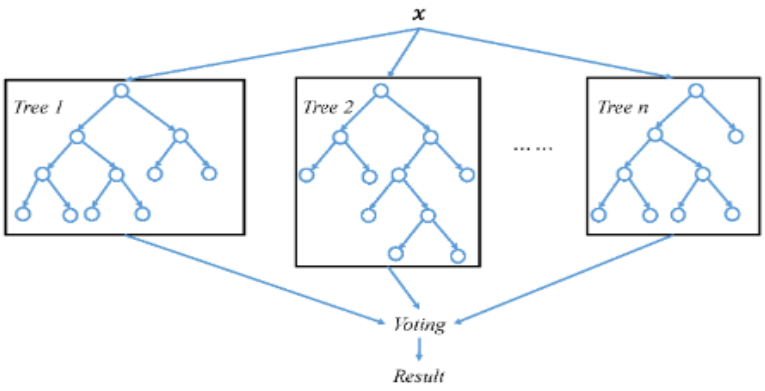

Figure 3.4. The structure of the RF technique [48].

3.2.4. K-Nearest Neighbor (KNN) Algorithm

This method is founded by Evelyn Fix and Joseph Hodges in 1951, and then enhanced by Thomas Cover [51]. Known for its simplicity, it's typically employed in conjunction with other methods for solving classification and regression issues. The algorithm receives as input the k closest training examples in the dataset, and the outcome is based on whether or not the ANN is utilized for classification. (yielding class membership) or regression (yielding attribute values for objects).

KNN operates in two primary steps.

- Locate the K samples that are most similar to the unknown object.
- Select the most popular categorization of these K instances.

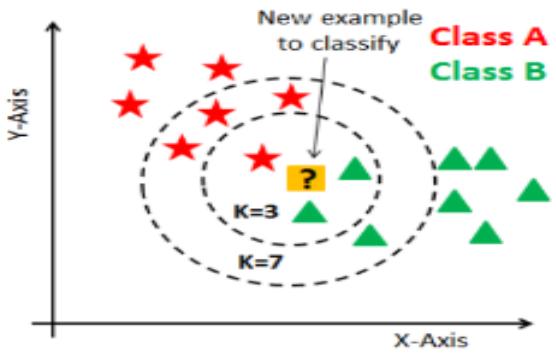

Figure 3.5. KNN algorithm [52]

In this study, the target outcomes are categorical, leading to the use of the KNN algorithm as a classification method. The KNN classifier is considered a non-Parametric Statistics because it lacks a distribution assumption about the variables involved. As illustrated in Figure 3.5, the new object (target) is categorized by the majority of its constituents.

3.2.4.1. Advantages Of the KNN Technique

- Easy to understand and interpret.

- Versatile, suitable for both regression problem and classification tasks.
- Offers high accuracy.
- Very good with data that is not linear, as it does not require additional assumptions about data, fine-tuning of multiple parameters, or creation of a model.
- Fast implementation.

3.2.4.2. KNN Disadvantages

- May be time-consuming for large datasets.
- Sensitivity to data volume and irrelevant features.
- Requires significant memory due to the need to store all training data.
- Computationally expensive, as it retains all training instances.

PART 4

METHODOLOGY

The model being proposed, illustrated in Figure 4.1, comprises four distinct steps. Initially, the dataset employed in this research is outlined. Following that, the data undergoes preparation through pre-processing during the second stage. In the third stage, the pre-processed data is enhanced with the implementation of the suggested GAN. Lastly, in the fourth stage, classification techniques are employed to forecast student performance using both the original pre-processed data and the enhanced data, and the effectiveness of various classifiers is compared by determining the accuracy, precision, recall, and f1-score of each method.

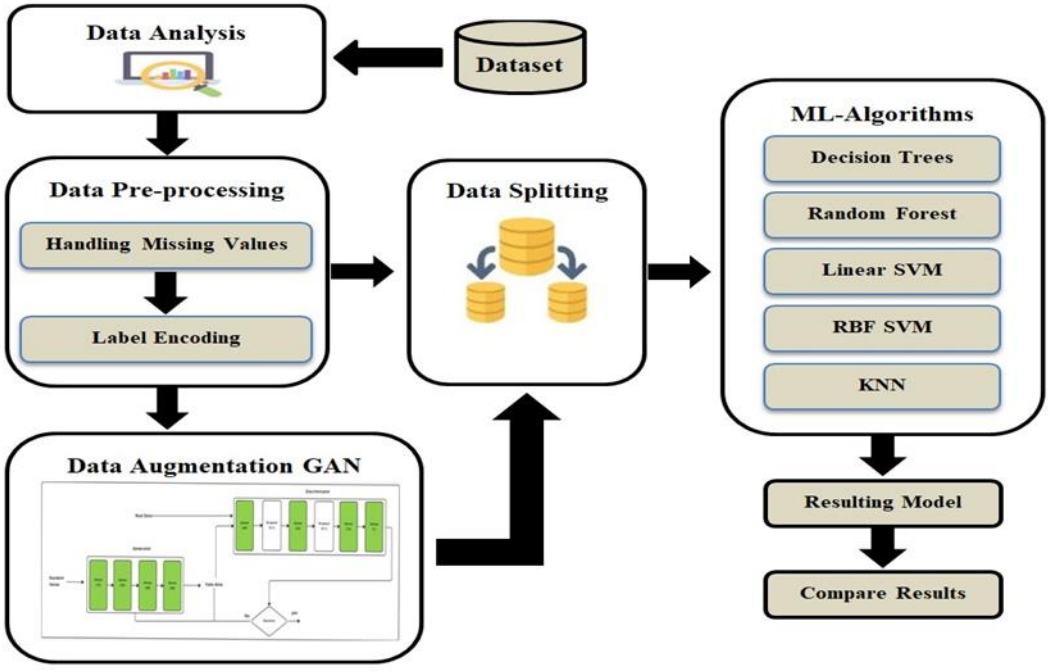

Figure 4.1. Flowchart of the method.

Platform Used. In this research, the machine learning algorithms were implemented using Google Colab Platform and Python (version 3.9.16). Panda's libraries (version 1.4.4) was used for processing, while the scikit libraries (version 1.2.2) were employed for developing the machine learning algorithms. Additionally, the TensorFlow platform (version 2.12.0) was utilized for this work.

4.1. DATA COLLECTION

The dataset [53] is composed of two separate files. (The course of mathematics) (The course of Portuguese) with in csv format. These datasets, all from a school in Portugal, were incorporated into one combined dataset with 1044 rows and 33 columns.

These data were formed from two exporters. paper-based school reports with limited attributes such as three hours of instructional grades and absences, and questionnaires that supplemented school reports. These questionnaires have closed-ended questions that concern various demographic information (e.g., the mother's education, income), social/emotional information (e.g., alcohol consumption), and the previous educational failures of the student. The questionnaire was evaluated by experts in education and tested on a small sample of 15 students that yielded feedback. The final version had 37 questions on a single A4 sheet and was responded to by 788 students during the classroom. Later, 111 responses were removed because of the lack of sufficient information regarding their identity (necessary to complete school reports). These numbers were ultimately gathered into two datasets for math (395 samples) and Portuguese (649 samples). In the preprocessing stage, some features are removed because they lacked a unique value. For instance, few participants disclosed their income from home (this may be for privacy reasons). For detailed description of all functions, see Table 4.1.

Table 4.1. Describe the dataset's features in detail.

Feature Name	Type	Description
1-School	Object	Gabriel Pereira or Mousinho da Silveira
2- Sex	Object	female or male
3-Age	Numeric	15-22
4-Adress	Object	student’s home address type (urban or rural)
5- Pstatus	Object	parent’s cohabitation status (living together or apart)
6- Fedu	Numeric	father’s education (from 0 – 4)
7- Medu	Numeric	mother’s education (0 – 4a)
8- Mjob	Object	mother’s job
9- famsize	Object	family size (≤ 3 or > 3)
10- guardian	Object	student’s guardian (mother, father, or other)
11- failures	Numeric	number of past class failures (n if 1 ≤ n < 3, else 4)
12- internet	Object	Internet access at home (yes or no)
13- travel time	Numeric	home to school travel time (1– < 15 min., 2– 15 to 30 min., 3– 30 min. to 1 hour, or 4– >1 hour).
14- study time	Numeric	weekly study time (1– < 2 hours, 2– 2 to 5 hours, 3– 5 to 10 hours, or 4– > 10 hours)
15- reason	Numeric	reason to choose this school (close to home, school reputation, course preference, or other)
16- schoolsup	Object	extra educational school support (yes or no)
17- paid class	Object	extra paid classes (yes or no)
18- famrel	Numeric	quality of family relationships (from 1– very bad to 5– excellent)
19- free time	Numeric	free time after school (from 1– very low to 5– very high)
20- Walc	Numeric	weekend alcohol consumption (from 1– very low to 5– very high)
21- romantic	Object	with a romantic relationship (yes or no)
22- Dalc	Numeric	workday alcohol consumption (from 1– very low to 5– very high)
23- health	Numeric	current health status (from 1– very bad to 5– very good)
24- nursery	Object	attended nursery school (yes or no)
25- famsup	Object	family educational support (yes or no)
26- higher	Object	wants to take higher education (yes or no)
27- go out	Numeric	going out with friends (from 1– very low to 5– very high)
28- activities	Object	extra-curricular activities (yes or no)
29- Fjob	Object	father’s job
30- absences	Numeric	number of school absences 0 -93
31- G1	Numeric	f first period grade 0-20
32-G2	Numeric	Second-period grade 0-20
33-G3	Numeric	f in grade 0-20

4.2. DATA PRE-PROCESSING

Real-world data often exhibits inconsistencies, and noise, and may include missing, redundant, or irrelevant information. These issues can adversely impact algorithm performance, leading to inaccurate knowledge and incorrect learning outcomes. Data pre-processing is a crucial step in machine learning methodologies. It attempts to deconstruct, simplify, and convert data into a format that is suitable for the employed algorithms.

First, we create a new column named “Final grade” and filled with ‘na’ values. It assigns one of three categories (‘Poor’, ‘Fair’, or ‘Good’) based on the following score ranges (Figure 4.2.).

- ‘Poor’. A score between 0 and 9 (inclusive).
- ‘Fair’. A score between 10 and 14 (inclusive).
- ‘Good’. A score between 15 and 20 (inclusive).

Then we apply the rest of the preprocessing steps.

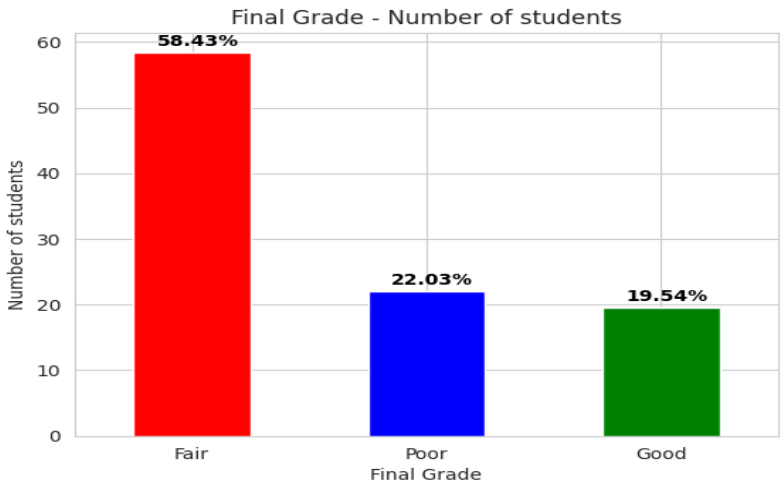

Figure 4.2. Final grade – Number of students.

4.2.1. Data Cleaning

Upon conducting a thorough examination of our dataset for missing or duplicate values, it has been determined that the data is complete and free from any such discrepancies.

4.2.2. Transforming Data

We used label encoding method to change categorical attributes into a single numerical value. In label encoding, each unique category value is assigned a unique integer. For example, consider the "sex" feature in the dataset. The original feature contains two categorical values. "female" and "male". To apply label encoding, we would first create a mapping of the categories to integers, such as.

- "female" -> 0
- "male" -> 1

Then, we would replace the original categorical values in the "sex" feature with their corresponding integer values. The resulting encoded feature would have the values.

- 0 for "female"
- 1 for "male"

This process is suitable for machine learning models without losing the underlying information in the data. However, it's important to note that label encoding does not add any new information to the data and may not be appropriate for all types of categorical data.

4.2.3. Data Augmentation

The Generative Adversarial Network (GAN), which was introduced in 2014 [54], is a network that is trained to address the challenge of generating data. GAN has two complex neural networks that are multi-layered, the discriminator (Disc) and the

generator (Gen). The primary goal of the generator is to create synthetic data by converting random noise into samples that have a meaningful origin, the purpose of the discriminator is to differentiate between genuine and imitated data samples. [55]. Both models are trained simultaneously, progressively enhancing their capabilities.
To train the GAN model, we utilize the following parameters.

Table 4.2. GAN Hyperparameters.

Parameter name	Description	Value / Ranges
optimizer	Updating GAN weights	Adam
batch size	Number of samples in each training	10
learning rate	Updating model	0.0005
noise shape	Dimension of input shape	32
epochs	number of epochs for training model	500
Number of layers	Number of hidden layer in GEN/DESC	4/4 dense layers
activation function	Decide whether function active or not	ReLU
Loss function	Measure performance model	Binary cross entropy

After the data is generated, the dataset is enlarged from 1044 to 46044 rows. The proposed GAN model is depicted in Figure 4.3.

Generator: The Gen has a series of four dense (fully connected) layers that increase in size. The first layer is comprised of the rectified linear function (ReLU) and has 12 neurons. The second and third layers have 24 and 48 neurons, respectively, and both also utilize ReLU activation. The final over-dense layer has 58 neurons, which are the exact same size as the generated data.

Discriminator: The Disc is made up of a series of dense layers with ReLU functions that are also used to dropout to avoid overfitting. The first dense layer is composed of 48 neurons, with the number of neurons halving in each subsequent layer. The output of the final dense layer is a sigmoid function that produces a probability value that indicates the degree to which the input data is genuine or manufactured.

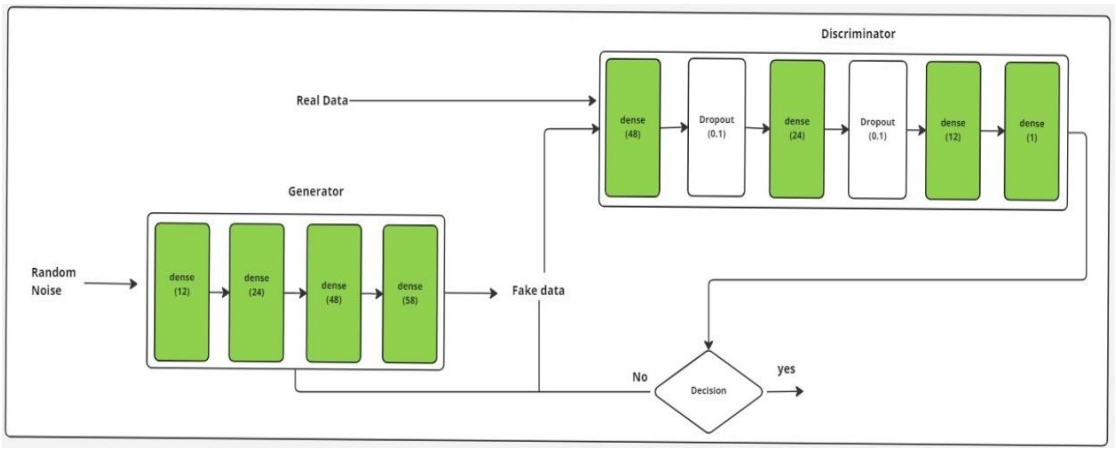

Figure 4.3. The proposed GAN model.

4.3. USING OF ML ALGORITHMS

Here we will describe the manner in which of the machine learning methods employed in the study were implemented, these methods were created using the scikit-learn library. Scikit-learn is a powerful platform that can be utilized to create machine learning models and conduct pre-processing and validation. Before employing the ML methods, the dataset was split into training and test sets, with 70% of the data utilized for training and 30% for testing. The methods of ML were then used to predict student success before and after additional data was incorporated. The ML methods were executed in two steps. In the first step, they utilized the first data, which is preprocessed original data. In the second step, the ML methods were utilized on a hybrid data set that is a combination of Dataset 1 and the data generated using a GAN.

4.3.1. Using of (DT) Algorithm

This algorithm was employed for student performance predicting. One of ML algorithms categorizes or predicts results based on specific data properties. It's similar to a tree, with internal nodes that represent attributes, branches that represent decisions, and leaves that represent results. The algorithm chooses the most beneficial attribute to describe the data as.

its ability to reduce entropy. The information is then organized into groups based on the value of the selected attribute; the process is repeated on each group until the tree has achieved its maximum size.

First, we import the “Decision Tree Classifier” class from the “sklearn. tree” module, which is an ML model used for classification tasks. The “Decision Tree Classifier” is a type of DT algorithm that learns from the data by dividing it into smaller subsets, according to features. Then we create an instance of the “ Decision Tree Classifier” class. This instance can then be dedicated to training on a labeled dataset and making predictions on new, untrained data. The DT has been employed using the default parameters to predict student performance in two separate stages before and after data augmentation.

For A 'Poor' class, there are 150 true positive predictions, meaning the model correctly identified 150 instances as 'Poor'. However, it misclassified 7 instances as 'Fair' and 20 instances as 'Good' that were actually 'Poor'.

- For the 'Fair' class, there are 49 true positive predictions, indicating that the model accurately identified 49 instances as 'Fair'. It misclassified 9 instances as 'Poor' that were in fact 'Fair'.
- For the 'Good' class, there are 60 true positive predictions, meaning the model correctly recognized 60 instances as 'Good'. Nonetheless, it misclassified 19 instances as 'Poor' that were actually 'Good'.

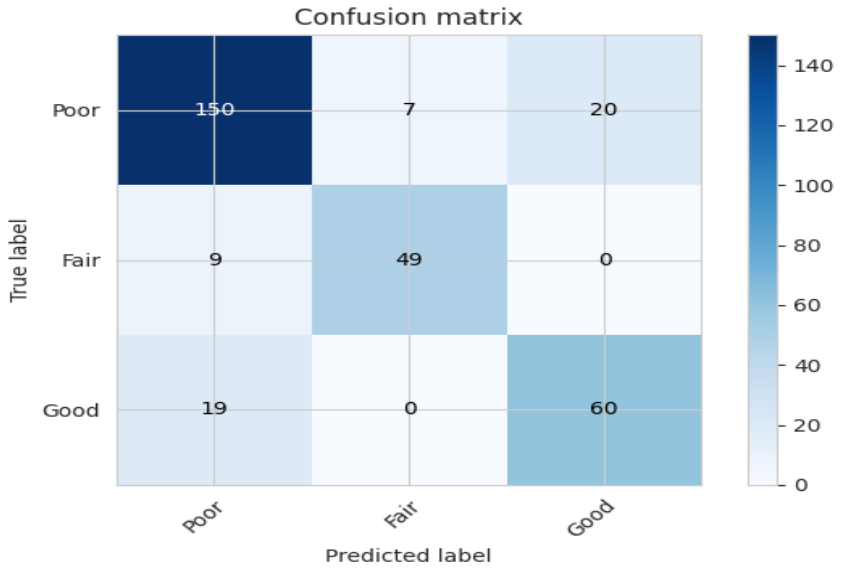

Figure 4.4. Confusion matrix for using the DT algorithm on first dataset.

Step two, DT was employed on the second dataset.

- For the 'Poor' class, there are 4532 true positive predictions, meaning the model correctly identified 4532 instances as 'Poor'. However, it misclassified 17 instances as 'Fair' that were actually 'Poor'.
- For the 'Fair' class, there are 4638 true positive predictions, indicating that the model accurately identified 4638 instances as 'Fair'. It misclassified 17 instances as 'Poor' and 10 instances as 'Good' that were actually 'Fair'.
- For the 'Good' class, there are 4594 true positive predictions, meaning the model correctly recognized 4594 instances as 'Good'. Nonetheless, it misclassified 6 instances as 'Fair' that were actually 'Good'.

After data augmentation, the Decision Tree model exhibits remarkable performance in accurately categorizing instances into their corresponding classes. The model demonstrates only a minimal number of misclassifications between the 'Poor' and 'Fair' classes, as well as between the 'Fair' and 'Good' classes. This signifies a substantial enhancement in classification accuracy when compared to the model's performance before data augmentation.

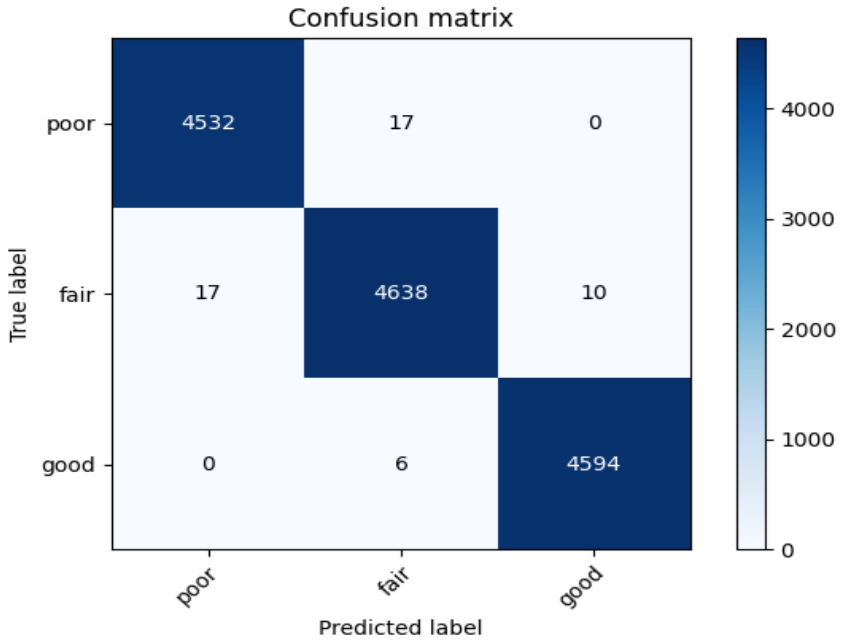

Figure 4.5. Confusion matrix using for the DT algorithm second data.

4.3.2. Using of Random Forest (RF) Algorithm

This algorithm is employed for estimate student's performance. It's one of the most common taxonomy algorithms.

- For A 'Poor' class, there are 160 true positive predictions, meaning the model correctly identified 160 instances as 'Poor'. However, it misclassified 4 instances as 'Fair' and 13 instances as 'Good' that were actually 'Poor'.
- For the 'Fair' class, there are 47 true positive predictions, indicating that the model accurately identified 47 instances as 'Fair'. It misclassified 11 instances as 'Poor' that were in fact 'Fair'.
- For the 'Good' class, there are 53 true positive predictions, meaning the model correctly recognized 53 instances as 'Good'. Nonetheless, it misclassified 26 instances as 'Poor' that were actually 'Good'.

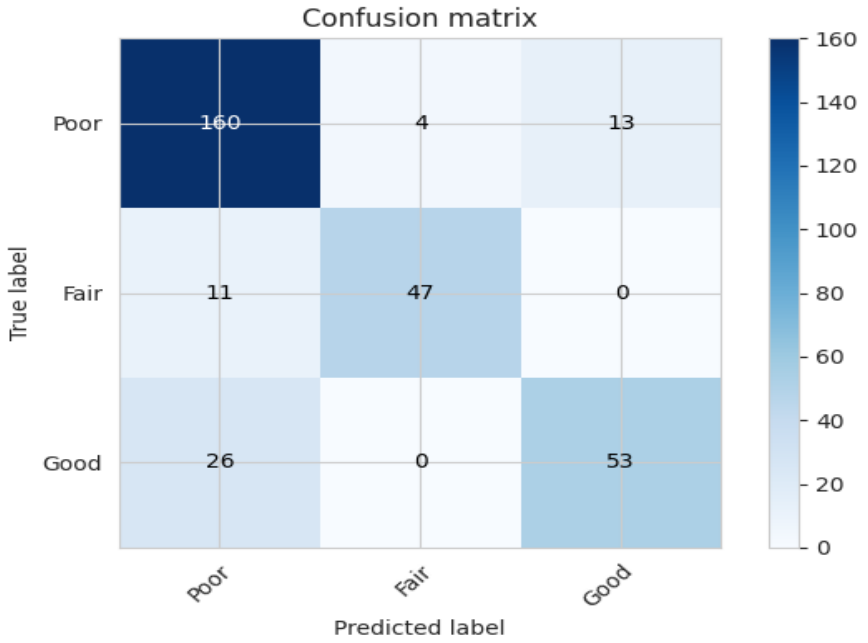

Figure 4.6. Confusion matrix using for RF algorithm on first data.

Step two, RF was employed on the second data.

- For A 'Poor' class, there are 4526 true positive predictions, meaning the model correctly identified 4526 instances as 'Poor'. However, it misclassified 23 instances as 'Fair' that were actually 'Poor'.
- For the 'Fair' class, there are 4646 true positive predictions, indicating that the model accurately identified 4646 instances as 'Fair'. It misclassified 15 instances as 'Poor' and 4 instances as 'Good' that were actually 'Fair'.
- For the 'Good' class, there are 4591 true positive predictions, meaning the model correctly recognized 4591 instances as 'Good'. Nonetheless, it misclassified 9 instances as 'Fair' that were actually 'Good'.

The Random Forest model demonstrates outstanding performance in classifying the instances into their respective classes after data augmentation. The model exhibits only a few misclassifications between 'Poor' and 'Fair', as well as between 'Fair' and 'Good' classes. This indicates a significant improvement in classification accuracy compared to the model's performance before data augmentation.

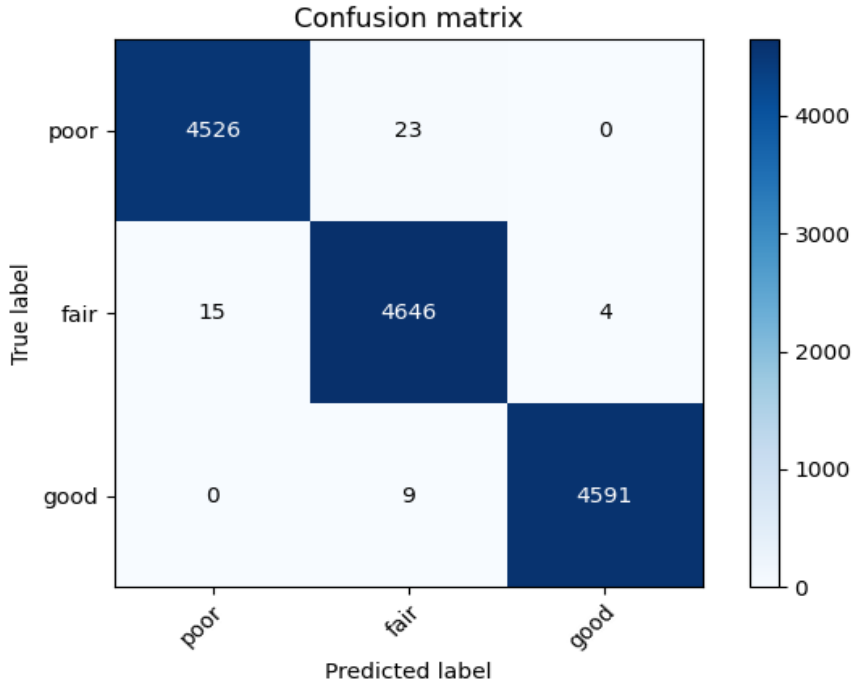

Figure 4.7. Confusion matrix using for RF algorithm on second data.

4.3.3. Using of (KNN) Algorithm

The K-Nearest Neighbors technique was employed to estimate student's performance. KNN is a simple algorithm that is primarily used in voting in its work. It is employed to estimate a student's success in two steps. Firstly, it's employed on the original data.

- For class 'Poor', there are 158 true positive predictions, meaning the model correctly identified 158 instances as 'Poor'. However, it misclassified 7 instances as 'Fair' and 12 instances as 'Good' that were actually 'Poor'.
- For the 'Fair' class, there are 49 true positive predictions, indicating that the model accurately identified 49 instances as 'Fair'. It misclassified 9 instances as 'Poor' that were in fact 'Fair'.
- For the 'Good' class, there are 55 true positive predictions, meaning the model correctly recognized 55 instances as 'Good'. Nonetheless, it misclassified 24 instances as 'Poor' that were actually 'Good'.

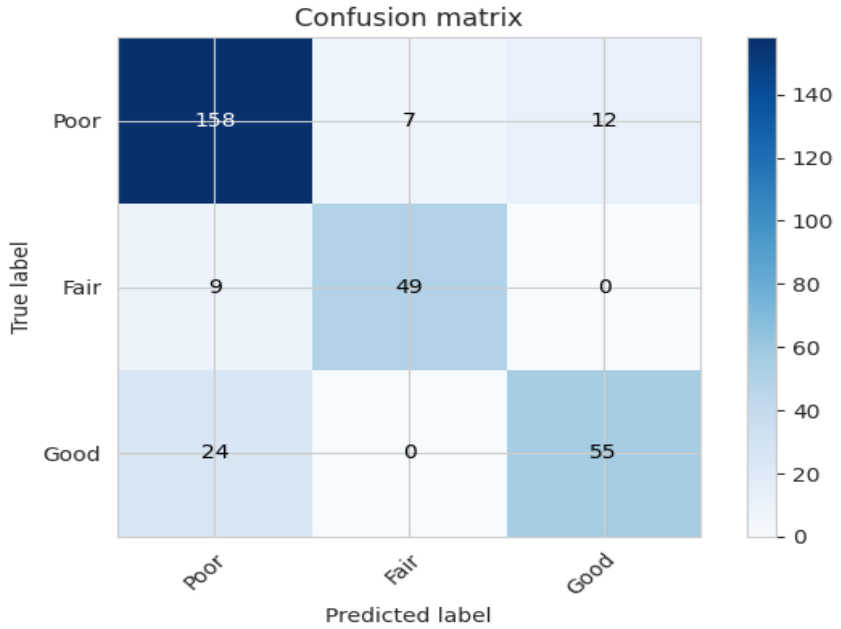

Figure 4.8. Confusion matrix using for KNN algorithm first data.

Second step, Its employed on the augmented data.

- For the 'Poor' class, there are 4526 true positive predictions, meaning the model correctly identified 4526 instances as 'Poor'. However, it misclassified 23 instances as 'Fair' that were actually 'Poor'.
- For the 'Fair' class, there are 4641 true positive predictions, indicating that the model accurately identified 4641 instances as 'Fair'. It misclassified 17 instances as 'Poor' and 7 instances as 'Good' that were actually 'Fair'.
- For the 'Good' class, there are 4582 true positive predictions, meaning the model correctly recognized 4582 instances as 'Good'. Nonetheless, it misclassified 18 instances as 'Fair' that were actually 'Good'.

In the aftermath of data augmentation, the KNN model showcases an impressive performance in distinguishing instances and assigning them to their appropriate classes. The model displays a limited number of misclassifications between the 'Poor' and 'Fair' classes, as well as between the 'Fair' and 'Good' classes. This highlights a notable improvement in classification accuracy when contrasted with the model's performance before data augmentation.

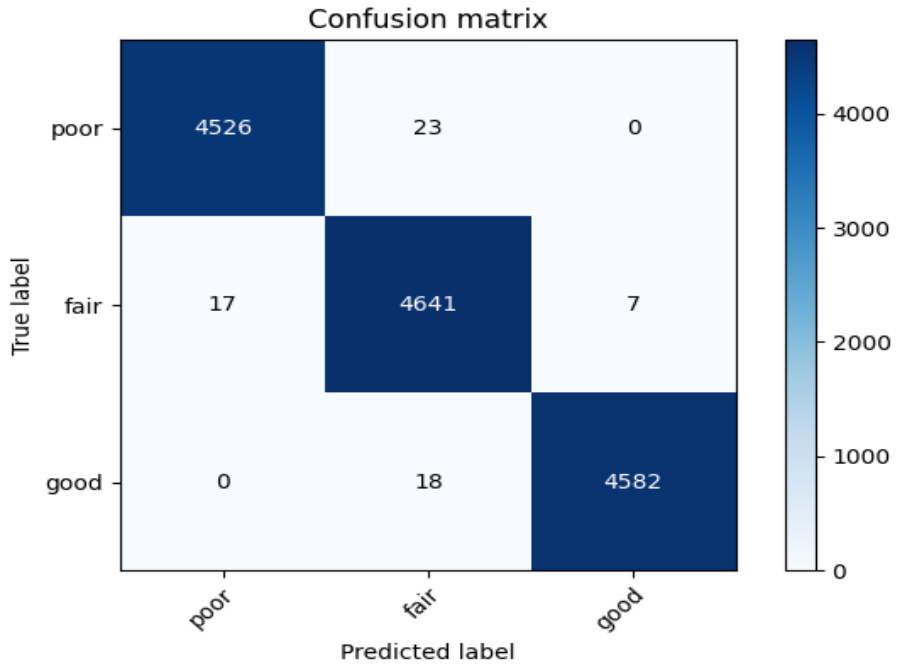

Figure 4.9. Confusion matrix using for the KNN algorithm second data.

4.3.4. Using of (Linear SVM) Algorithm

The using of these algorithm on the original data is explain as bellow.

- For the 'Poor' class, there are 150 true positive predictions, meaning the model correctly identified 150 instances as 'Poor'. However, it misclassified 12 instances as 'Fair' and 15 instances as 'Good' that were actually 'Poor'.
- For the 'Fair' class, there are 46 true positive predictions, indicating that the model accurately identified 46 instances as 'Fair'. It misclassified 12 instances as 'Poor' that were in fact 'Fair'.
- For the 'Good' class, there are 64 true positive predictions, meaning the model correctly recognized 64 instances as 'Good'. Nonetheless, it misclassified 15 instances as 'Poor' that were actually 'Good'.

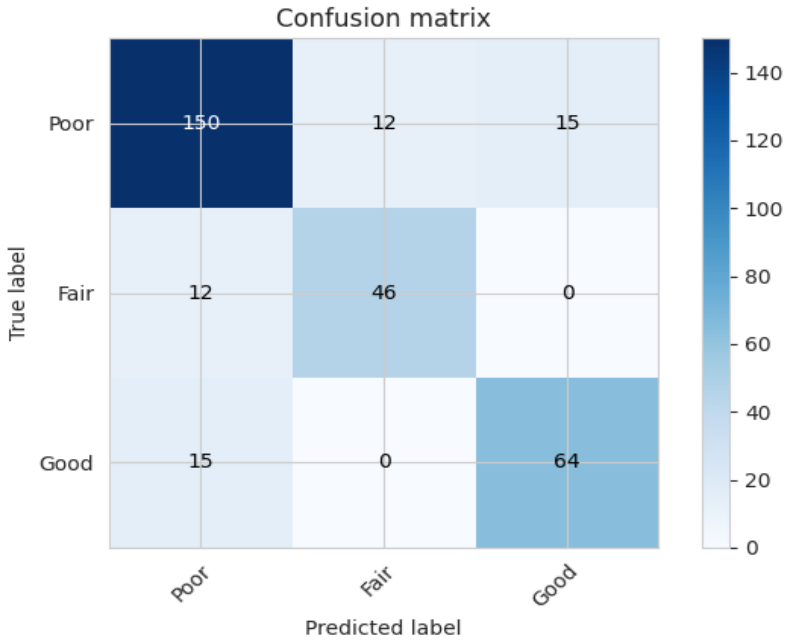

Figure 4.10. Confusion matrix using for Linear SVM algorithm on first data.

- Step two, these algorithms employed on the augmented data.
- For the 'Poor' class, there are 4536 true positive predictions, meaning the model correctly identified 4536 instances as 'Poor'. However, it misclassified 13 instances as 'Fair' that were actually 'Poor'.
- For the 'Fair' class, there are 4638 true positive predictions, indicating that the model accurately identified 4638 instances as 'Fair'. It misclassified 18 instances as 'Poor' and 9 instances as 'Good' that were actually 'Fair'.
- For the 'Good' class, there are 4592 true positive predictions, meaning the model correctly recognized 4592 instances as 'Good'. Nonetheless, it misclassified 8 instances as 'Fair' that were actually 'Good'.

Following data augmentation, the Linear SVM model demonstrates exceptional performance in accurately classifying instances into their respective classes. The model exhibits a minimal number of misclassifications between the 'Poor' and 'Fair' classes, as well as between the 'Fair' and 'Good' classes. This reveals a significant enhancement in classification accuracy compared to the model's performance before data augmentation.

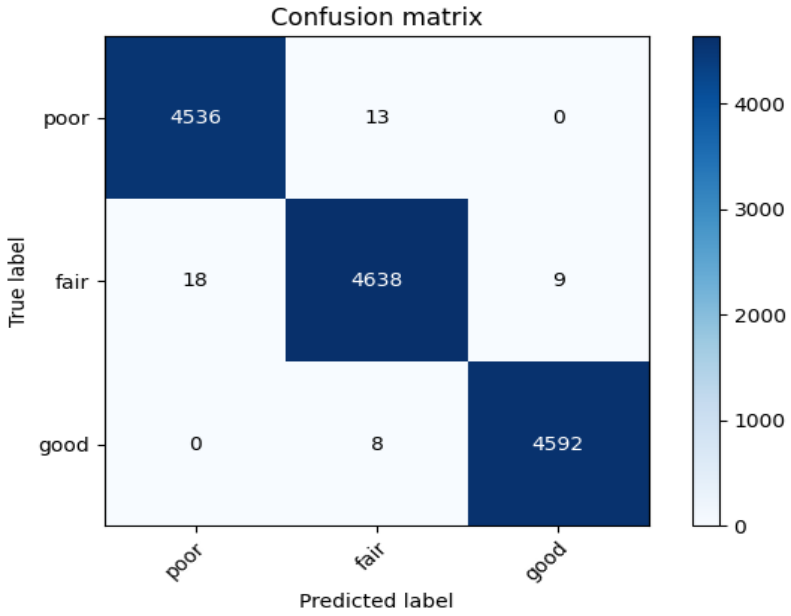

Figure 4.11. Confusion matrix using for SVM algorithm on the second data.

4.3.5. Using (RBF SVM) Algorithm

In the first stage, the RBF(SVM) algorithm was employed on the real data.

- For class 'Poor', there are 164 true positive predictions, meaning the model correctly identified 164 instances as 'Poor'. However, it misclassified 3 instances as 'Fair' and 10 instances as 'Good' that were actually 'Poor'.
- For the 'Fair' class, there are 47 true positive predictions, indicating that the model accurately identified 47 instances as 'Fair'. It misclassified 11 instances as 'Poor' that were in fact 'Fair'.
- For the 'Good' class, there are 56 true positive predictions, meaning the model correctly recognized 56 instances as 'Good'. Nonetheless, it misclassified 23 instances as 'Poor' that were actually 'Good'.

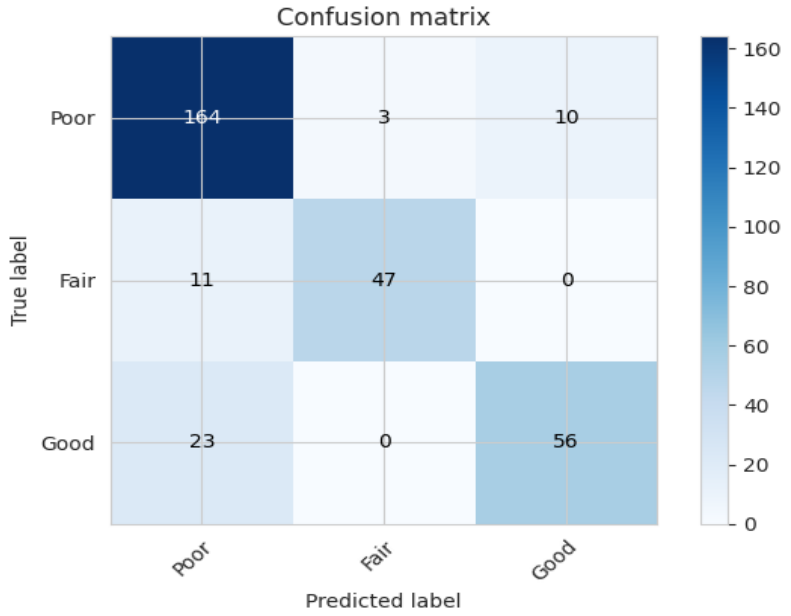

Figure 4.12. Confusion matrix using for the RBF SVM on first data.

In step two, the RBF SVM algorithm was employed on dataset 2. Figure 4.13. illustrates the confusion matrix for the RBF SVM on dataset 2.

- For the 'Poor' class, there are 4530 true positive predictions, meaning the model correctly identified 4530 instances as 'Poor'. However, it misclassified 19 instances as 'Fair' that were actually 'Poor'.
- For the 'Fair' class, there are 4647 true positive predictions, indicating that the model accurately identified 4647 instances as 'Fair'. It misclassified 15 instances as 'Poor' and 3 instances as 'Good' that were actually 'Fair'.
- For the 'Good' class, there are 4590 true positive predictions, meaning the model correctly recognized 4590 instances as 'Good'. Nonetheless, it misclassified 10 instances as 'Fair' that were actually 'Good'.

After data augmentation, the RBF SVM model exhibits outstanding performance in accurately classifying instances into their corresponding classes. The model demonstrates only a limited number of misclassifications between the 'Poor' and 'Fair' classes, as well as between the 'Fair' and 'Good' classes. This indicates a considerable improvement in classification accuracy when compared to the model's performance before data augmentation.

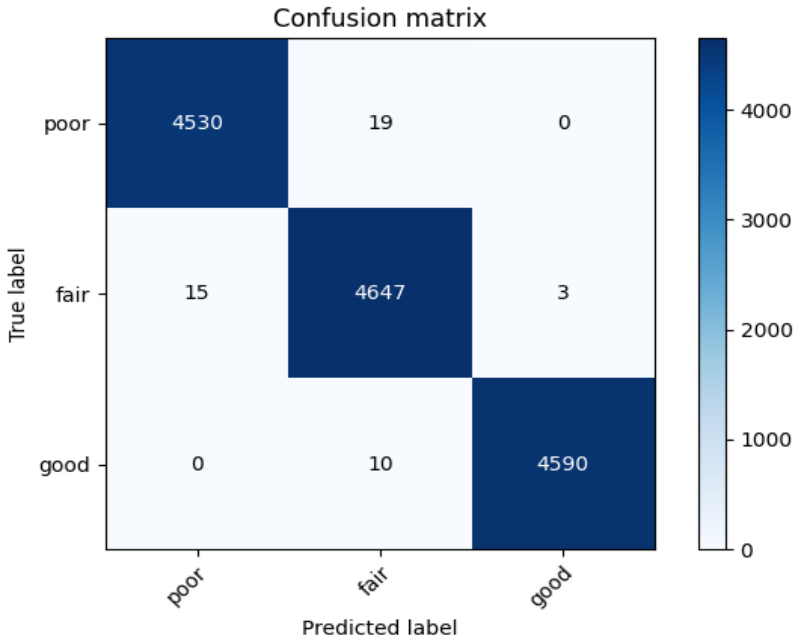

Figure 4.13. Confusion matrix using for the RBF SVM on the second data.

4.4. PERFORMANCE EVALUATION

Several metrics were used to assess the effectiveness of each method, including precision, recall, accuracy, and F-Score. The confusion matrix is a visual depiction of the algorithm's performance. The rows of the matrix are the actual classifications, while the columns are the predicted classifications. The components of the matrix are the number of samples that were divided into each possible combination of real and predicted labels.

For the binary predictions in Table 4.2, the matrix is comprised of two rows and two columns, representing the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

Table 4.3. Confusion matrix for binary classification.

	Predicted class 1	**Predicted class 2**
Actual class 1	True Positives (TP)	False Positives (FP)
Actual class 2	False Negative (FN)	True Positive (TN)

For the multiple-class classification that involves three classes (Table 4.3), the confusion matrix is a 3x3 matrix that represents the number of correctly and incorrectly classified samples for each class. In this matrix, the true positive (TP) value for each class is the number of samples that were correctly classified as belonging to that class. The false positive (FP) value represents the number of samples that were incorrectly assigned to a particular class. The false negative (FN) value represents the number of samples that were misclassified as not belonging to a particular class.

Table 4.4. Confusion matrix for multi-class classification

	Class 1 Predictions	**Class 2 Predictions**	**Class 3 Predictions**
Actual Class1	True Positives (TP1)	False Negatives (FN1)	False Negatives (FN1)
Actual Class2	False Positives (FP2)	True Positives (TP2)	False Negatives (FN2)
Actual Class3	False Positives (FP3)	False Positives (FP3)	True Positives (TP3)

- Accuracy is a numerical system that calculates the percentage of correctly identified occurrences, including those that are classified as anemics, as well as those that are not classified as such. It's calculated by taking the total number of tested samples and dividing it by the total number of TN and TP.

$$\text{Accuracy} = (TN + TP)/(TN + TP + FN + FP)$$

- Precision is a numerical system that calculates the percentage of true positives (anemics that are actually anemics) among the total number of positive samples. It's calculated by taking the number of correct predictions plus the total number of positive samples that were predicted.

$$\text{Precision} = TP/(TP + FP)$$

- Recall is a percentage that calculates the true positives number among overall of the actual positives. It's calculated by taking the number of correct predictions plus the total number of actual positive samples.

F-Score is a numerical system that aggregates precision and recall into one number that is then used to evaluate the degree to which a system is accurate. The rising value is 1, and the lowest is 0. It's calculated according to the equations listed in the description.

$$\text{F1 Score} = 2TP/(2TP + FP + FN)$$

PART 5

RESULTS AND DISCUSSION

5.1. EXPERIMENTS AND RESULTS

Five machine learning methods are derived from the scikit-learn library, which is a robust platform for implementing ML models and to pre-process and validate models. The real dataset (dataset 1) has 1044 variables, (610) for Fair students, (230) for poor students, and 204 samples of good students. After augmenting the data, the dataset is now composed of 15610 samples of Fair students, 15230 samples of poor students, and 15204 samples of good students. 70% of data was divided for train and 30% for test.

5.1.1. Statistical Analysis of Data

This part includes all of the sociodemographic features associated with students, the Social/Emotional Status, and the school-related variables.

5.1.1.1. Social-Demographical Variables

The fundamental information regarding social and demographic data found in this part like (educational level of the mother, and the educational level of the father).

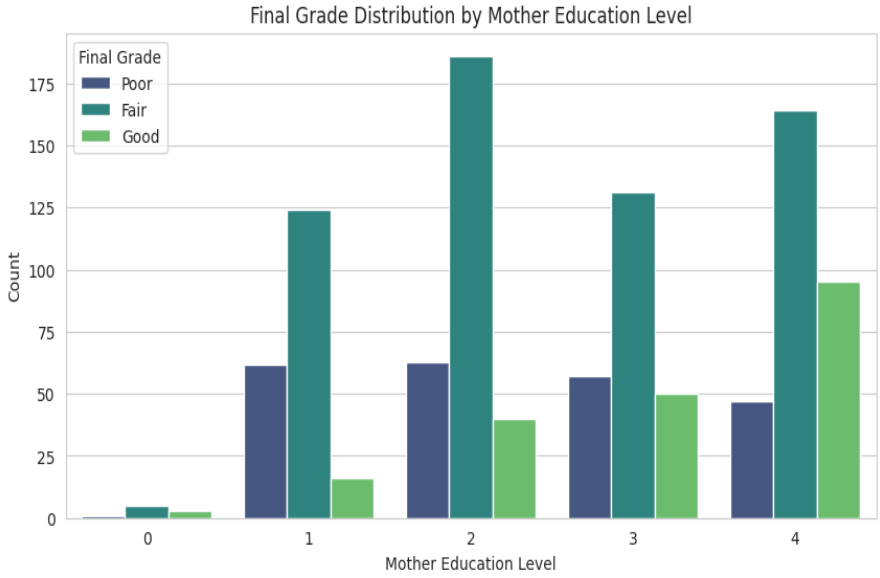

Figure 5.1. Students' performance depends on mother's education level.

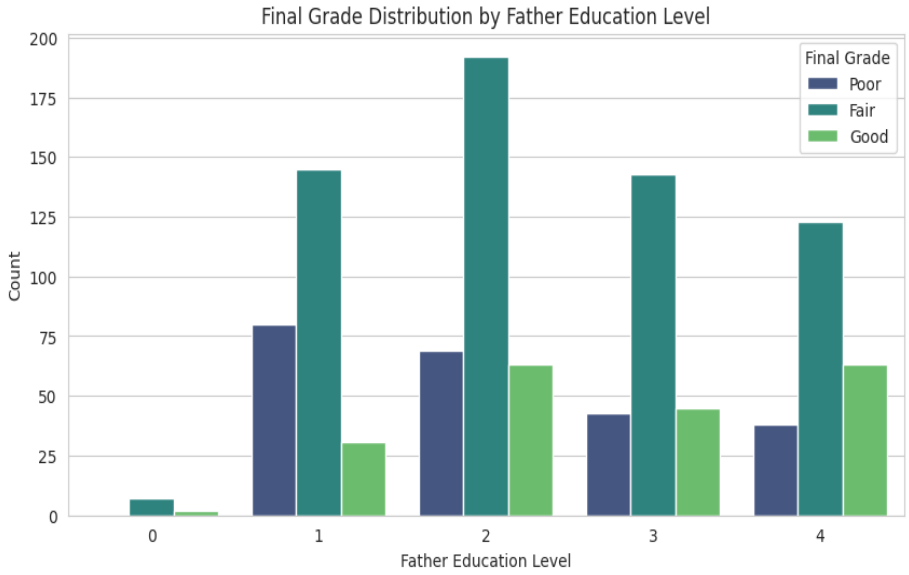

Figure 5.2. Students' performance depends on father's educational level.

5.1.1.2. Social/Emotional

The social and emotional status of the students participating in this study was mentioned, such as (alcohol consumption and romantic status).

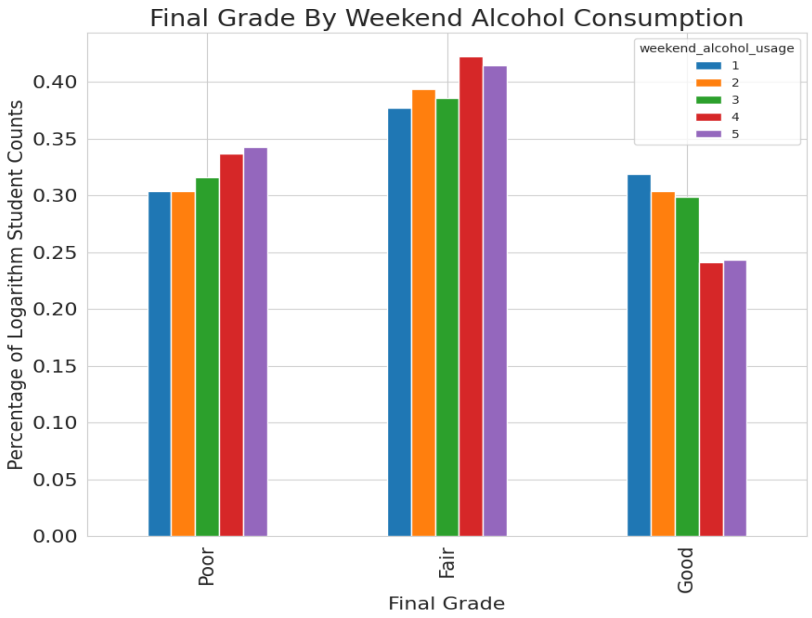

Figure 5.3. Students' performance depends.

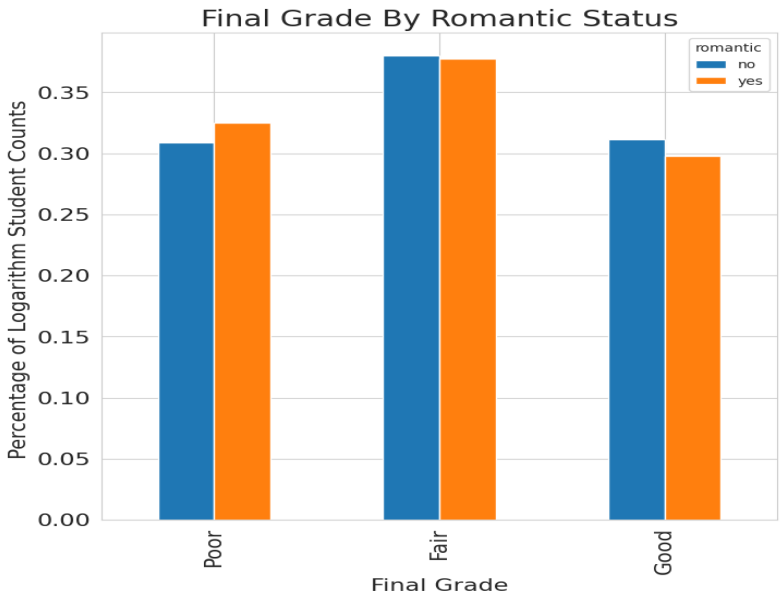

Figure 5.4. Students' performance according to romantic case.

5.1.1.3. School-Related

Students' information that is pertinent to this study is mentioned, including (the time of study, the desire to get higher education, past failures, absences, etc.).

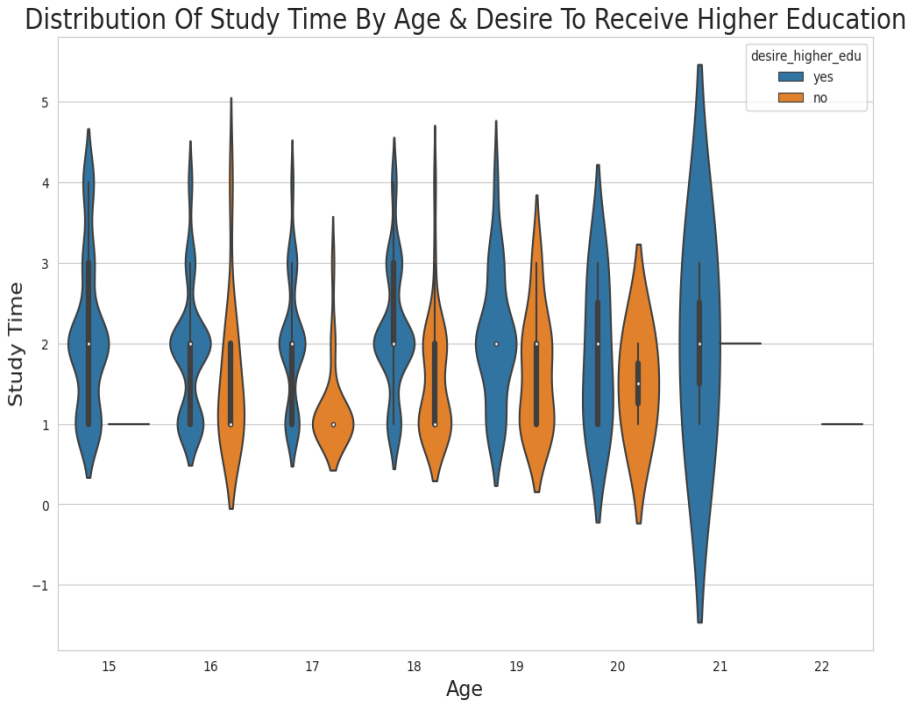

Figure 5.5. The distribution of study time by age and desire to higher education.

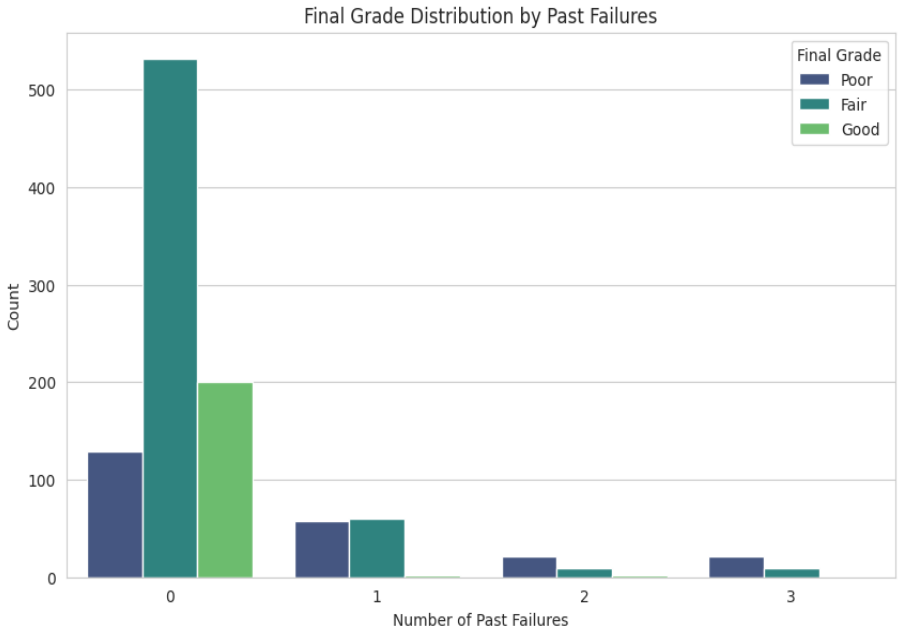

Figure 5.6. Students' performance depends on past failures.

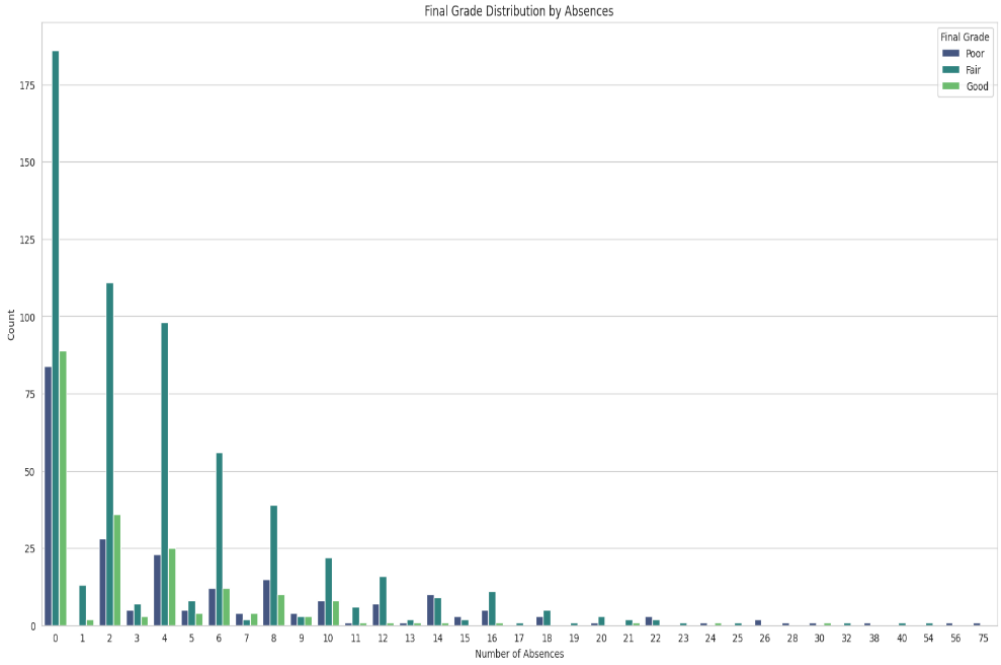

Figure 5.7. Students' performance depends on absences.

5.1.2. Experimental Results on the first data

The students' performance was forecasted when employed real data, which includes the real dataset before data augmentation. The result meant that the RBF SVM had the highest accuracy of 85%. and DT had the lowest accuracy of 82.4%.

Table 5.1. Performance of ML methods implemented on real data.

Algorithms	Accuracy	Class	Precision	Recall	F1 score
DT	82.4%	Poor	0.84	0.85	0.85
		Fair	0.88	0.84	0.86
		Good	0.75	0.76	0.75
RF	82.8%	Poor	0.81	0.90	0.86
		Fair	0.92	0.81	0.86
		Good	0.80	0.67	0.73
KNN	83.4%	Poor	0.83	0.89	0.86
		Fair	0.88	0.84	0.86
		Good	0.82	0.70	0.75
Linear SVM	82.8%	Poor	0.85	0.85	0.85
		Fair	0.79	0.79	0.79
		Good	0.81	0.81	0.81
RBF SVM	85%	Poor	0.83	0.93	0.87
		Fair	0.94	0.81	0.87
		Good	0.85	0.71	0.77

The charts of the ROC AUC curve of the ML classifiers before the data augmentation process will be shown below.

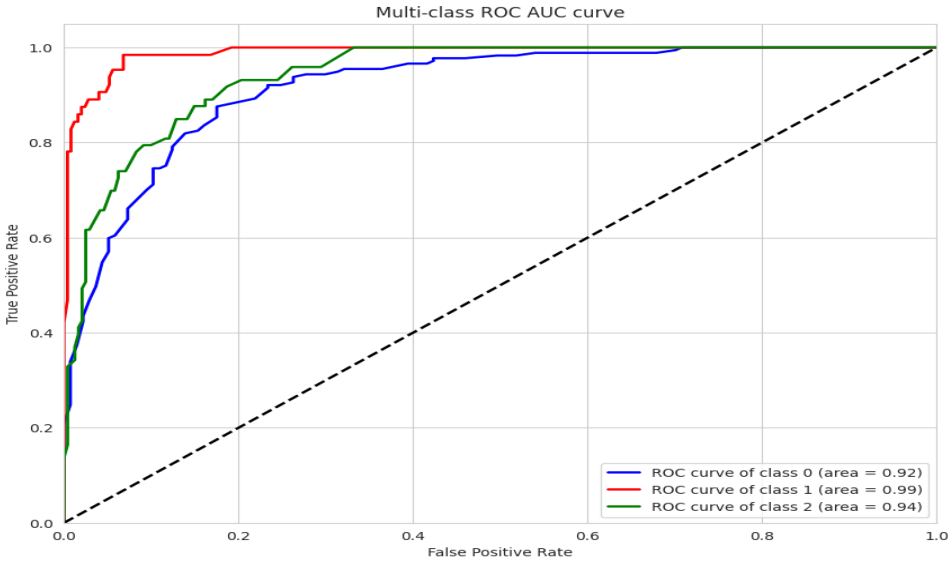

Figure 5.8. RF classifier ROC AUC cuve.

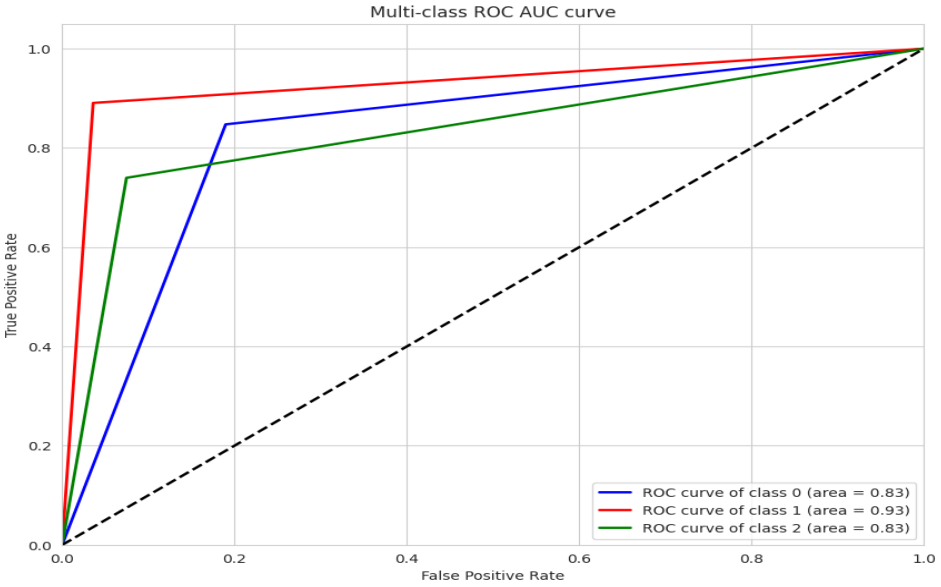

Figure 5.9. DT classifier ROC AUC curve.

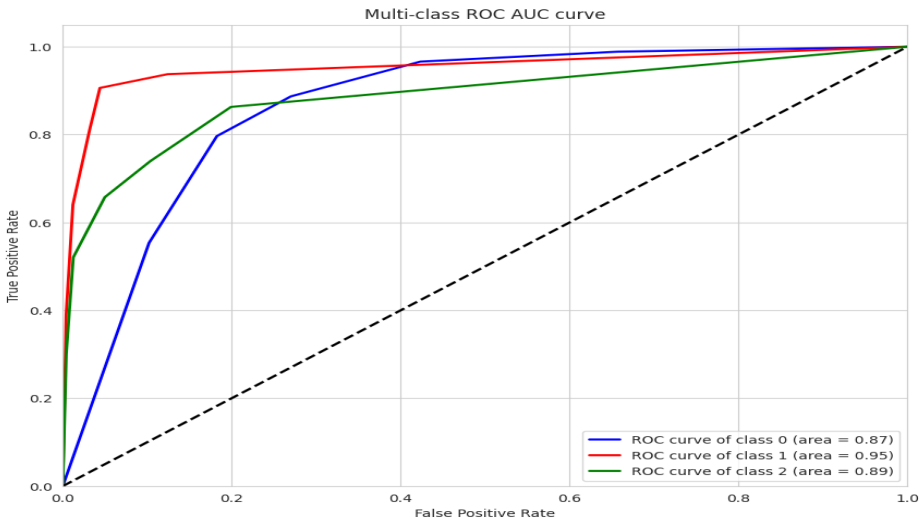

Figure 5.10. KNN classifier ROC AUC cuve.

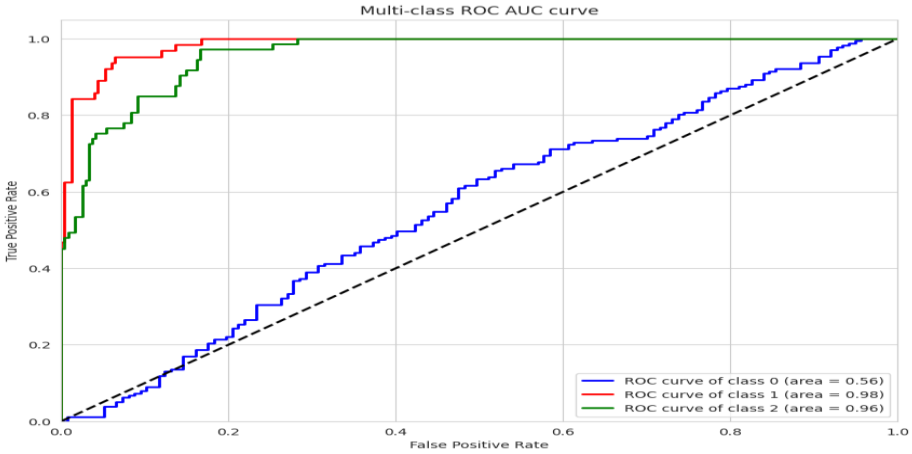

Figure 5.11. Linear SVM classifier ROC AUC cuve.

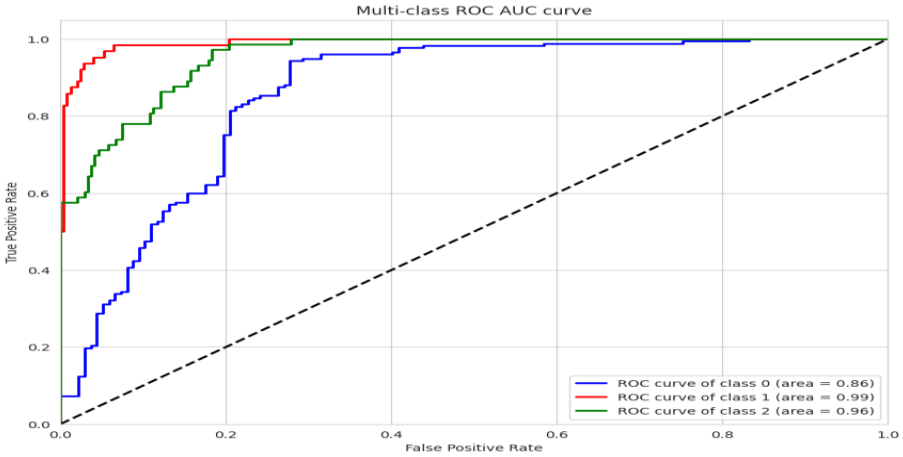

Figure 5.12. RBF SVM classifier ROC AUC cuve.

5.1.3. Experimental Results on Dataset 2

The augmented data was employed for student's performance predicting using GAN. The results were greatly enhanced for all of the algorithms.

Table 5.2. Results of ML method implemented on augmented data.

Algorithms	Accuracy	Class	Precision	Recall	F1 score
DT	99.63%	Poor	1.00	1.00	1.00
		Fair	1.00	0.99	1.00
		Good	1.00	1.00	1.00
RF	99.63%	Poor	1.00	0.99	1.00
		Fair	0.99	1.00	0.99
		Good	1.00	1.00	1.00
KNN	99.52%	Poor	1.00	0.99	1.00
		Fair	0.99	0.99	0.99
		Good	1.00	1.00	1.00
Linear SVM	99.65%	Poor	1.00	1.00	1.00
		Fair	1.00	0.99	0.99
		Good	1.00	1.00	1.00
RBF SVM	99.65%	Poor	1.00	1.00	1.00
		Fair	0.99	1.00	0.99
		Good	1.00	1.00	1.00

The charts of the ROC AUC curve of the ML classifiers after the data augmentation process will be shown below.

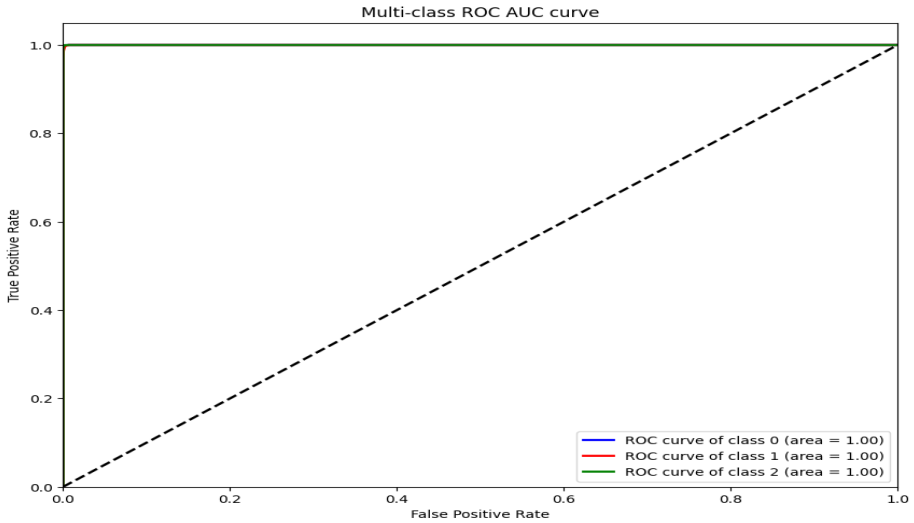

Figure 5.13. RF classifier ROC AUC cuve.

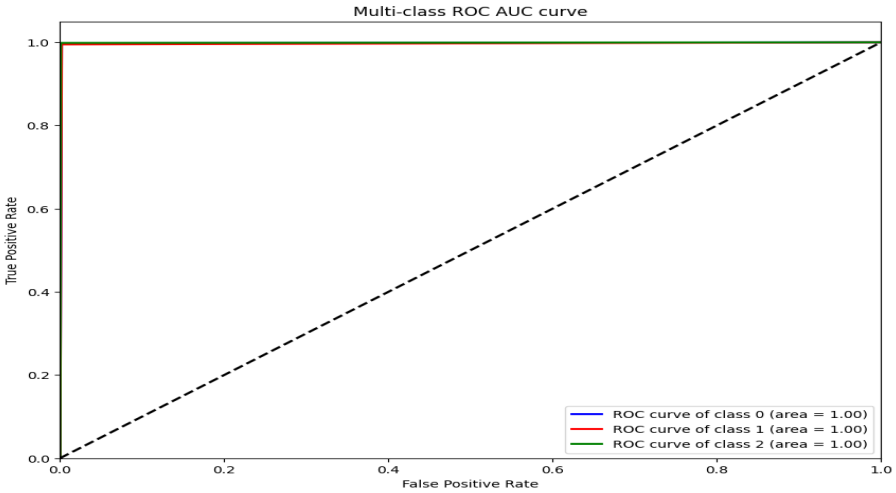

Figure 5.14. DT classifier ROC AUC cuve.

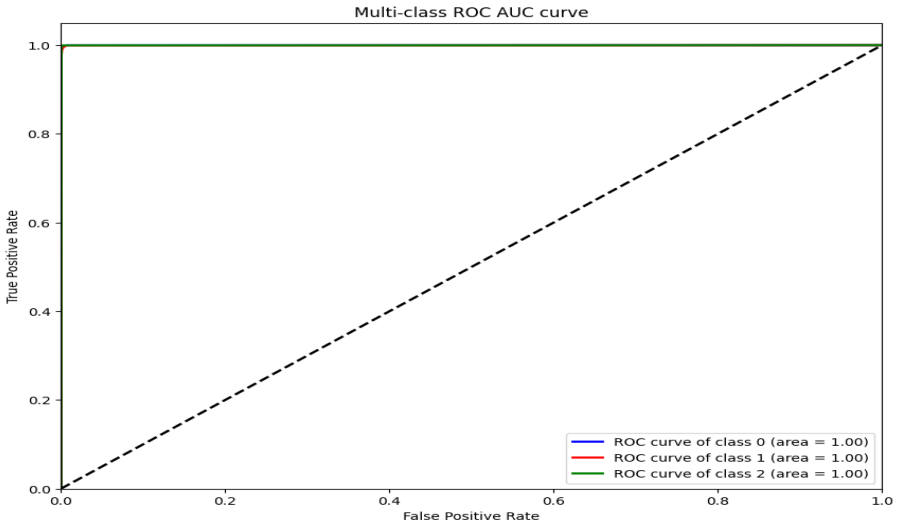

Figure 5. 15. KNN classifier ROC AUC cuve.

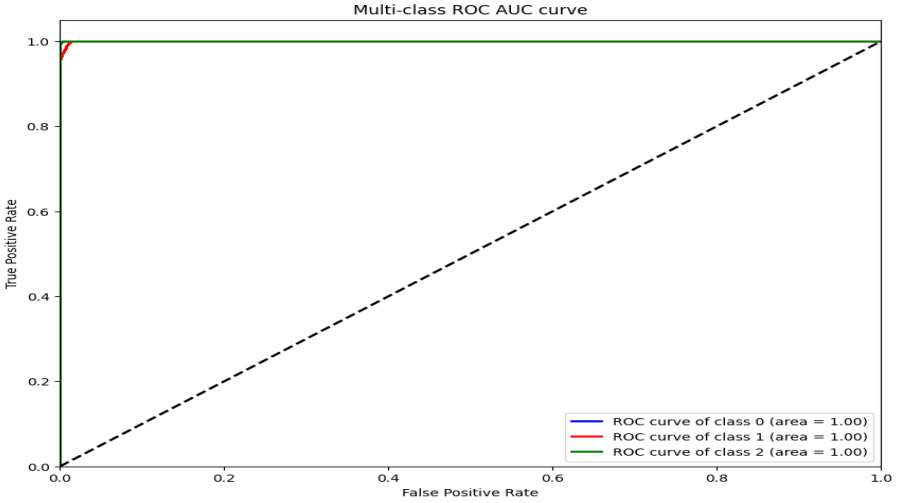

Figure 5.16. Linear SVM classifier ROC AUC cuve.

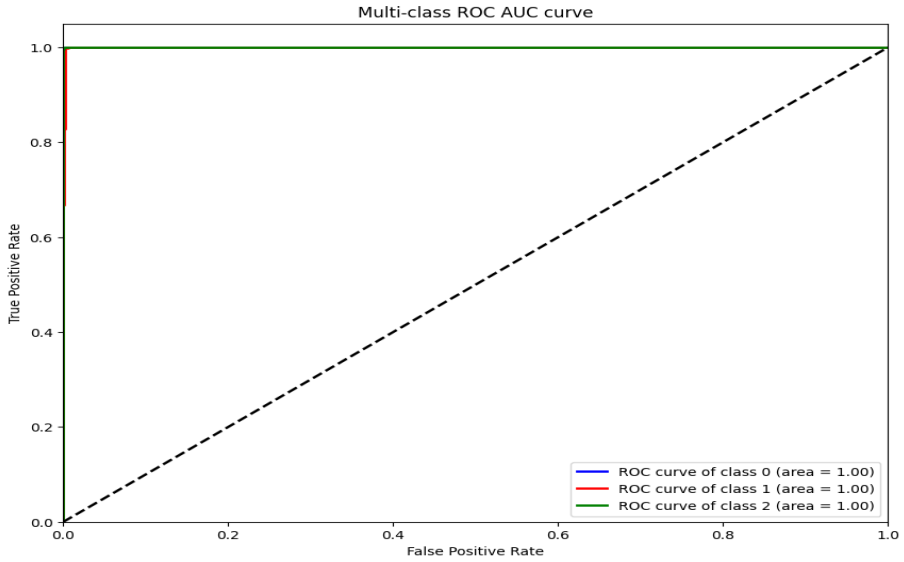

Figure 5.17. RBF classifier ROC AUC cuve.

5.2. DISCUSSION

Before using GAN, the ML classifiers had a high degree of accuracy, with rates between 82.8% and 85 % (Table \ref{tab.tab1}). However, the precision and recall rates were different for the three performance groups, and classifiers had a greater success in predicting poor academic performance than they had in predicting good results.

After implementing GAN, all classifiers' performance increased significantly, with the accuracy rate increasing from 99.52% to 99.65% (Table \ref{tab.tab2}). Additionally, all classifiers had flawless accuracy, recall, and F1 score in all categories of performance, this demonstrated the models' capacity to accurately predict the performance of students in all academic disciplines.

The increased performance following the introduction of GAN is attributed to the larger size and more diverse dataset. GAN generates fake data that resembles the original data exactly, this increases the size of the dataset while also providing new examples that the classifiers have not encountered before. This larger variety of data may have benefited the classifiers by providing additional information that enhanced their ability to generalize and make more accurate predictions.

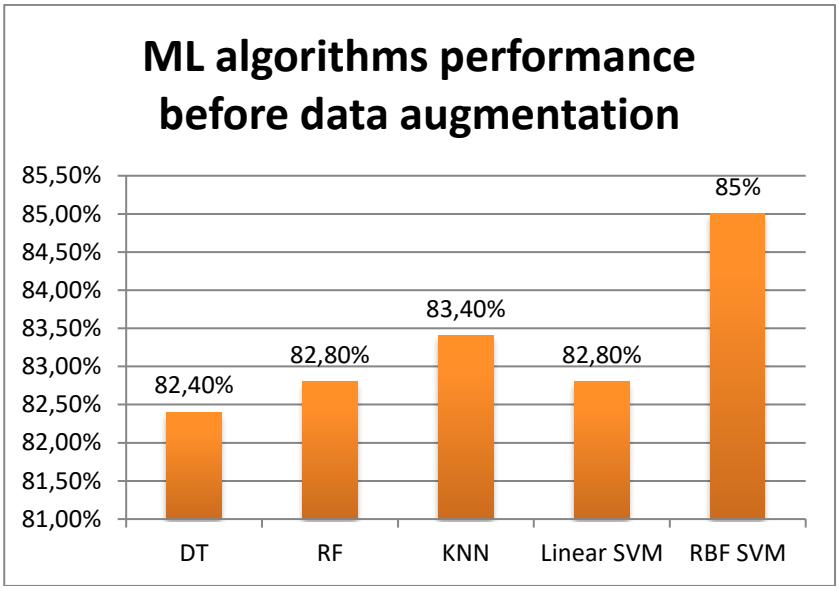

Figure 5.18. Results of ML techniques on data set 1.

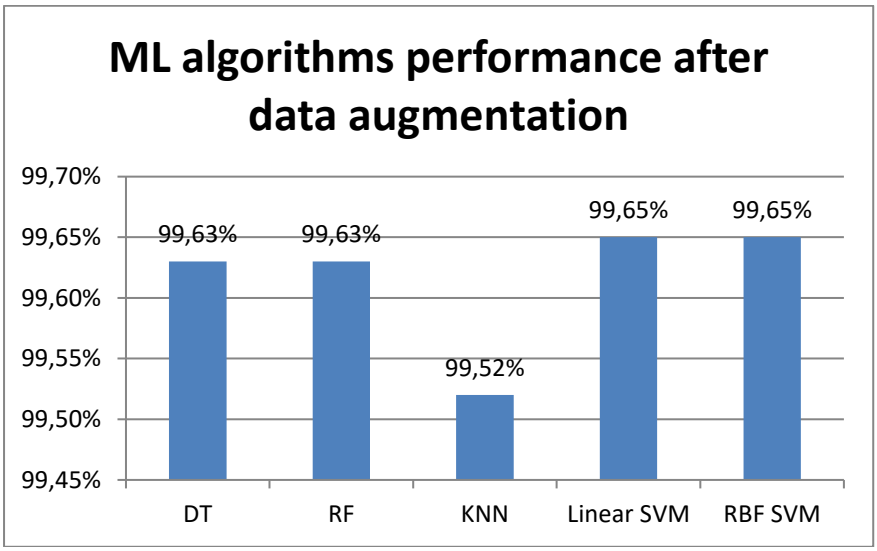

Figure 5.19. Results of ML techniques on data set 2.

PART 6

CONCLUSION

Predicting students' success has become significant in the evaluation of educational accomplishments. The capacity to predict academic success is beneficial in addressing various academic concerns, including the long time required to graduate and drop out of college. Additionally, student performance predictions help to analyze students' behavior during the learning process, increase the environmental quality of learning, address issues associated with student learning, and facilitate data-based decision-making. Our study introduced a promising method of predicting student performance. By utilizing GAN algorithms to augment the data, the results show that data augmentation has a significant effect on the accuracy and reliability of these predictions (RF 99.6, DT 99.6, KNN 99.5, Linear SVM 99.6, RBF SVM 99.6). This research has the potential to significantly impact the educational field by providing a more comprehensive and accurate model for predicting student performance, promoting informed decision-making, and improving educational results.

REFERENCES

1. Haoxiang, W. and Smys, S., 2021. Big Data Analysis and Perturbation using Data Mining Algorithm. **Journal of Soft Computing Paradigm (JSCP)**, 3(01), pp.19-28.

2. Shang, H., Lu, D. and Zhou, Q., 2021. Early warning of enterprise finance risk of big data mining in the Internet of things based on fuzzy association **rules. Neural Computing and Applications,** 33(9), pp.3901-3909.

3. Albreiki, B., Zaki, N. and Alashwal, H., 2021. A systematic literature **review of student performance prediction using machine learning techniques.** Education Sciences, 11(9), p.552.

4. Adekitan, A.I. and Noma-Osaghae, E., 2019. Data mining approach to predicting the performance of first-year students in a university using the admission requirements. **Education and Information Technologies, 24(2), pp.1527-1543.**

5. Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R. and Van Erven, G., 2019. Educational data mining. Predictive analysis of the academic performance of public school students in the capital of Brazil. **Journal of Business Research, 94, pp.335-343.**

6. Ghorbani, R. and Ghousi, R., 2020. Comparing different resampling methods in predicting students' performance using machine learning techniques. **IEEE Access**, 8, pp.67899-67911.

7. Oladipupo, O.O. and Olugbara, O.O., 2019. **Evaluation of data analytics-based clustering algorithms for knowledge mining in student engagement data.** Intelligent Data Analysis, 23(5), pp.1055-1071.

8. Jalota, C. and Agrawal, R., 2019, February. Analysis of educational data mining using classification. In 2019 **International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)** (pp. 243-247). IEEE.

9. Francis, B.K. and Babu, S.S., 2019. Predicting the academic performance of students using a hybrid data mining approach. **Journal of medical systems,** 43(6), pp.1-15.

10. Hamza, H.A.A. and Kommers, P., 2018. **A review of educational data mining tools & techniques. International Journal of Educationa3l Technology and Learning,** 3(1), pp.17-23.

11. Romero, C. and Ventura, S., 2020. Educational data mining and learning analytics. An updated survey. Wiley Interdisciplinary **Reviews. Data Mining and Knowledge Discovery**, 10(3), p.e1355.

12. Hernández-Blanco, A., Herrera-Flores, B., Tomás, D. and Navarro-Colorado, B., 2019. A systematic **review of deep learning approaches to educational data mining.** Complexity, 2019.

13. Chaudhury, Pamela, Sushruta Mishra, Hrudaya Kumar Tripathy, and Brojo Kishore."Enhancing the capabilities of Student Result Prediction System." In Proceedings of the Second **International Conference on Information and Communication Technology for Competitive Strategies, p. 91. ACM, 2016.**

14. Mishra, T., Kumar, D., & Gupta, S. (2014, February). Mining students' data for performance prediction. In Fourth **International Conference on Advanced Computing & Communication Technologies (pp. 255-262).**

15. Huang, S., & Fang, N. (2012, October). Work in progress. Early prediction of students' academic performance in an introductory engineering course through different mathematical modeling techniques. In 2012 Frontiers in Education **Conference Proceedings (pp. 1-2). IEEE.**

16. Xing, W., Guo, R., Petakovic, E., & Goggins, S. (2015). Participation-based student final performance prediction model through interpretable Genetic Programming. Integrating learning analytics, educational **data mining and theory. Computers in Human Behavior, 47, 168-181.**

17. Gunnarsson, B. L., & Alterman, R. (2012, April). Predicting failure. A case study in co-blogging. In Proceedings of the **2nd international conference on learning analytics and knowledge (pp. 263-266).**

18. Jayaprakash, S., Krishnan, S. and Jaiganesh, V., 2020, March. Predicting student's academic performance using an improved random forest classifier. In 2020 **international conference on emerging smart computing and Informatics (ESCI) (pp. 238-243). IEEE**

19. Kiu, C.C., 2018, October. Data mining analysis on student's academic performance through exploration of student's background and social activities. In 2018 **Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA) (pp. 1-5). IEEE**

20. Chen, S., & Ding, Y. (2023). A Machine Learning Approach to **Predicting Academic Performance in Pennsylvania's Schools. Social Sciences, 12(3), 118**

21. Pallathadka, H., Wenda, A., Ramirez-Asís, E., Asís-López, M., Flores-Albornoz, J., & Phasinam, K. (2023). **Classification and prediction of student**

performance data using various machine learning algorithms. Materials today. proceedings, 80, 3782-3785

22. Xu, K., & Sun, Z. (2023). Predicting academic performance associated with physical fitness of primary school students **using machine learning methods. Complementary Therapies in Clinical Practice, 101736.**

23. Holicza, B., & Kiss, A. (2023). Predicting and Comparing Students' Online and Offline **Academic Performance Using Machine Learning Algorithms. Behavioral Sciences, 13(4), 289.**

24. Sarwat, S., Ullah, N., Sadiq, S., Saleem, R., Umer, M., Eshmawi, A. A., ... & Ashraf, I. (2022). **Predicting Students' Academic Performance with Conditional Generative Adversarial Network and Deep SVM. Sensors, 22(13), 4834.**

25. Nabil, Aya, Mohammed Seyam, and Ahmed Abou-Elfetouh. "Prediction of students' academic performance based on courses' grades using deep neural networks**." IEEE Access 9 (2021). 140731-140746.**

26. Gajwani, J., & Chakraborty, P. (2021). Students' performance prediction using feature selection and supervised machine learning algorithms. In **International Conference on Innovative Computing and Communications. Proceedings of ICICC 2020, Volume 1 (pp. 347-354). Springer Singapore.**

27. R. Ghorbani and R. Ghousi, "Comparing different resampling methods in predicting students' performance using machine learning techniques," **IEEE Access, vol. 8, pp. 67 899–67 911, 2020.**

28. H. Waheed, S.-U. Hassan, N. R. Aljohani, J. Hardman, S. Alelyani, and R. Nawaz, "Predicting the academic performance of students from vile big data using deep learning models**," Computers in Human Behavior, vol. 104, p. 106189, 2020.**

29. S. Hashim, W. A. Awadh, and A. K. Hamoud, "Student performance prediction model based on supervised machine learning algorithms," in **IOP Conference Series. Materials Science and Engineering, vol. 928, no. 3. IOP Publishing, 2020, p. 032019.**

30. Chui, K. T., Liu, R. W., Zhao, M., & De Pablos, P. O. (2020). Predicting students' performance with school and family tutoring using generative adversarial network-based deep support vector machine. **IEEE Access, 8, 86745-86752.**

31. H. Altabrawee, O. A. J. Ali, and S. Q. Ajmi, "Predicting students' performance using machine learning techniques," **JOURNAL OF the UNIVERSITY OF BABYLON for Pure and applied sciences, vol. 27, no. 1, pp. 194– 205, 2019.**

32. H. Al-Shehri, A. Al-Qarni, L. Al-Saati, A. Batoaq, H. Badukhen, S. Alrashed, J. Alhiyafi, and S. O. Olatunji, “Student performance prediction using support vector machine and k-nearest neighbor,” in 2017 **IEEE 30th Canadian conference on electrical and computer engineering (CCECE). IEEE, 2017, pp. 14.**

33. Nkasu, M. M., "Investigation of the Effects of Critical Success Factors on Enterprise Resource Planning (ERP) Systems **Implementation in the United Arab Emirates", Smart Innovation, Systems and Technologies, 611–623 (2020).**

34. Jordan, M. I. and Mitchell, T. M., **"Machine learning. Trends, perspectives, and prospects", 349 (6245). (2015).**

35. Peng, J., Jury, E. C., Dönnes, P., & Ciurtin, C. (2021). **Machine learning techniques for personalized medicine approaches in immune-mediated chronic inflammatory diseases. applications and challenges. Frontiers in pharmacology, 12, 720694.**

36. "Types of Machine Learning | MLK - Machine Learning Knowledge", **https.//machinelearningknowledge.ai/types-of-machine-learning/ (2021).**

37. Ak, M. F. (2020, April). A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications. In **Healthcare** (Vol. 8, No. 2, p. 111). MDPI.

38. Edition, S., "**Data Mining and Knowledge Discovery Handbook",**

39. Kingsford and S. L. Salzberg, “What are decision trees?” **Nature Biotechnology,** vol. 26, no. 9, pp. 1011–1013, 2008.

40. Chiu, M. H., Yu, Y. R., Liaw, H. L., & Chun-Hao, L. (2016). The use of facial micro-expression state and Tree-Forest Model for predicting conceptual-conflict based conceptual change. **Chapter Title & Authors, 2016.**

41. "1.10. Decision Trees — Scikit-Learn 0.24.2 Documentation", **https.//scikit-learn.org/stable/modules/tree.html#mathematical-formulation (2021).**

42. Meena, K., Tayal, D. K., Gupta, V., and Fatima, A., "Using classification techniques for statistical analysis of Anemia", Artificial Intelligence In **Medicine,** 94 (February 2018). 138–152 (2019).

43. Cortes, C. and Vapnik, V., **"Support-Vector Networks", 297. 273–297 (1995).**

44. Yang, Y**., "THE RESEARCH OF THE FAST SVM CLASSIFIER METHOD"**, (1). 121–124

45. J. Smola and B. Sch¨olkopf, “A tutorial on support vector regression,” **Statistics and Computing,** vol. 14, pp. 199–222, 2004.

46. García-Gonzalo, E., Fernández-Muñiz, Z., Garcia Nieto, P. J., Bernardo Sánchez, A., & Menéndez Fernández, M. (2016). Hard-rock stability analysis for span design in entry-type excavations with learning classifiers**. Materials, 9(7), 531.**

47. T. K. Ho, "Random decision forests," in **Proceedings of 3rd international conference on document analysis and recognition, vol. 1. IEEE, 1995, pp. 278–282**

48. Wang, Y., Pan, Z., Zheng, J., Qian, L., & Li, M. (2019). A hybrid ensemble method for pulsar candidate classification. **Astrophysics and Space Science, 364, 1-13.**

49. Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J., **"Classification and Regression Trees", (1984).**

50. Boulesteix, A., Janitza, S., and Kruppa, J., "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics**", 2 (December). 493–507 (2012).**

51. Taylor, P., Altman, N. S., and Altman, N. S., "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression**",** (December 2014). 37–41 (2012).

52. Nechuta, S.J.; Caan, B.; Chen, W.Y. The after breast cancer pooling project. Rationale, methodology, and breast cancer survivor characteristics. Cancer Causes Control 2011, 22, 1319–1331

53. P. Cortez and A. M. G. Silva, **"Using data mining to predict secondary school student performance," 2008**

54. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," **Communications of the ACM,** vol. 63, no. 11, pp. 139–144, 2020.

55. Yang, Y. Shen, Y. Xu, and B. Zhou, "Data-efficient instance gen- eration from instance discrimination**," Advances in Neural Information Processing Systems,** vol. 34, pp. 9378–9390, 2021.

RESUME

Aws Mohammed KHUDHUR, he completed high school education at (ALSALAM) high school in Salah al din /Iraq. then, He obtained a bachelor's degree from Tikrit University / Computer Science. To complete their M.Sc., he moved to Karabuk/Turkey in 2021. He started his master's education at the department of computer engineering in Karabuk University.