Arab American University

Faculty of Graduate Studies

Student Majors Prediction and Classification Using Machine Learning Techniques

Prepared By

Jihad Mousa Amin Shekh Mousa

Supervisor By

Prof. Mohammed Awad

This thesis was submitted in Partial Fulfillment of the Requirements for theMaster's Degree in Computer Science

February / 2022

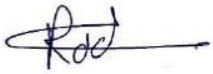
Thesis Approval

Student Majors Prediction and Classification Using Machine Learning Techniques

By

Jihad Mousa Amin Shekh Musa

This thesis was defended successfully on **14 /02/2022** and approved by:

Committee Members	Signature
1. Supervisor: **Prof. Dr. Mohammed Awad**	
2. Internal Examiner: **Dr. Radwan Qasrawi**	
3. External Examiner: **Dr. Ahmad Ewais**	

Declaration

I acknowledge that this thesis entitled "Student Majors Prediction and Classification Using Machine Learning Techniques" is the result of my research and work, and I write it alone. It is devoid of any other research work except for the reference, and it has not been submitted to any other scientific or any scientific degree.

Name: Jihad Mousa Sheikh Mousa

Signature: 

Date: 14. 2. 2022

Dedication

I dedicate this thesis to: My father a strong and gentle soul who taught me to believe in hard work and that the best kind of knowledge to have is that which is learned for its own sake.

My mother whose affection, love, encouragement and prays of day and night make me able to get such success and honor.

My brothers and sisters for being my guardians during my educational career.

My wife and sons for their continuous support and understanding when undertaking my research and writing my thesis.

Your prayer for me was what sustained me this far.

My eternal love my homeland_ Palestine

Acknowledgment

I would like to acknowledge and give my warmest gratitude to my supervisor (Muhammad Awad), who made this work possible. His guidance and advice carried me through all the stages of writing my thesis. A debt of gratitude is also owed to my university AAUP and my respected teachers for their invaluable help in preparing this research.

I would also like to give special thanks to my parents, my wife, my children, my brothers, and my family as a whole for their continuous support and understanding when undertaking my research and writing my thesis. Your prayer for me was what sustained me this far.

I humbly extend my thanks to all concerned persons who cooperated with me in this regard.

Finally, I would like to thank God for letting me through all the difficulties. I have experienced your guidance day by day. You are the one who let me finish my degree, and I will keep on trusting you for my future.

Abstract

Student Majors Prediction and Classification Using Machine Learning Techniques

Prepared By: Jihad Mousa Amin Shekh Musa

Supervisor By: Prof. Mohammed Awad

Predicting students' majors is very important in developing the performance of the Palestinian Ministry of Education, initial recovery plans for weak students, determining the future needs of the Ministry of Education system, contributing to building structures for the teachers' organization within the Ministry of Education and many others, all of the above mentioned are the result of early anticipation of the students' different specializations.

This study represents the first of its kind in Palestine. This thesis is concerned with determining the students' future specializations in the scientific, literary, industrial and commercial branches. The classification of majors is based on machine learning techniques and their classification ability. Identifying the essential features and the size of the dataset collected is very important to enhance classification accuracy and classification matrices.

So, Finding and examining the best machine learning algorithms and their ability to classify as accurately as possible and at reach, high rates in different classification matrices depending on the selected feature and the collected data are the core of this thesis. The first step to achieving classification as accurately as possible is features selection. Specialists in the field of education determine these features. Those features included required information about the student, required information about the

student's family, and other information about the student's community, in addition to his academic achievement in the tenth grade. This process is followed by creating a questionnaire judged by professionals, and then it is distributed to the various directorates of education after obtaining prior permission to do so. The collected dataset includes more than 1200 students in different branches; this dataset is subjected to the pre-processing process phase by cleaning, scaling, and encoding it to be ready for machine learning techniques.

Sex classification learner algorithms are used, in addition to neural networks algorithms. The learner algorithms used are: k-nearest neighbors' algorithm (KNN), Support vector machines (SVM), linear discriminant, decision tree, ensembled boosted tree, and ensembled bagged tree, while the neural network used is multiple-layer perceptron neural network (MLPNNs).

Experiments show that neural networks significantly outperformed the rest of the other Machine learning techniques, as the classification accuracy reached 96.9%, Whereas the classification accuracy in ensembled algorithm id 93.8%, which is the best within the category of classification learner algorithms. The experiments also show poor accuracy when using marks in the curriculum alone for classification on all algorithm uses. This comparison shows the need to include other features that have been studied, the most important of these features are the student's and father's trends and the availability of specialization.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

KNN	K-Nearest-Neighbor
ANNs	Artificial Neural Networks
NNs	Neural Networks
SVM	Support Vector Machine
MLP	Multi-Layer Perceptron
TP	True Positive
FP	False-Positive
FN	False-Negative
TN	True Negative
MOE	Ministry of Education
ML	Machine Learning
MLT	Machine Learning Techniques
MFFNNs	Multilayer Feed Forward Neural Networks
BP	Back Propagation
ANNs	Artificial Neural Networks
LM	Levenberg-Marquardt
ES	Ensemble
MSE	Mean Squared Error

Chapter One

Introduction

1.1 Introduction

Generally, improving the quality of education is a significant research problem. Many methods have been proposed to increase the quality of Palestinian learning and teaching. Researchers have focused on improving educational performance [1], especially in the advanced stages. Predicting students' future majors has a significant role in improving many factors, and it is undoubtedly one of the areas that strongly influence the quality of education. It is worth noting that predicting student majors is a complex matter, as many factors affecting the selection of a major underline it. Therefore, predicting student majors and classifying them in universities or schools is more difficult because of the many features that can be used to determine the appropriate major for students [2][3].

The Palestinian secondary education system includes several branches, and these branches are identified in the tenth grade. These branches are the scientific branch, the literary branch, the commercial branch, and the industrial branch. The scientific branch allows the student to join most specializations, especially the scientific ones. In most cases, students' degrees in this branch are high in scientific subjects such as mathematics, chemistry, and others. The literary branch is joined by students who wish to study literary branch and also allows them to study commercial branch, and among these disciplines, for example, the specializations of law, commerce, and art. Also,

educational courses such as history, geography, and others, When the student decided that his principal would be commercial only, he would join the commercial branch specialization so that he would be able to study the majors that support commercial sciences in the future. As for students who wish to study specific industrial disciplines such as engineering, as well as those who do not possess the scientific competence necessary to obtain high marks in the scientific branch, they can benefit from the industrial branch through which they can form a basic industrial knowledge based on the industrial curriculum.

It is worth noting that the selection of majors for students is not necessarily based on their future inclinations. Some students enroll in one of the branches to increase academic achievement, some for friendship and kinship, and others for the unavailability of the branch in their area and other reasons. These reasons play an essential and influential role in students' enrollment in the different majors [2]. Identifying these reasons must be based on previous information and the opinions of specialists in the educational field [4].

Machine learning techniques have been used in educational issues, many machine learning techniques have been applied to enhance educational performance. The newest study, which concerns enhanced education, focused on machine learning techniques for prediction and classification because it has an excellent ability to classify and predict's performance with high accuracy and the least possible square error. Different domains of education benefit from machine learning techniques, such as expecting the student's graduation, courses marks, university average marks, the year of graduation, etc[4][5].

Various machine learning techniques have been used to classify academic performance, as they have had an influential role in measuring students' performance at university levels [5]. Machine learning techniques depend on datasets about students, their academic achievement, and other important information. It is found that MOE databases in schools in Palestine contain millions of records of data on thousands of students over the past years.

Institutions can adopt machine learning techniques instead of traditional assessment mechanisms to enhance classification and forecasting; This helps decision-makers develop an appropriate strategy and policy in the educational fields [6]. Decision-makers face annual challenges related to the expecting number of students in different disciplines, determining the capacity of schools in higher disciplines as well as universities, which the popular traditional methods did not contribute to the development of accurately and effectively long-term future visions, so the need to use modern forecasting mechanisms has become necessary.

This study aims to anticipate and classify a student's major based on academic factors and surrounding circumstances. It will help the decision-makers set the best plans for education. In simple words, the intelligent system can help decision-makers get detailed statistics on the number of students, their performance, and the majors expected. As well as, it is contributing to the development of annual plans capable of adapting to developments on the one hand and improving educational performance on the other hand.

Many factors affect a student's enrollment in a particular major. Determining these features requires research, and the opinions of specialists in the field of education,

especially since defining the student's majors is the first research of its kind in Palestine and perhaps rarely applied in the world due to the inability to find previous studies on the same subject specifically. Choosing a major depends mainly on the marks of the curriculums, and some curricula are essential for determining the major. The student's society, surrounding conditions, and competition among his relatives are vital factors in changing his future major. So, the dataset must include exclusive features on different fields that can affect the student's decision. Briefly, the student's major is predicted according to their academic history based on previous educational attainment and the student environment.

Related works show the ability of machine learning techniques in classification. These techniques vary in their ability to classify based on the features determined [5][6]. It is noticed that machine learning techniques give more accurate classification results than traditional methods [6]. The accuracy of machine learning techniques is affected by the chosen features, the number of samples used, and the pre-processing mechanisms. From this point, specialists are hired to determine the features that affect the selection of majors, sufficient samples, and choosing appropriate preprocessing mechanism [6].

This work use curriculum marks in the tenth grade as features, in addition to this student’s environment and personality. These features are formed in a questionnaire and distributed electronically to Ministry of Education directorates from the north to the south. The dataset collected from those directorates reaches more than 1,200 records about students from different majors. These data are treated with the necessary techniques to be ready for classification. Collecting data from the Palestinian educational community is not an easy process, especially after the spread of the Corona

epidemic, which posed the most significant challenges that impeded obtaining official approval for data collection and the ease of collecting it.

Our thesis uses machine learning techniques to classify students' majors. Many machine learning techniques are applied. These techniques are decision trees, ensembled algorithms, KNN, SVM, multi-layer perceptron neural networks, and others, so the capabilities of Different machine learning techniques for determining the majors of students in different branches are examined, as well as the most excellent accuracy that can be obtained after applying those techniques is determined. All classification algorithms supported by the MATLAB program are tested, and the most accurate classifications are selected according to the group to which it belongs. A neural network implemented by MATLAB software is also used for classification.

Furthermore, a comparison between different machine learning is performed. The comparison shows which machine learning techniques can classify and give better results. For more tests, a dataset containing the students' marks is involved in classification. The best two algorithms in terms of accuracy and classification matrices are chosen. This test aims to check the dataset of student marks to classify student majors. These techniques have been applied to demonstrate the importance and effectiveness of the features that have been studied

1.2 Objectives

This study aims to establish an machine leering system capable of predicting educational outcomes related to students' majors through determining the essential features and collecting required information about students, then processing the

collected data by applying optimization tools such as removing outliers, cleaning data, preventing duplications, etc., which would provide more accurate and safe future predictions and identify the most critical factors that affect the outcome of the educational process and students majors—by applying different classification techniques, then designing and building the machine learning.

The general summarizes the objectives in the following points:

- Determining the most important features that affect a student decides to choose his future major.
- Examining the strength of the machine learning technology used and which one is the best for educational data.
- Establishing the foundation stone for changing the traditional methods used in the Ministry of Education to modern methods.
- Showing the ability of machine learning techniques to predict the future, they could move forward in enhancing the country's educational level and developing plans for future adaptation.
- Helping the students determine the future majors that fulfill their dreams and preferences. Moreover, giving them the essential instructions to impact and improve their educational behavior in a way that supports their goals.
- Developing a general perception of the academic situation for the governments, colleges, and universities.

1.3 Contribution

This study is built on the data of the educational institution of the Palestinian Ministry of Education. Through the Ministry of Education system and relying on professionals in this field, the most important feature that affects the student to determine his future majors is determined. Professionals in statistics can use these features to get high accuracy on future works that depend on classification. The best classification machine learning algorithms are chosen, and the best ones are determined. This is performed after using different machine learning algorithms in MATLAB software. MLP neural networks are also used to perform the classifications to compare the ability of the neural networks in the classification and their superiority over the machine learning classification learner algorithms. Forecasting students' majors help decision-makers in the Ministry of Education make appropriate decisions for the infrastructure needs and evaluate students' academic performance. It is worth noting that predicting the future specialization using machine learning is the first research of its kind in Palestine, and the specialized research sites in research prove this.

1.4 Overview

The rest of this thesis is arranged to start from the second chapter to the fifth chapter and place the appendices supporting it. Chapter 2 provides a background on the work of the study for the Palestinian educational field and mentions the literature review for previous participations relevant to the educational field. The third chapter explains the data collection, selection, processing, and normalization mechanism. And then an explanation of the techniques, the most general techniques of machine learning

specialized in the field of classification, which will be used, as well as methods for measuring the accuracy of classifications for the categories of multiple classifications. In the fourth chapter, the results of the algorithms that were used, presenting their results, analyzing them, and comparing the algorithms among themselves to show the accuracy of each algorithm besides the points of the superiority of each algorithm over another algorithm. In study procedures: official correspondence questionnaire forms were used.

Chapter Two

Background

2.1 Introduction

The Palestinian high school exam represents the defining stage, the end of school life, and the beginning of university life. University acceptance rates are affected by students' grades and numbers in each branch separately. From this point of view, universities should know the number of students in each batch of branches to prepare their infrastructure to accommodate them. The prior knowledge of their numbers in each specialization also contributes to building an educational system to adapt to the increasing numbers in the Ministry of Education

Based on the Ministry of Education statistics shown in figure 2.1 about this subject, it is found that the highest success rate is among students in the scientific branch and the lowest in the literary branch.

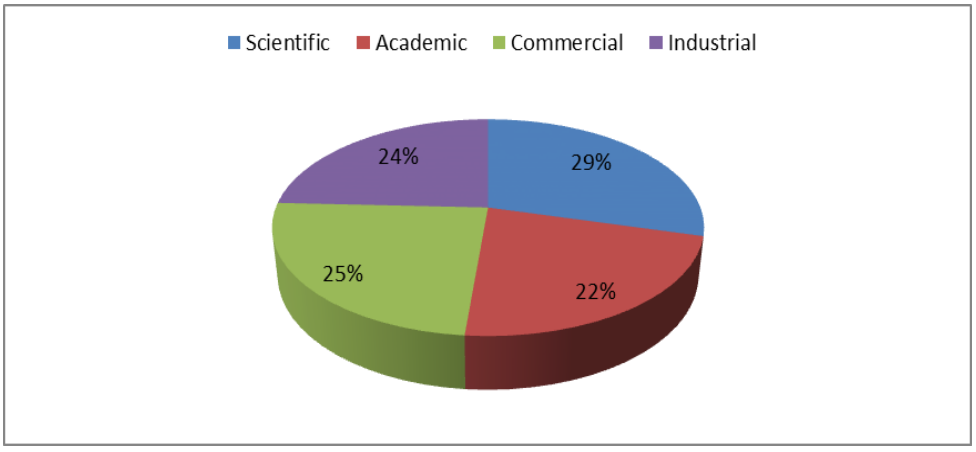


Figure 2.1 The General Palestinian secondary exam 2021 chart

The number of students enrolled in each branch has been studied for a more authentic look. The required data is gotten from statistics of the Palestinian General Secondary Certificate for the year 2019, where the literary branch has the most extraordinary luck in the number of students, followed by the scientific branch, followed in relative proportions by the commercial and industrial branches, this is illustrated in figure 2.2.

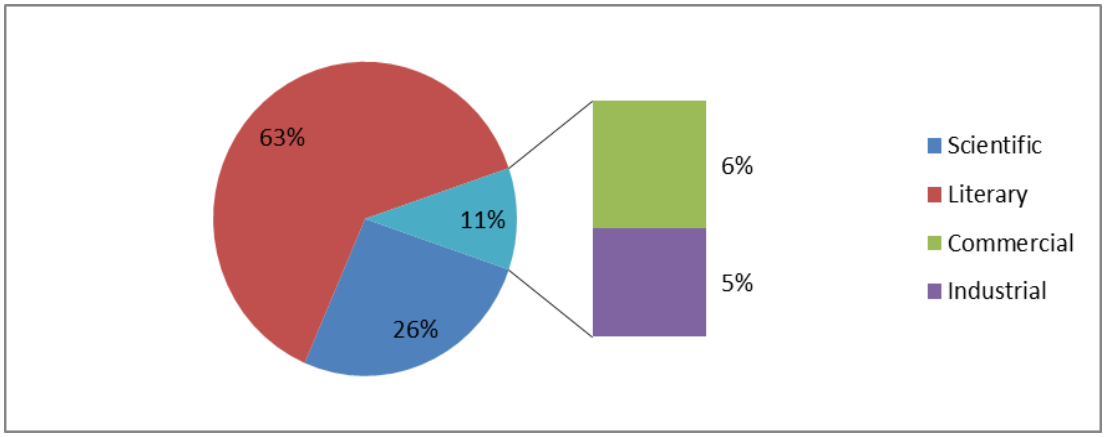


Figure 2.2 The number of students enrolled in each branch chart

It is also noted that after comparing the success rates for four years against the other, it is found a large discrepancy between the year 2018 and the year 2021, and this is shown in figure 2.3

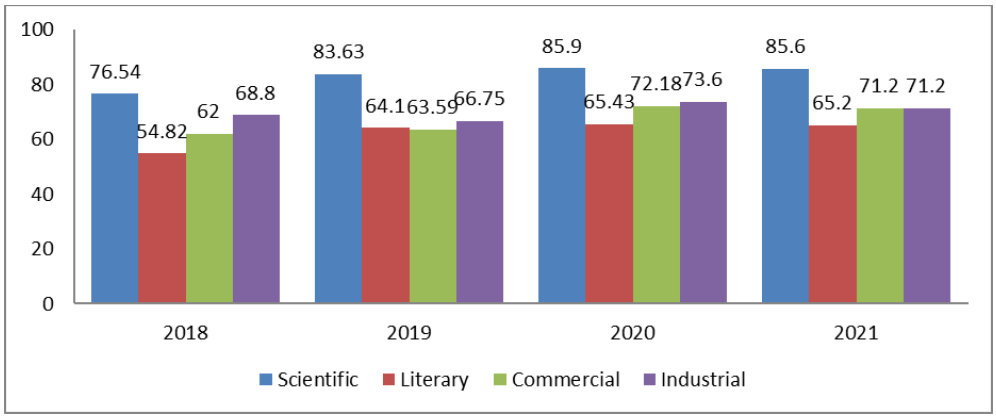


Figure 2.3 The general Palestinian secondary exam comparison chart

Predicting the number of students in each branch is necessary because it significantly impacts decision-makers. As it is mentioned earlier, success rates vary from year to year. Accordingly, the Palestinian educational environment should be carefully studied to obtain accurate classifications of the exit expectation from the high school exam. Here, the need to adopt modern mechanisms appeared in the prediction that machine learning techniques are recently adopted in many studies and show their ability to classify educational data with high ability. This encourages performing this study to classify students' specialization in Palestine.

This thesis deals with data on the education system in Palestine. It has not been conducted in advance; so many difficulties are determining the features that affect the student choosing his major, with the help of educational specialists in this field. These most critical features that affect the student in determining his specialization have been identified, and those features are placed in an electronic questionnaire. After obtaining approval, the questionnaire is published in several education directorates, see appendix A. It is published as follow: three directorates are selected in the north of Palestine, which are Jenin, Qabatiya, and Tubas, and three in the south, namely Hebron, North Hebron, Bethlehem, and the center, represented in the Directorate of Ramallah and Al-Bireh and the suburbs of Jerusalem. In this section, the features adopted from specialists have been mentioned and discussed throw questionnaire statistics. Also, the related work which is closed to our studies has been discussed in this chapter.

2.2 Data Description

In the classification, the supervised learning approach is the one that depends. A questionnaire containing specific questions and answers for each student with the

current majors were created. The questionnaire was established with the assistance of educational experts. The questionnaire includes the students' academic marks in the tenth grade with other general information that affects the student's decision to choose his\her future major. The number of curricula taught by the student in the tenth grade is eleven; in addition to that, ten other classifications were taken related to the student's life, orientation, work of his\her relatives, and the field of their work, as well as the classifications that were carefully selected to obtain an effective classification in this field.

The questionnaire was adopted and audited by specialists at the Arab American University and specialist educators outside the university. The Arabic language was adopted in the questionnaire. It was linguistically checked by specialists so that it would be easier for the student to answer the questions of the questionnaire quickly and conveniently. It is worth noting that some of the added features were omitted from the questionnaire to preserve the privacy of the student, which are related to family problems, health, and financial matters. Multiple-choice question forms is used for ease classification and the Ministry's platform were used to obtain correct academic information.

The questionnaire was designed on four main parts, see appendix B, which are first information related to the student himself, second, information related to the student's family. Third, information related to the student community, Fourth, information related to the student's academic achievement in the tenth grade [2].

The following explains each of these parts with their features and discusses the general statistics resulting from each feature alone.

In the first part, the focus was on particular information about the student himself, demographically and behaviorally as well, as specialists approved these features, and some features were similar in some previous studies, which are as follows

1- Academic level: This feature is suggested by experts, as some other features are related to this feature. This feature requires two answers, and he is either in the eleventh grade or in the twelfth grade. The voting rate of eleventh students is much higher than that of twelfth students, as shown in Figure 2.4.

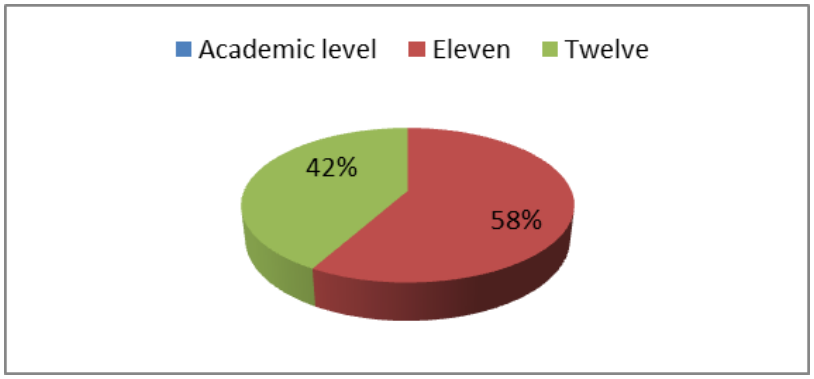


Figure 2.4 Academic level chart

2- Gender :This feature to determine voter gender, either male or female. Figure 2.5 shows that females are the most voted on the questionnaire.

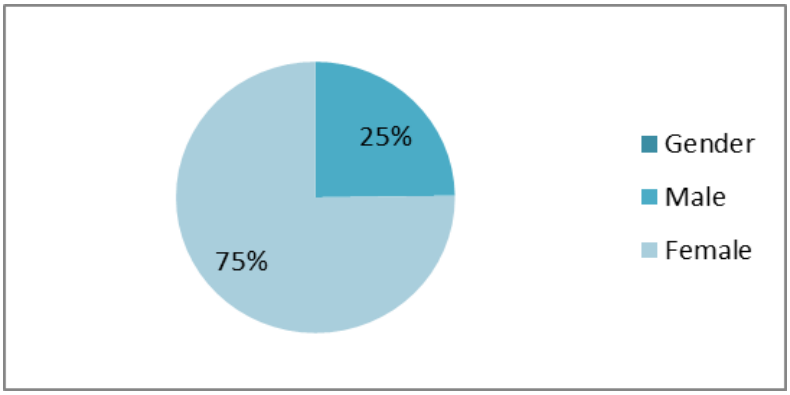


Figure 2.5 Gender chart

3- Current major: This feature will represent the output of the different machine learning algorithms, and this option has four possibilities, either scientific, literary, commercial, or industrial. As shown in Figure 2.6, the questionnaire's statistics showed that most students are enrolled in the scientific and literary branches, which is typical compared to their numbers.

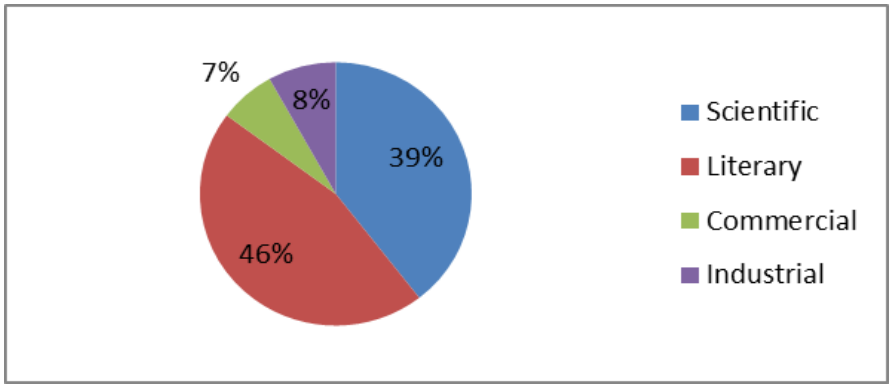


Figure 2.6 Current major chart

4- The student tends to specialize: This feature tests the president's inclinations to enroll in one of the branches. It has four options: scientific, literary, commercial, and industrial. Inclinations to the scientific branch are the most popular among students, and this is what the statistic shows in Figure 2.7.

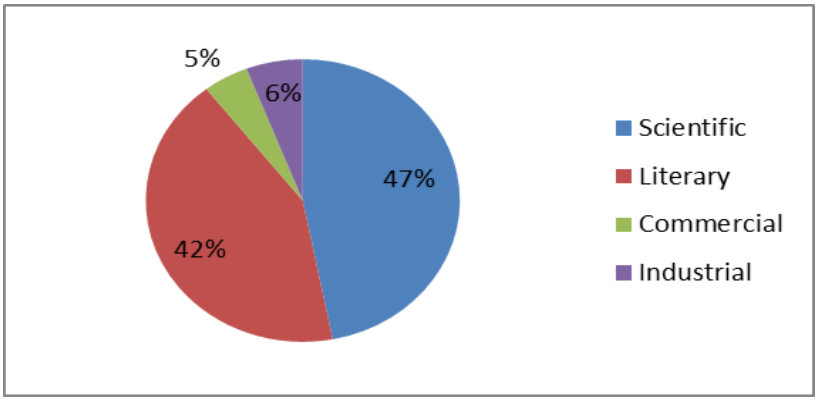


Figure 2.7 The student tends to specialize chart

5- In terms of social relations, the student is classified as: This feature tests the student's social relationships. The options were categorized into four sections, starting with effective social relationships and ending with a limited relationship. The questionnaire showed that almost half of the students have normal social relations. Note the Figure 2.8.

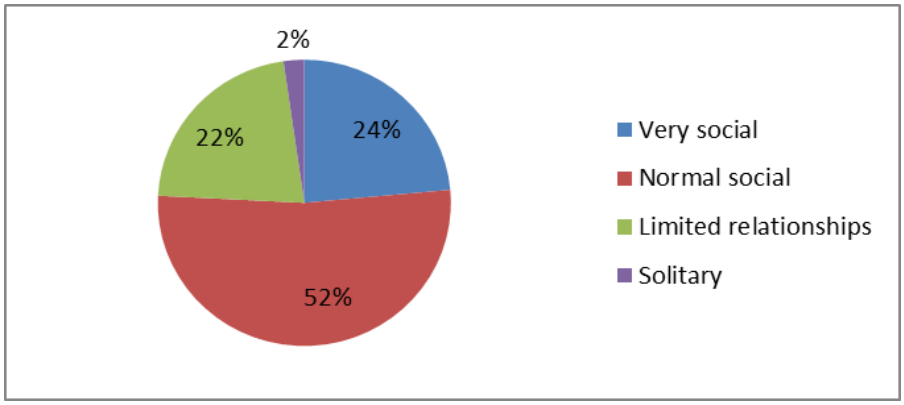


Figure 2.8 In terms of social relations, the student is classified as a chart

6- The student is full time to study: This feature show whether the student has enough time to study, the classification was divided into four groups, starting from one hour and ending with six hours and more, the questionnaire reveals that 60% of students have enough time to study, and this is shown in the following figure 2. 9.

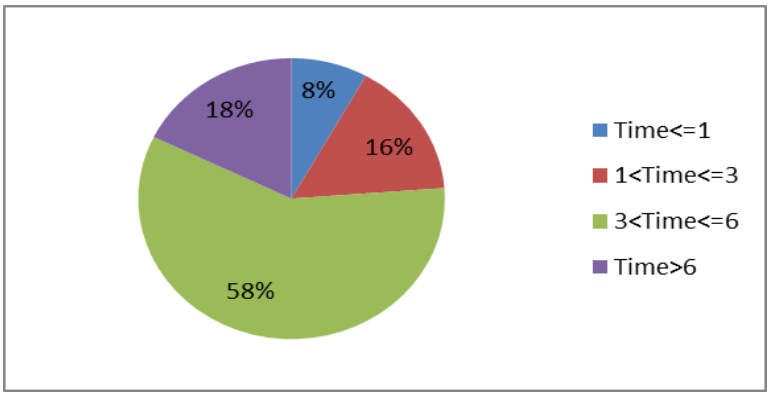


Figure 2.9 The student is entirely full time to study chart

7- The number of daily study hours: This feature shows the actual number of hours the student studies. The classification was divided into four groups, starting from one hour and ending with six hours and more. The questionnaire reveals that 36% of students study between one to three hours per day, shown in Figure 2.10.

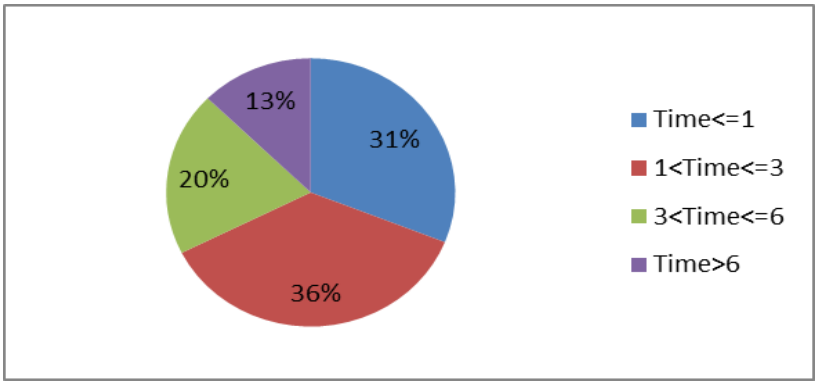


Figure 2.10 The number of daily study hours chart

8- The number of hours spent online and mobile daily :This feature shows the number of hours a student spends on mobile. The classification was divided into four groups, starting from one hour and ending with six hours and more. Most of the students voted that they use the Internet and the phone for a period ranging from one to three hours. Notice Figure 2.11.

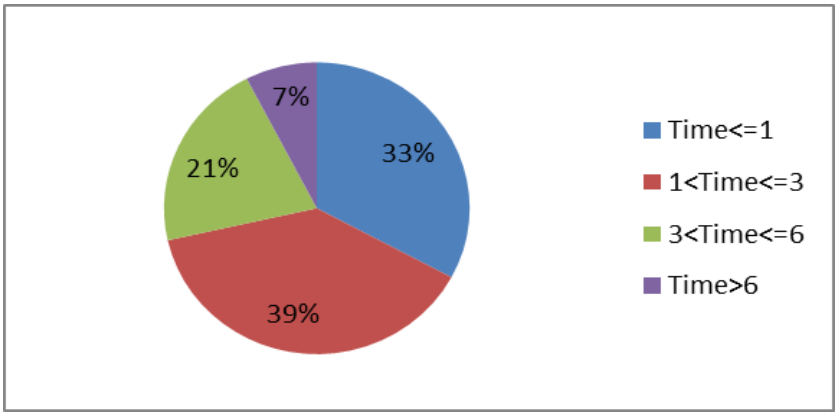


Figure 2.11 The number of hours spent online and mobile daily chart

9- The number of hours spent on playing and physical exercise per day: This feature measures a student's number of hours on physical exertion. The classification was divided into four groups, starting from one hour and ending with six hours and more. The questionnaire revealed that our students do not exercise, evident in Figure 2.12.

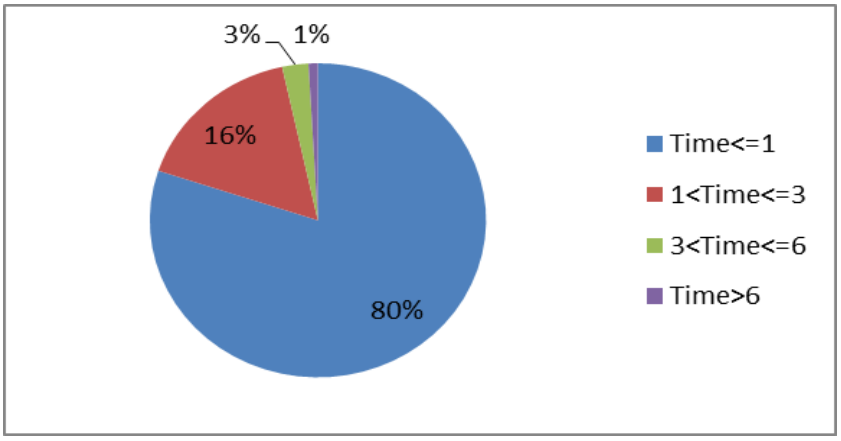


Figure 2.12 Number of hours spend on playing and physical exercise per day chart

And also include information about the family of the claimant within the approved features As the specialists admitted that there is also an influence for the student's family in determining their future major

10- Father works: This feature tests whether the father is working or not. Where experts believe that the father's work is reflected in the student's choice, it may, directly and indirectly, affect the choice. The questionnaire shows that most fathers work. See figure 2. 13.

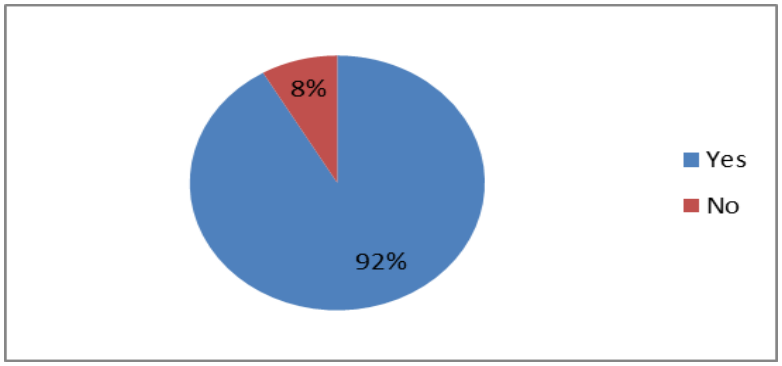


Figure 2.13 Father works chart

11- Mother works: This feature tests whether the mother is working or not, Where experts believe that the mother's work is reflected in the choice of the student and the hours she frees up to educate her children, Most of the mothers are unemployed, as the questionnaire shows note the figure 2.14.

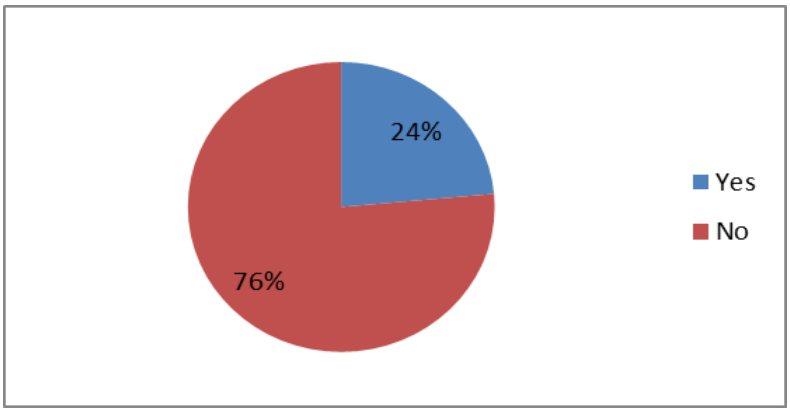


Figure 2.14 Mother works chart

12- Father specialization: This feature tests the father's specialty within the displayed academic branches. Some students are greatly affected by their father's work, and they like to join a specific specialty, following their father's example. The father's specialization was in line with the number of students in the literary branch, as most fathers enrolled in the literary branch. This is what the questionnaire revealed. See figure 2.15.

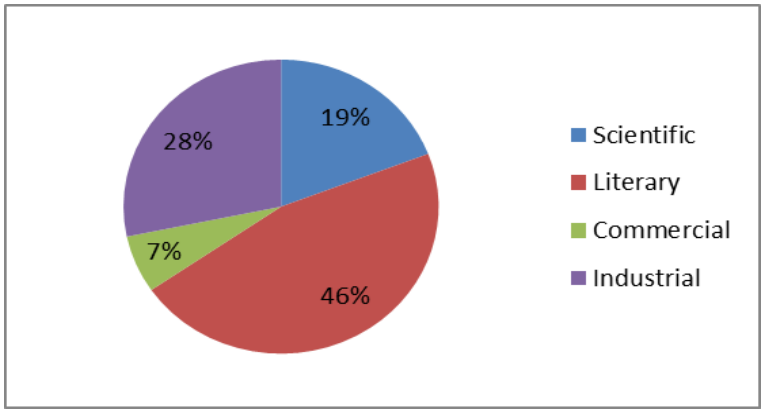


Figure 2.15 Father specialization chart

13- Mother specialization: This feature tests the mother's specialty within the offered academic branches. There are some families where the mother is the role model, so specialists believe this is reflected in their children's choice of major. The results were somewhat similar between the specialization of the father and the specialty of the mother as well, note figure 2.16.

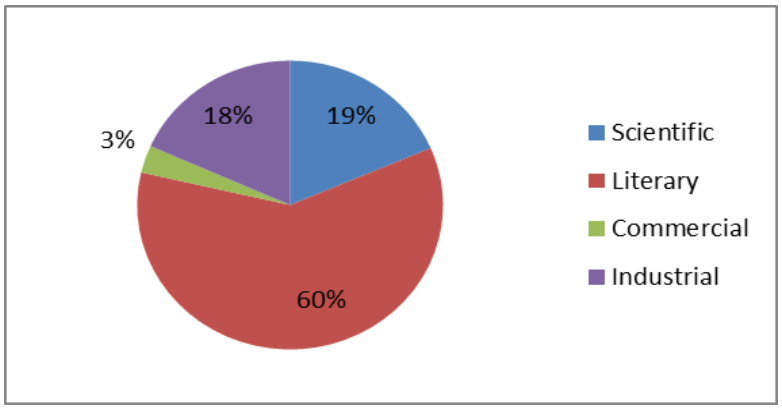


Figure 2.16 Mother specialization chart

14- Father tendency: This feature tests the father's tendencies for the branch he would like the student to join. If their father's opinion influences the student, their inclinations play a significant role in choosing a major. The questionnaire

shows that the father tends to the literary and commercial branches. Note figure 2. 17.

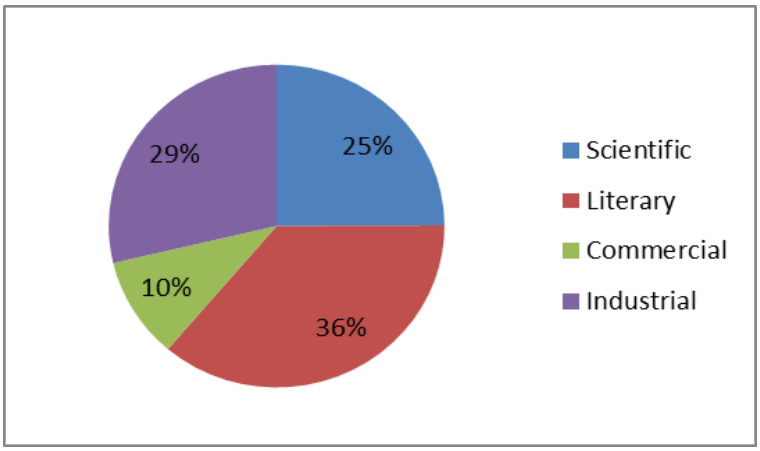


Figure 2.17 Father tendency chart

15- Mother tendency: This feature tests the mother's preferences for the branch you would like the student to join. The mother's tendencies may influence the student's selection for a major, especially if the mother is the role model for that student. The questionnaire shows the mothers' tendency to the literary branch significantly. Perhaps the reason is the fear that their children will not succeed, shown in figure 2.18.

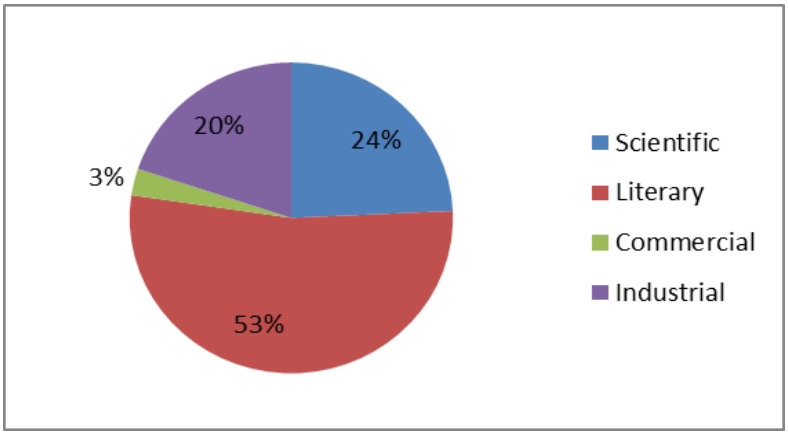


Figure 2.18 Mother tendency chart

16- Father is dedicated to his children: This feature tests if the father is only devoted to his children and their affairs. The questionnaire confirms that half of the fathers are dedicated to their children. Note Figure 2.19.

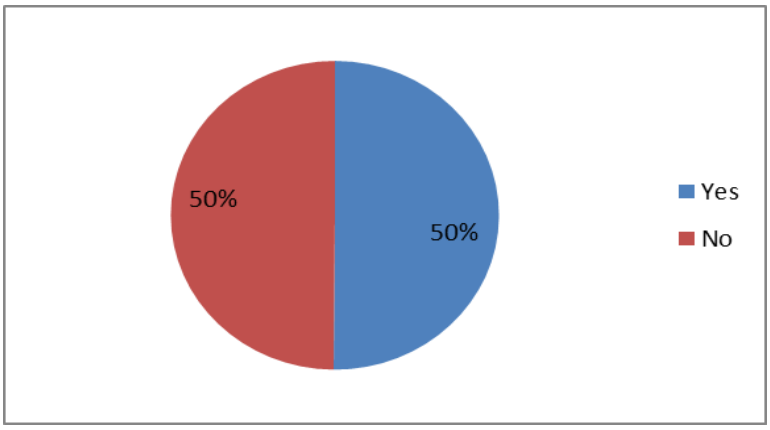


Figure 2.19 Father is dedicated to his children chart

17- Mother dedicated to her children: This feature tests whether the mother is free for her children and their affairs only. Most mothers are entirely free for their children, which appears from Figure 2.20, representing the statistic of a questionnaire in this feature.

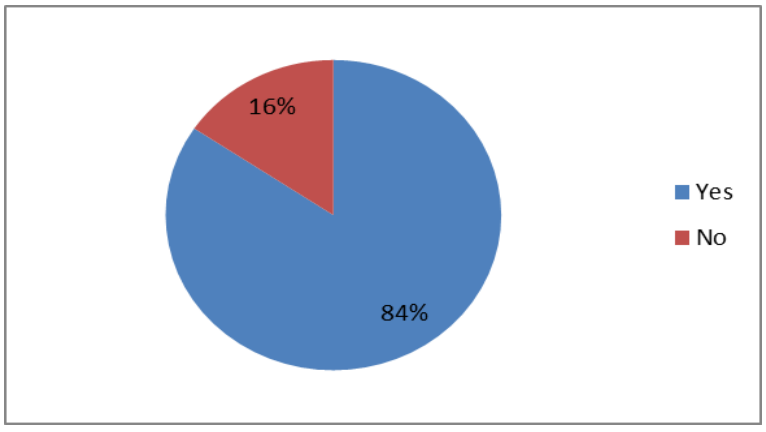


Figure 2.20 Mother dedicated to her children chart

18- Competition of relatives: This feature tests if there is a competition to join specialties between relatives, more than 40% of parents are influenced by the specialties of their relatives' children, Note Figure 2. 21

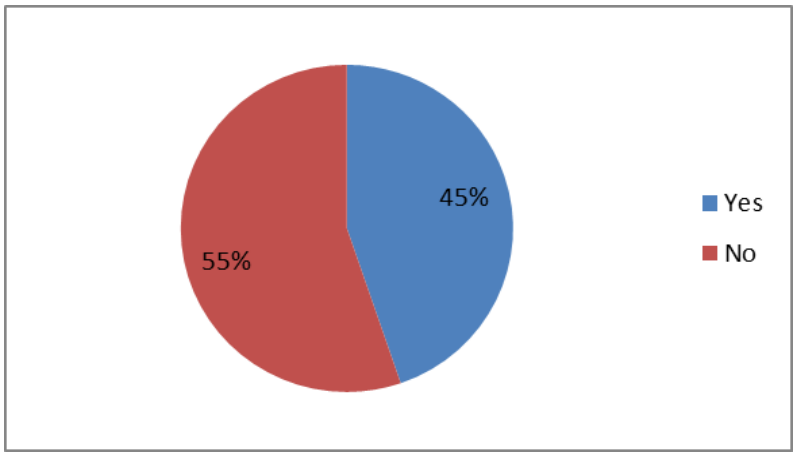


Figure 2.21 Competition of relatives chart

It also included general information about the student's community, as follows.

19- Availability of primary/specialization in his current place :This feature tests the branch's availability in the student's residence, and this feature impacts females more than males in some areas. Statistics show that 13% of students study outside their hometowns, as shown in Figure 2. 22.

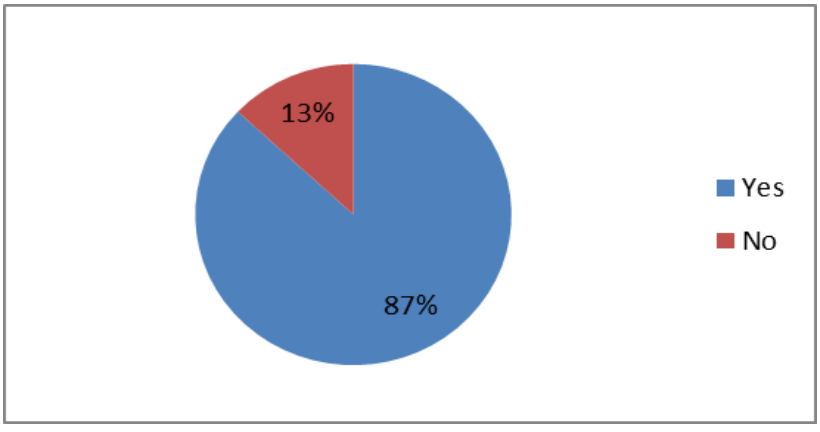


Figure 2.22 Availability of primary/specialization in his current place chart

20- Place of living :Here, the student's place is tested. Does he live in a city, village, or camp? About half of the students who voted live in a village. Note the Figure 2. 23.

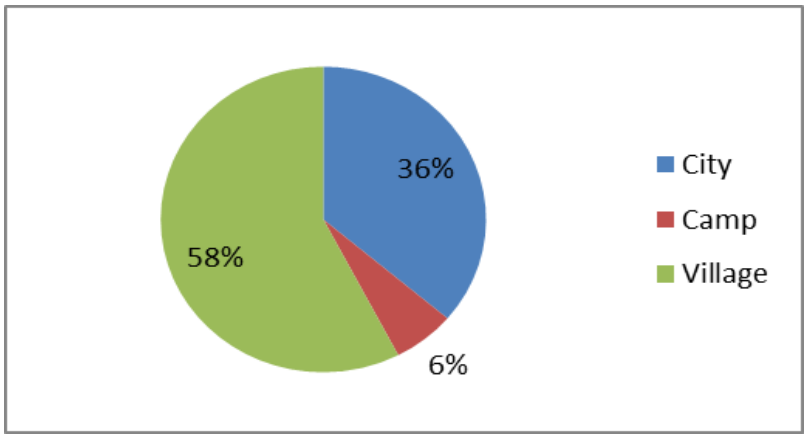


Figure 2.23 Place of living chart

21- Community orientation :The orientation of society, in general, is measured by this feature and whether society has industrial tendencies, for example. The statistics revealed that society's orientation agrees with the family's orientation in that the most popular specialization is the literary branch. Note the Figure 2. 24.

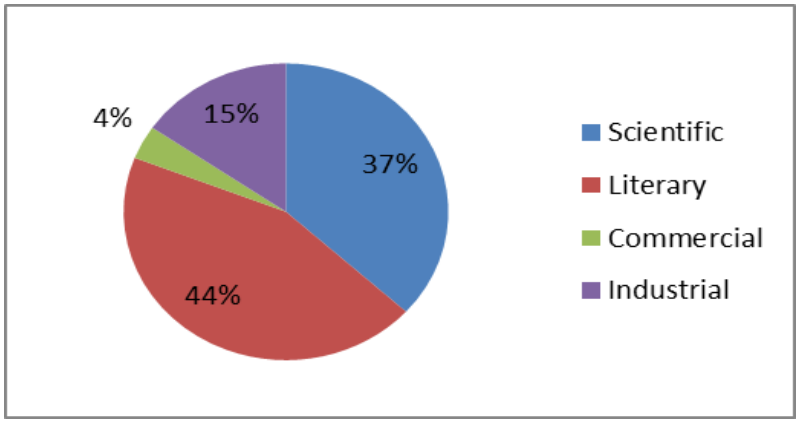


Figure 2.24 Community orientation chart

Finally, the questionnaire included the students' tenth-grade grades, as in the eleven curricula of the tenth-grade studies. The number of features used to determine the student's tendencies is 31 features and one output bearing four options. It was emphasized in the curriculum marks that had to be filled must be depending on the electronic platform of the Ministry of Education E-School, curriculum names were as follows: Islamic education, Arabic, English, Maths, Physics, Chemistry, Life Sciences, Palestine history and geography, arts and crafts, Physical education and technology education. The statistics related to the questionnaire revealed that the average marks of students in the scientific branch are higher than the rest of the branches. Note the Figure 2. 25.

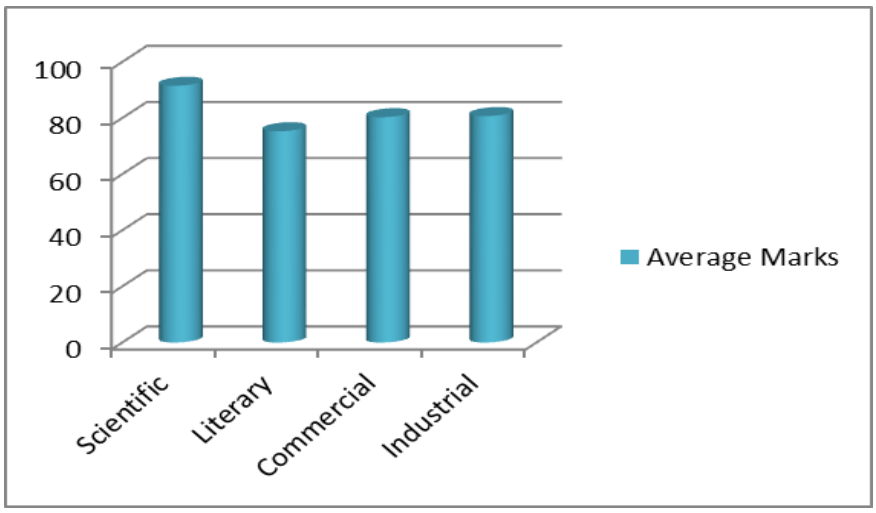


Figure 2.25 Average curriculums marks for each branch chart

2.3 Related Work

This study confirmed that the student's performance was one of the most important things that had to be worked on to increase the efficiency of education, and to improve the quality of education, according to the researcher's opinion, by identifying students

with low achievement and this was done by classifying the performance of students. The education system puts mechanisms to heal and move education towards the better. This study used two data sets to classify the low-performing students, the first containing 395 rows and the second containing 649 rows. This data set consisted of 33 tribunes, and the assembled decision tree was also used, which is one of the machine learning techniques to obtain more accurate classifications. The researcher used three types to support this algorithm (C5.0, AdaBoost.M1, and AdaBoost.SAMME). Two models were also used in classification, binary classification, and multiclass classification. Note the table.

Table 2. 1 Grade in binary classification

Classification Goals	Grade
Pass	10-20
Fail	1-9

Table 2. 2 Grade in 5-levels classification

Classification Goals	Grade
Very Good	16-20
Good	14-15
Satisfactory	12-13
Sufficient	10-11
Fail	1-9

The study was categorized into three scenarios: the first is an increase in floods, where the two algorithms AdaBoost. SAMME and AdaBoost.M1 outperformed C5.0 in classification, and the second is a modification to a different number of training data, in which AdaBoost.M1 excelled in the third. The third is using specific training parameters and testing them with other training parameters. The study found that training

parameters can now be used and tested by other training parameters. The algorithms vary in strength according to the situation and system to which they are applied [9].

In this study, the researcher wants to increase the students' efficiency in the SSC exam. Machine learning techniques are applied to predict the students' achievement in this exam. As the researcher sees, the prediction of achievement helps the continent's adopters in the educational system develop a treatment plan for students with low achievement. The dataset used in machine learning consisted of data for 400 students. This data consisted of 31 characteristics, which included the marks of the most critical seven academic subjects, as shown in Table 2.1. In addition to his name and identity, the rest of the columns are the scientific background on some educational subjects and topics. The student's GPA in the SSC exam was placed in the last column, and This data is a summary of its processing operations normalization. The encoder was used to prepare the data for machine learning algorithms.

Table 2. 3 Features description

Feature name	Description	Values	Data type
English	Marks of English in Exams	0-100	integer
Bangla	Marks of Bangla in Exams	0-100	integer
General Math	Marks of General Math in Exams	0-100	integer
Physics	Marks of Physics in Exams	0-100	integer
Chemistry	Marks of Chemistry in Exams	0-100	integer
Biology	Marks of Biology in Exams	0-100	integer
BDS	Marks of BDS in Exams	0-100	integer

Many machine learning algorithms have been used to compare between them. Also, to determine which of these algorithms can give more high accuracy than others, the machine learning algorithms used are: Naive Bayes, K-nearest Neighbors, Support

Vector Machine, XG-boost, Multi-layer Perceptron, One of the most critical classification factors that the researcher used to judge the ability of these algorithms in predicting student outcomes is F 1 score. It showed the superiority of neural networks over others in classifying the exam outcomes. And the weaknesses of XGBoost in classifying the exam outcomes compared to other algorithms see Table 2.4 [10].

Table 2. 4 Performance evaluation for each classifier

Classifier	Precision	Recall	F1-Score	Accuracy
NB	0.83	0.82	0.82	82%
KNN	0.84	0.82	0.81	82.5%
SVM	0.82	0.81	0.81	81.25%
XGBoost	0.82	0.81	0.81	81%
MLP	0.87	0.86	0.86	86.25%

This study was conducted on the learning management platform adopted by some education systems. This platform gives many management and evaluation features. The information collected by the platform aims to develop a complete picture of the functioning of the educational organization in that system. The learning management system contains many vital data related to students' academic performance and behavior. This prompted the researcher to adopt this platform to make an early prediction on the performance and behavior of students based on the information provided by this platform. This study was conducted on 480 students through which categorical and related features were used so that a total of 16 features was used. These features were divided according to the category to which they belong. Some are demographic, such as impurity, academic ones, such as the average, and some are behavioral as the classmates and raising the hands. The researcher's features in assessing students' performance from the two tables 2.5, 2.6 is noticed.

Table 2. 5 Summary statistic for continuous variable

Raise Hand	Visit Resources	View Announcement	Discussion Minimum
1st Quartile	1st Quartile	1st Quartile	1st Quartile
Median	Median	Median	Median
Mean	Mean	Mean	Mean
3rd Quartile	3rd Quartile	3rd Quartile	3rd Quartile

Table 2. 6 Statistic for categorical variable

Categorical	Variable
Gender	Male, Female
Nationality	Kuwait, Jordan, Palestine, Iraq, Lebanon
Birthplace	Kuwait, Jordan, Palestine, Iraq, Lebanon
Semester	First Second
Relation	Father-Mother
Parent answering survey	Yes No
Parent satisfied with the school	Good Bad
Absent days	Above seven days Under seven days
Subject	IT French Science Arabic English
StageID	High School Lower level Middle school
GradeID	G-02 G-04 G-06 G-07 G-08
SectionID	A B C
Class (response variable)	High Low Medium

The researcher used Learning Vector Quantization (LVQ) to reduce the dimensions of the features matrix. He found that increasing resources and raising hands are critical features in determining students' achievement and the most important. It also used multiple machine learning algorithms such as Linear Discriminate Analysis (LDA), Classification and Regression Tree (CART), k-Nearest Neighbor (KNN), Support Vector

Machine (SVM) and Random Forest (RF) To check the accuracy and classify students' performance and compare and contrast the researcher's work with previous research in this field, see the attached table [11].

Table 2. 7 Comparison study with previous work

Algorithms	[11]		[12]		[13]	
	Accuracy (%)	F-measure	Accuracy (%)	F-measure	Accuracy (%)	Kappa
J48	75.8	75.9	75.8	75.9	-	-
Naïve Bayes	67	67.1	67.7	67.1	-	-
ANN	79.1	79.1	-	-	-	-
LDA	-	-	-	-	80	70
CART	-	-	-	-	75	62
KNN	-	-	-	-	71	56
SVM	-	-	-	-	81	71
RF	-	-	76.6	76.6	90	85

Predicting the number of students who may not graduate from the bachelor's stage was the focus of this researcher's study. According to the statistics center in their country, it was found that half of the students who enroll at this stage do not graduate. Hence the problem of research in the future prediction of those dogs who are expected not to graduate helps the decision-makers in that university to develop appropriate treatment plans and guide students so that they do not face this problem in the future. In his research, the researcher focused on the most accurate machine learning algorithm in classification and prediction, where he applied group models in predicting students' graduation. In his research, the researcher used binary classification criteria and the flow matrix to study the prediction accuracy and the algorithm's accuracy in prediction. The researcher found that group models would increase prediction accuracy and improve the accuracy in the classification outputs [14].

This study used machine learning techniques to evaluate massive online courses, known as MOOCs. Assessments based on machine learning helped increase the efficiency of those training courses. Increasing the efficiency of training courses was important because these platforms constitute a source of income for their creators. Hence,

selecting coefficients to be collected from that cycle had an essential impact on increasing classification accuracy. In these transactions, the study aimed for any of these factors that significantly impact the quality of the course. The boosting trees model was used to assess the extent of students' satisfaction with the course. The boosting trees model was used because it was the strongest in classifying than other classification algorithms. He relied on the following classification criteria: recall, accuracy, and F1-Score to select the best classification algorithm. He also used 10-fold in the random search method to test classification power. It was found through the analysis that the course teacher, the content of the instructional material, and the timetable play an essential role in increasing the efficiency of the course. At the same time, the duration of the video and the specialization of interaction on it do not significantly affect student satisfaction. Therefore, this study discussed the factors that affect student satisfaction at the learner and the course level, which can predict student satisfaction [15].

This study uses the MANTIS-S model in Student Academic Performance Prediction, which handles the training strategy. And the parameter set, the result showed increasing accuracy, enhanced the computational time, and achieved simplicity [16]. This study was conducted to assist in early intervention, and this is achieved by early prediction of students' performance. This study showed the superiority of deep learning over basic logistic regression, as the percentage of improvement increased by 10% [17]. The research aimed to measure using a learning management system on student performance. It was performed on 3518 university students. The neural network used homework score, time spent on the content, several entries to content, score, and gender in predicting student performance, 80.47% accuracy percent was achieved using this model [18]. This study was conducted in three colleges, and this study aimed to improve

students' performance by knowing their future performance expectations. The deep learning model was used to make the prediction. The Adam optimization method is used to carry out optimization processes, and other techniques are performed to make classifications, such as AdaBoost, to compare it with deep learning models [19]. This study aimed to compare conventional statistical evaluations to predict students' performance and the neural networks. Eleven variables are used. I'm modeling the neural network with two hidden layers. The results showed the superiority of neural networks over the traditional methods in the prediction process [20]. This Survey discusses the mechanism to increase the growth of education from another point of view. It studied the expectation of students' dropout from schools and the most general factors affecting this, and it was based on calculating the expectation on machine learning techniques. It also discussed the delay of developing countries in applying machine learning to improve their expectations [21]. In this work, semi-supervised learning algorithms have been applied for predicting student performance, this work reveals a new method that can be applied to make the system more reliable, and this can be achieved by using many unlabeled data and few labeled data the result show improvement in classification accuracy [22]. This study aims to overcome student prediction challenges: different in select course and background, depending on the course only does not give an accurate prediction, and the evolution of student must be implemented into prediction. The proposed new model supports a data-driven approach and a bilayered structure. The proposed method shows superior performance [23]. The feed-forward neural network is adopted in this study to analyze and establish the nonlinear relationship that exists between psychological and cognitive that affects academic performance [24]

Some models' employee another algorithm such as artificial bee colony (ABC), ABC algorithm is performed to enhanced radial basis function neural network for big data environments, the search strategy of ABC is replaced with mutation and crossover operator, this improves searching and optimal candidate food source [25] In this approach, the center is selected randomly from the calibration set, it aims to find the best position center depending on the score matrix from the whole space, this can be achieved by determining the generated scores of changing the location of a center which is taken from a genetic algorithm [26] In another study, inter-homogeneous and intra-heterogeneous groups are achieved by applying a genetic algorithm approach, this method allowed to get a deep insight into the character of a student as desired [27].

From the above, previous studies close to our research revealed their adoption of additional criteria other than grades in their work and their use of different machine learning techniques in improving education progress. Many fields of education have been targeted in classification and prediction processes, such as expecting graduation, determining students' future grades, and other fields. In our study, machine learning techniques reveal the students' expected specializations in the future. The used features approve by specialists in this field, and it is found similarity of some features with some previous studies in this regard, such as gender, signs, country, etc. After that, the pre-processing procedure is performed on the data set before it is used by machine learning techniques to support multiple classifications.

Chapter Three

Methodology

3.1 Introduction

This chapter will discuss the mechanism used to collect data from the Palestinian community, how the collected data was processed, cleaning the data from unacceptable values, and applying data encoding operations with more than one option. The data is also arranged according to its similarity by applying the standard degree. After the data is ready for use, it is entered into the different machine learning platforms used for classification and MLP neural networks to classify the student majors, choose the best results, and compare the applied ML techniques.

The following figure 3.1 illustrates the methodology used in this research. The process starts with Data collection from the ministry of education schools and databases. Then data pre-processing will be applied: The data will need to be processed, altered, excluded, and other steps before it is entered into a machine learning technology, as the processing process will affect the nature and accuracy of the result, then the Feature scaling is being used, then cross-validation will be applied to divide the data into training and testing. Finally, ML models and Artificial neural networks will be applied.

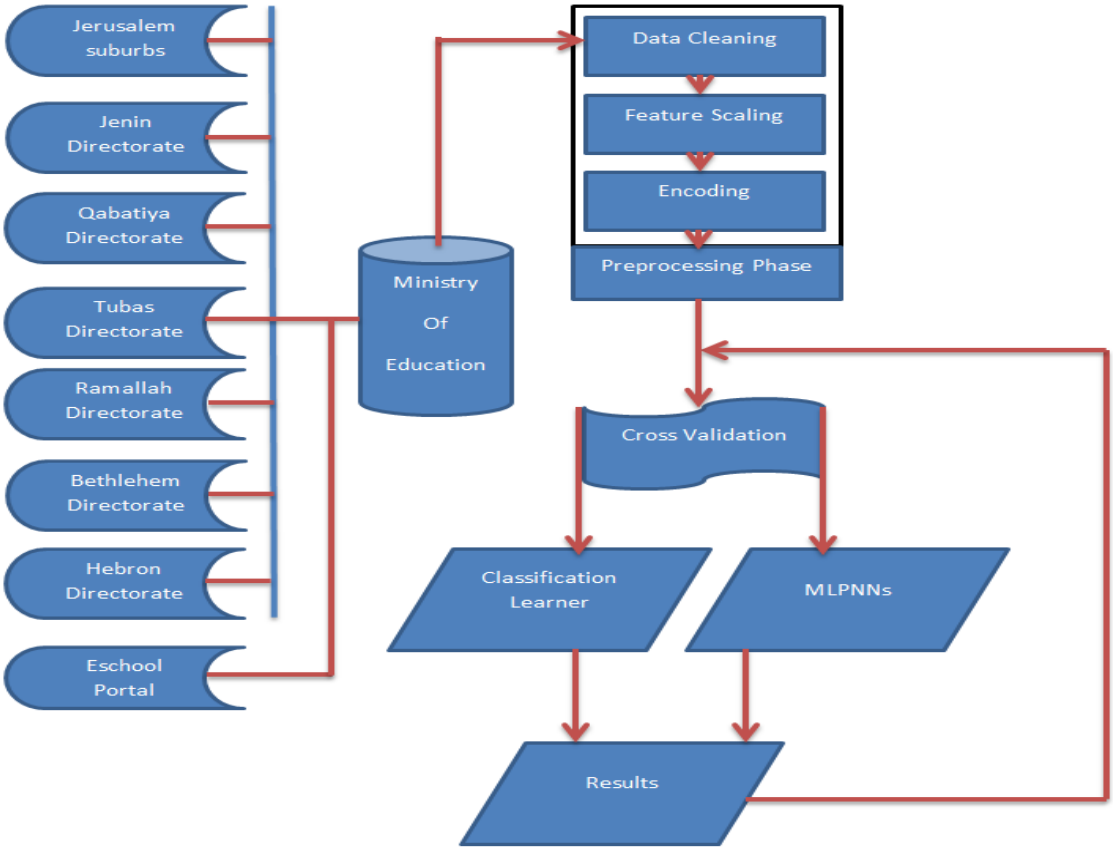


Figure 3.1 The general method procedure flow chart

3.2 Data Collection

The data was collected from the ministry of education schools, which students filled on a Microsoft form after taking the required permission to do this from MOE, see appendix A. The questionnaire was distributed in each of the following directorates: Jenin, Nablus, Ramallah, Jerusalem, Bethlehem, and Hebron, as well as approximately 1,300 random samples, were collected, the participants turnout in the North was the highest, reaching 45 %, in the middle, it was 20 %, whereas, in the South, 35 %, and the same was true for the eleventh grade. The electronic questionnaire was adopted as a compulsory, not a choice, and as a preventive measure in the wave of the new epidemic

spreading globally. Many problems were encountered in collecting data from the Ministry of Education, which lasted for more than three months due to health conditions and other administrative matters. The questionnaire was established with the assistance of educational experts. The questionnaire includes the students' academic marks in the tenth grade and other general information that affects the student's decision to choose his\her future primary, see appendix B.

The questionnaire was adopted and audited by specialist educators, and the Arabic language was adopted in the educational questionnaire. Specialists linguistically checked it to make it easier for the student to answer the questionnaires quickly and conveniently. It is worth noting that some of the added features were omitted from the questionnaire to preserve the privacy of the student, which is related to family problems, health, and financial matters. The principle of multiple choices has been adopted and not open answers in all classifications so that unification of all answers could be obtained.

Collecting data from the principals took three consecutive weeks, and communication and coordination were made with each directorate separately. At the same time, students were encouraged to fill out the questionnaire with the help of their teachers and the Marks platform in the Ministry of Education School. Despite the absolute care for the integrity of the data, the mechanism of its collection, and the binding restrictions placed on the data model, some random entries were necessary to enter the second stage, which is the stage of data processing.

3.3 Preprocessing Phase

3.3.1 Cleaning Data

Data cleaning is an important step performed as the first stage of the pre-processing phase for machine learning algorithms [28]. In the stage of data processing, and after a deep examination of the data, it was found that some students entered failed marks in the questionnaire; data which was filled based on the promotion laws was omitted so that the dataset was only for the successful student in academic subjects. In addition to that, some data is also omitted, which is not completing the entry of marks logically if it was found that he \she entered the rest of the marks by placing the Curriculum mark 1, besides excluding in the total marks, all the student's marks have been 100 in all marks, and this could not be primarily in physical fitness and arts.

3.3.2 Feature Scaling

Feature scaling has to be applied to the data. This critical step in preparing data is necessary for classification algorithms and neural networks algorithms, and this phase helps machine learning handle the input data.

There are many ways to scale the features: data normalization and standardization. In the normalization of the data, the standard technique used for this is the min-max model, in which the data is converted to values between 0-1.

The standardization model treats scattering in the data, especially in random data, and the values are distributed between -3 and 3. This process is suitable for machine learning platforms and is widely used [29], as shown in equation 3.1.

$$x' = (x - \text{average}(x))/(\text{std}(x)) \qquad 3.1$$

3.3.3 Data Encoding

In our approved questionnaire, there are two options for answering one question in every existing feature, and there are also some questionnaires with more than two options, and it is also a string value.

These options are prepared to be suitable for machine learning by encoding those options. The value of the first option has a value of zero; the second is one, and so on. Specifically, in neural networks, data that includes two or more options must be encoded into an array of the binary number represent option index. For example, the first option has the value 001; the second option is 010, and so on [30].

3.4 Building Model Phase

3.4.1 K-Nearest Neighbor (KNN)

It is known that the KNN technique uses the space between two points in the classification process, this algorithm is easy to apply, but the classification results achieved are not satisfactory. This technique is affected by choosing K parameters; the technique becomes sensitive on small k. When k is large, the wrong classification may occur since the neighbor may include some points that belong to another class, which may directly impact the integrity of the classification [31].

Technique Steps:

- First, set the number of classes K.

- Generate initial point for centroids by rearranging the dataset first and second, choosing K point of data randomly without replacement focuses for the centroids
- Do emphasis to achieve and make sure that no change cannot be achieved to the centroids.

The following equation (3.2): is used in Computing the summation of the squared distance between all data points and centroid

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} (\|x_i - v_i\|)^2 \qquad 3.2$$

Where $\|x_i - v_i\|$ is the Euclidean distance formula used to calculate the distance between a centroid v_i and the point x_i , these operations repeated for all k points in thei_{ih}classes, and also repeated for all n class. c Is represent the number of classes centers and c_i represents the data points number ini_{ih} classes[31].

Appropriating stage implemented using equation (3.3): it is to appropriate each data point to the closest classes, where c_i is represented the count of data points in i_{ih}.

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i \qquad 3.3$$

Take the average of all data points which belong to each cluster and then compute the centroids for the clusters.

A new approach is developed to overcome this issue called weighted KNN. This model uses a kernel function that gives the k parameter weight. This weight varies along with points distribution. In other words, the weight of the nearest point is larger than the far points, which helps determine the point which has to belong to the correct classes. This method depends on the majority vote. The results of this algorithm are more accurate unless the nearest neighbors do not differ significantly in their distance [32].

The technique is working according to the following procedure:

- Let's assume the training set of the Xi observation with associated class Yi
- Let X is a new observation that would be finding its classes
- Calculate the distance using Euclidean distance between xi and x for i=0 … n points
- Select D' ⊆ D. This determined query point is nearest to the k set of the training data point.

The voting technique is used to determine the predicted class which it is called weight technique as formula 3.4.

$$y' = argmax_v \sum_{(xi,yi)\in D}^{v} \text{wi} * \text{I (v} = \text{yi)} \qquad 3.4$$

This formula works according to the following procedure: Calculate the inverse of the distance. Hence, the distance the inverse = 1\d, then calculate the sum of the inverse, divide each item by the resulting sum. After that, calculate the generated value which belongs to its class. Finally, choose the max value to make this class the predictable class.

3.4.2 Support Vector Machine (SVM)

In this technique, the data is divided by finding a plane that can separate the distributed elements in space based on the number of features each element has. Most importantly, the best level to distinguish those elements well is found [33]. This technique depends on determining the optimum plane to separate the data with the best separation distance between the plane and points with similar characteristics. The plane's most significant distance to the closed points is taken [34].

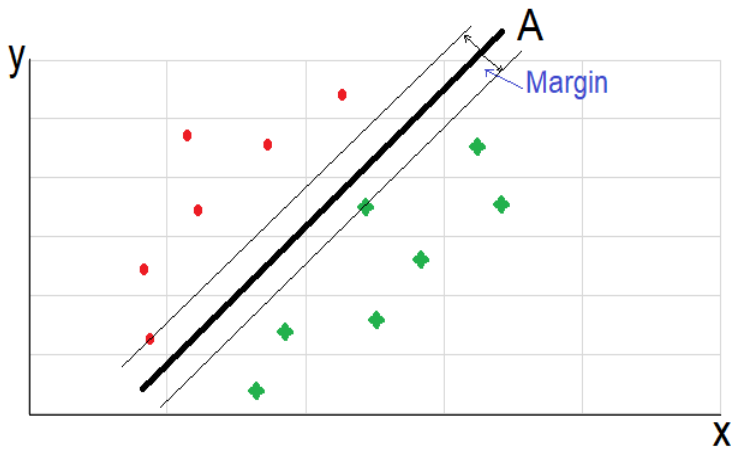


Figure 3.2 The SVM hyperplanes, margins, and support vectors

It is worth noting that this technique can ignore outliers. For example, if a star appears in groups of circles, this feature achieved by this algorithm gives it strength against outliers .Also, this technique can develop new features to create planes capable of separating the elements distributed in the space with high efficiency, adding its ability to adapt to the data through the kernel technique, which allows it to handle complex and nonlinear functions.

As is mentioned earlier, in this technique, the most significant possible margin between the hyperplane and the closest points to the elements is tended to be found, and this can be applied using the loss function c(x,y,f (x)) that gives a relation between the training data (Xj, Yj) for j = 1 ...N, with Xi ∈ Rd and Yj ∈ {−1, 1} on the one hand and learning a classifier f(x) on the other hand[35], see equations 3.5,3.6.

$$c\big(x, y, f(x)\big) = \begin{cases} 0 \ , & if\ y * f(x) \geq 1 \\ 1 - y * f(x), & if\ y * f(x) < 1 \end{cases} \qquad 3.5$$

$$c\big(x, y, f(x)\big) = (1 - y * f(x))_+ \qquad 3.6$$

The cost will be zero when the result is similar for both the expected and actual values. Otherwise, the loss cost is calculated if the expected and actual values are not similar. Regularization parameters can be used in the cost function, which balances losses. See (3.22) where λ is regularization parameters, the input is Xj, and the weight is W

$$\min_w \lambda \|w\|^2 + \sum_{i=1}^{n} (1 - y_i \langle x_i, w \rangle)_+ \qquad 3.7$$

In loss function, the weight has to be updated, this update can be achieved using a gradient, and this gradient comes from partial derivatives concerning the weights found for the data point, which this illustrated in equation 3.9, 3.10.

$$\frac{\delta}{\delta w_k} \lambda \|w\|^2 = 2\lambda w_k \qquad 3.9$$

$$\frac{\delta}{\delta w_k} (1 - y_i \langle x_i, w \rangle)_+ = \begin{cases} 0 \ , & if\ y_i \langle x_i, w \rangle \geq 1 \\ -y_i x_i, & if\ y_i \langle x_i, w \rangle < 1 \end{cases} \qquad 3.10$$

In the loss function, the weight has to be updated as seen in equation 3.11, and this update can be achieved using gradient; this gradient comes from partial derivatives, which concern the weights found for the data point.

$$\omega = \omega - \alpha.(2\lambda\omega) \qquad 3.11$$

When miss prediction occurs, this leads to miss classification, so one of the most solutions that could handle this issue is using equation 3.12, which includes loss and the regularization parameter.

$$\omega = \omega + \alpha.(y_i.x_i - 2\lambda\omega) \qquad 3.12$$

3.4.3 Decision Tree

The decision tree is one of the most popular techniques because it has an effective and influential role in the prediction and classification process. This technique is similar in its construction to the tree model so that its nodes represent a test of the features so that their results appear on the branches and that each end sheet of it bears a name from the existing names of the categories [36]. The decision tree learns by dividing the data set into subsets based on the tests it performs at its internal contracts. These operations are repeated similarly in a process called recursive partitioning. When the values of the target variable are equal to the subset, the recursive partition process stops. There is another case in which recursive division stands, in which it is not possible to obtain new results for the predictions. The application of this model does not require the preparation of some parameters or prior solid knowledge in the domain, and it is also characterized by its ability to deal with high-dimensional data, see figure 3.3. The decision tree is

almost accurate. It is good and also characterized by its speed. Thus, it can be used in the rapid induction of classifications [37].

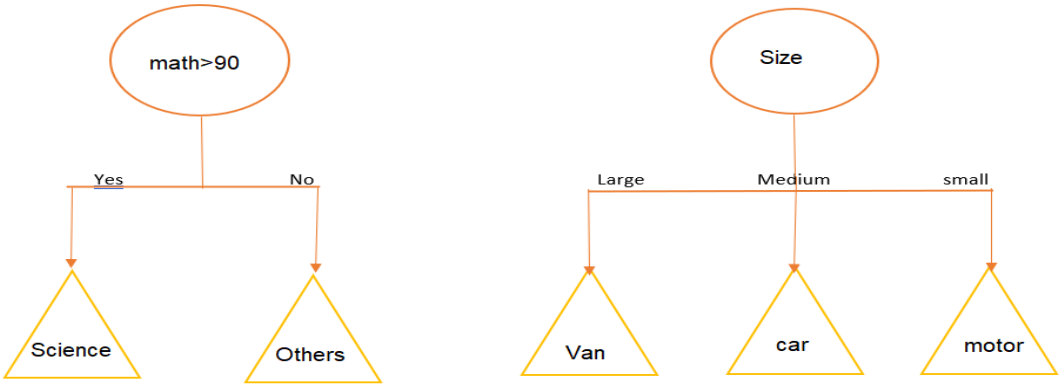


Figure 3.3 Decision tree two and three classes [37]

The decision tree adopts a top-down approach to the data to collect similarities. Decision trees are based on layers to divide data. By dividing the data, it seeks to divide it into two or more groups to include elements similar to each other. The multi-directional division that the decision tree adopts has an influential role in not repeating the traits within the decision tree, so the presence of repetitions of the traits within the decision tree is rarely noticed. The decision tree is characterized by creating easy-to-interpret rules that divide the data with simple arithmetic operations and indicate the essential fields in classification and prediction ،But an evident shortcoming in the various classifications in the decision tree is found, especially if the data for training is relatively few.

The Decision Tree technique can be summarized as the following steps

- The tree begins with a root node called R, and this node contains the whole dataset.

- Splitting the dataset into two main subtrees using the best attribute found in this level.
- Iterate the previous step for all subtrees on the decision tree.
- The best attribute is found on the decision node that is used for splitting
- The splitting procedure iterates to reach the final node and reach a good result or when the splitting procedure doesn't give any value to the classification.

3.4.4 Linear discrimination

This technique is used to reduce the dimensions in supervised classifications. It can be used to classify two classes and also used to classify several classes. This model assumes that the classes it can be separated linearly, based on the several previous hyperplanes, can be constructed using multiple linear discrimination functions for the future space to distinguish between the classes wells. The LDA merges the two hyperplanes representing two lasses on one hyperplane, as shown in figure 3.4. The data points on each hyperplane are projected on the new hyperplane, considering that the separation between two classes is maximized and the variation is minimized as much as possible [38].

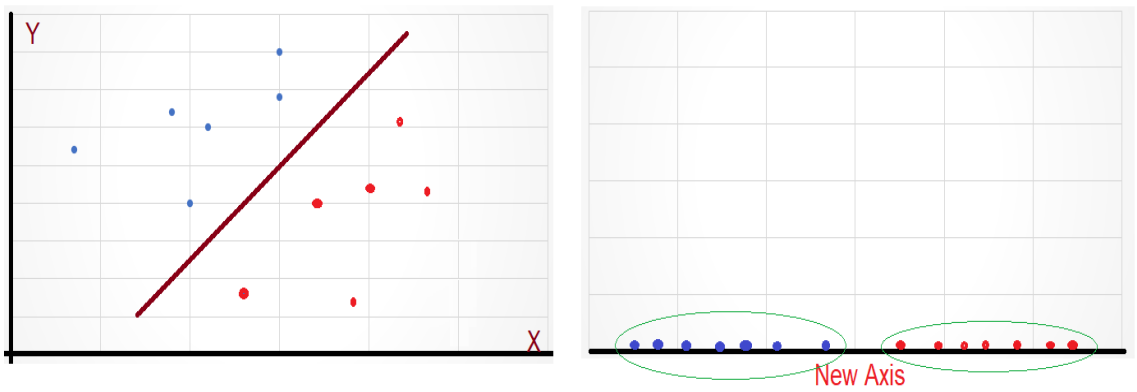


Figure 3.4 Linear discrimination dimension reduction

Hence, it is noted that this algorithm works efficiently from multiple classes. It is also used to reduce the number of features before introducing them to the learner. It is also used to extract valuable data in face recognition processes. But this technique may not work well with those not linearly separable classes.

LDA technique general steps:

- For different classes, find d-dimensional mean vectors.
- The scatter matrices value have to be found within, in, and between classes
- The eigenvectors and their corresponding have to be computed for the scatter matrices
- Eigenvectors must be sorted decreasing depending on eigenvalues, also find the largest eigenvalues for choosing eigenvectors to construct matric d*k whose column represents the eigenvectors.
- Use the previous vector to create the new space and transform the sample.

3.4.5 Ensemble Techniques

The ensemble technique is used more than one combined model instead of the individual model. This construction is used to obtain more accurate classifications compared to using one model alone, as it was found that more accurate classifications can be obtained if this construction is used. Undoubtedly, using a group of algorithms combined requires more complex mathematical operations and more mathematical operations in evaluating the model and checking the integrity of the prediction. This construction is suitable for improving those weak algorithms by merging them and increasing the calculations procedure to obtain high accuracy in classifications. Using

one model with increased complexity and resources may be an alternative to combining many algorithms. Still, it was found that the accuracy that resulted from the combination is better than using a single model. The application of this merging appeared in fast algorithms such as decision trees. It is also integrated with different machine learning platforms; this does not mean slow algorithms cannot be used [39].

To implement the Ensemble model, two basic methods are adopted. The first is a sequential assembly of algorithms. The base learner is used in sequential techniques implemented sequentially, and this sequence, constructed by the generation of base learners, aims to increase the relation and dependency between base learners. It gives higher weight for those not represented well to increase the model performance. It appears in figure 3.5.

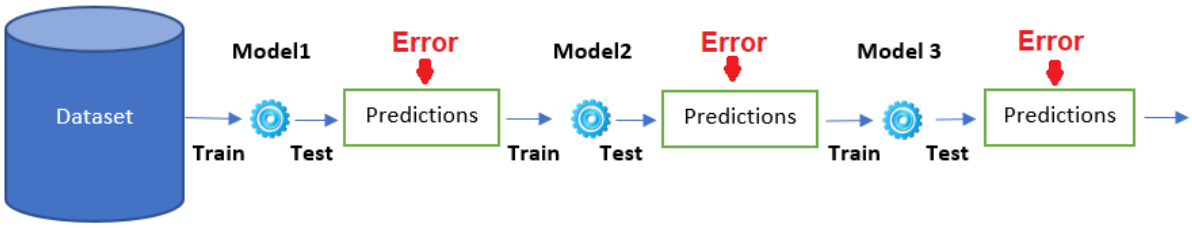


Figure 3.5 Ensemble algorithm sequential model

The second technique is a parallel assembly of algorithms. In this technique, constructing the bases learning to work parallel like the implementation assembled model that combined decision trees to construct random forest trees encourages independence between models. It uses a slightly different partial training data set to learn the error generated by the previous model, see figure 3.6 [40].

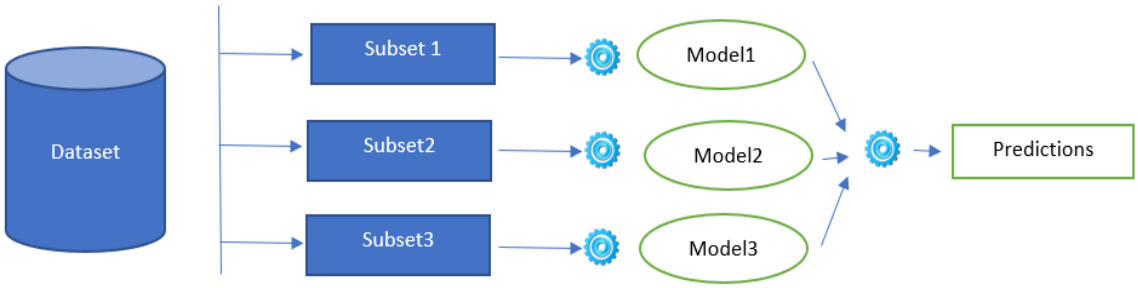


Figure 3.6 Ensemble technique parallel model

The techniques mentioned above adopt scientific terminology as follows:

Bagging

The word bagging comes from bootstrapping and aggregation. Using the same data set to train models is likely to produce similar results, so bagging comes to overcome this problem by separating the dataset into sub-datasets with slight changes in bootstrapping. Bootstrap aggregating is carried out to understand the distribution of data, illustrated in figure 3.7.

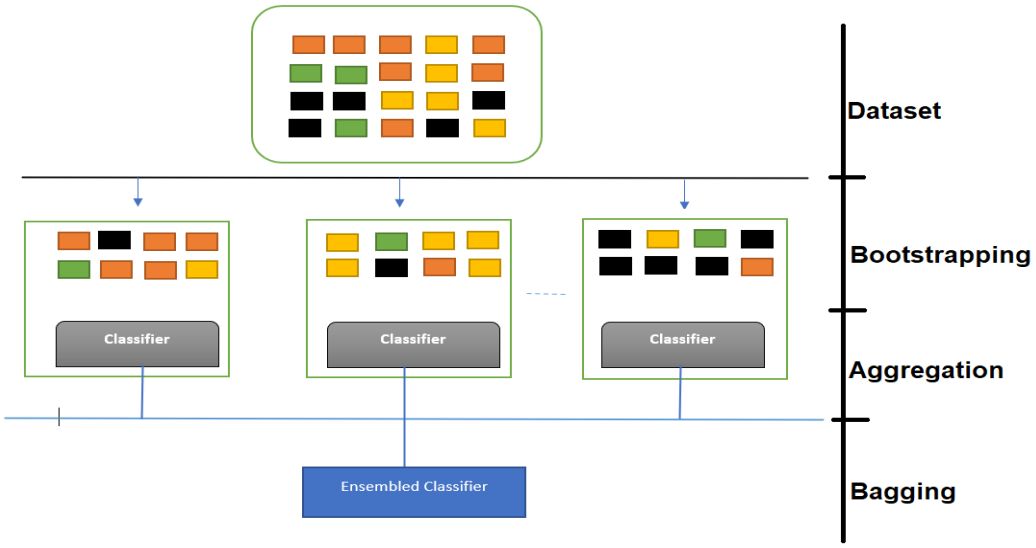


Figure 3.7 Ensemble algorithm bagging

Implementation Steps of Bagging:

- Step 1: Generate several subsets of the same size from the original dataset; the selected observation has replacement capability.
- Step 2: Each subset creates a base learning model and assigns a base learning model.
- Step 3: Every model is learned in an independent way and parallel.
- Step 4: Combining the prediction generated from the models to get the final predictions.

Random Forests Ensemble technique: This uses bagging techniques, decision tree technique is used for each data set, the final result is combined to construct an algorithm, the distribution of samples/features is distributed randomly in implementation, this called the Random Forests technique, the technique is executed through the following operation as follows:

- Dividing the dataset into a subset with the same size then choosing several features after that creates a decision tree to handle and classify it.
- Repeating the previous procedure using different variables covering the whole data set, the chosen variable could be used in several trees
- Doing a test for each tree after completing the procedure.
- Choosing the result from the generated results from the above steps.

Boosting

The boosting ensemble technique is a sequential procedure, and it is designed for a weak leaner algorithm such as a decision tree, which aims to decrease error in

classification. This algorithm decreases the error in classification by increasing the weight for misclassification data points and repeating this procedure to obtain the desired classification [41], note figure 3.8.

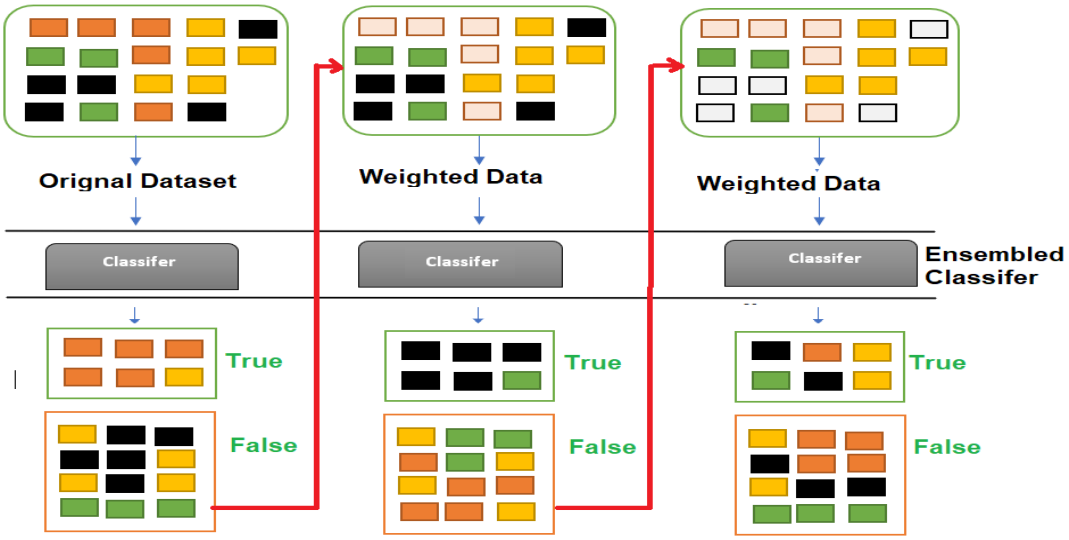


Figure 3.8 Ensemble algorithm boosting

Technique steps*:*

- Generate dataset and give an equal weight for whole data points
- Feed the model by the generated dataset above with corresponding's weights
- When the classification of some data point is wrong, increase its weight
- If suitable result -> end , else -> go to step 2

3.4.6 Multi-Layer Perceptron Neural Networks (MLPNNs)

The idea of neural networks came in machine learning to simulate the work of the neural networks in the human brain. It was found that the human brain depends on

responding to the events taking place around it by referring to the previous information that is previously stored in its memory and how it responded to it previously or how other people responded to it or how it can respond to it without prior knowledge; that what was previously mentioned represents the different machine learning methods in neural networks, some of them are learned through a supervising teacher, and others need to use mechanisms without the need for previous information about the field of event that you want to decide a process called Learning without supervision [42].

Neural networks have been employed in many fields, including medicine, industry, agriculture, education, and many others. It plays a significant role in predicting many fields, such as forecasting the cost of production, academic disciplines, their names, diagnosing diseases, etc.

It is worth noting that neural networks are learning techniques that can calibrate themselves based on the difference between actual output and predicted outputs to balance their networks until they come out with a firm structure capable of accurate prediction and classification, and not a computer program that cannot be calibrated and controlled.

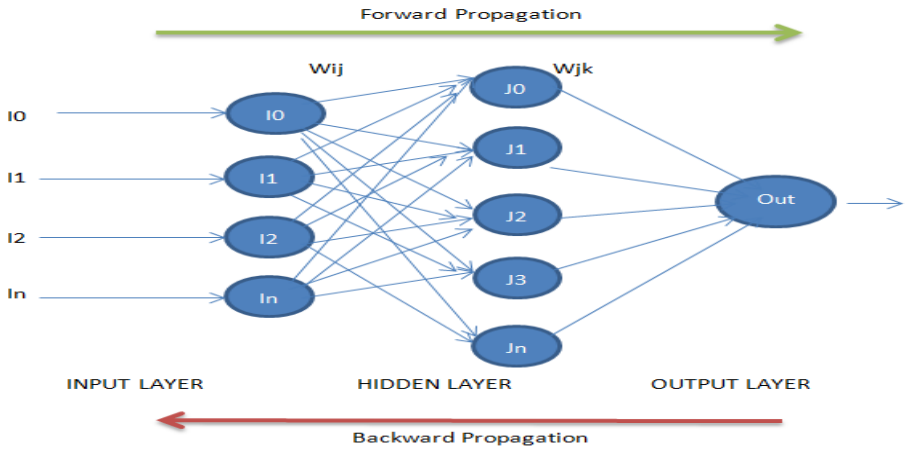


Figure 3.9 MLPNNs Topology

Figure 3.9 shows that the neural networks depend on a primary nucleus called the neuron; neural networks contain many neurons, and many connections between the neurons. These connections have weights based on the neural network's learning algorithms. These weights are adjusted by knowing the error resulting from the prediction.

The algorithms used in neural networks depend on errors between the predictable output and the actual output to adjust the weights to get the least possible error in predicting the output. The neurons are distributed over layers, some neural networks contain one layer, and others consist of many layers. The algorithms differ depending on the number of layers and the application [43].

Here the layers in multi-layer neural networks will be explained; these networks contain three main layers:

- Input layer: The input layer represents the source of nutrition for the neural network. This network is responsible for entering data into the neural network without performing any operations. It should be noted that the number of these cells represents the number of features that have been developed to study a particular field
- Output layer: The layer on which the prediction results predicted through the neural network processes appear. The outputs from this layer are used to compare them in the actual output. Through this comparison, the error resulting from the prediction for the output is calculated in the supervised learning process. This is used the mistake is made later to adjust the neural network's weights to predict more accurately. The number of neurons is the number of

outputs that the neural network predicts. This is taken from the domain study to which the neural network is being applied.

- The hidden layer is the layer between the input and output layers, which is used to re-adjust the weights in the neural network down to the weights between the hidden network and the neurons in the input layer.

Neural networks depend on prediction accuracy on adjusting the weights between their neurons. These weights are mathematically multiplied in the classes entered through the input layer. The output from this process enters into an arithmetic chain inside the hidden layer. It is processed until it goes out to the output layer to form the expected results from applying neural networks to them. This stage in the learning process of the network is called the forward propagation stage, as, during this stage, the outputs are predicted and compared with the accurate outputs that it is sought to reach. At this stage, the error resulting from the previous prediction process is taken advantage of to reflect its effect within the neural networks. The weights are adjusted through mathematical operations to reach the adjustment process for weighting between the hidden and the input layers. This process is repeated in a loop to reach the least possible error in prediction operations. The repetition of the process is sometimes stopped if it is found that it does not reduce the error and reach the desired prediction. It is worth noting that, at the beginning of the formation of the neural network, the weights are set randomly as starting values to be calibrated through the processes mentioned above [44].

Many mathematical equations are used on neural networks to calculate the network output for the aforementioned operations. The following equation3.13 is implemented:

$$y = \sum W_{ij} X_i + \theta_j \qquad 3.13$$

θ_j: called the bias node

Where w_{ij}: is the weight of the connection between neuron ith neuron exists in the input layer and the *j*th neuron exists in the hidden layer, and x_i: is the *i*th input features.

As mentioned previously, the results produced by neural networks are compared with the accurate output data sought to reach. This comparison is made by finding the difference between the output from the network and the actual output, which is called the error, which is symbolized with the symbol $\Delta_k$and it is rented using the following mathematical equation 3.14

$$\Delta_k = t_k - Y_k \qquad 3.14$$

y_d is a predictive output and y_i is the actual output

A mechanism is also used to measure the error in the neural networks called the least square error. Note the equation 3.15, that the result of this equation is the least error it can be strived to reach.

$$MSE = \frac{1}{2}\sum_{i}^{n}(y_d - y_i)^2 \qquad 3.15$$

The process of teaching the neural network enters into a loop. The error resulting from the prediction process is calculated, and the weights are reset to reach the lowest possible error [45]. The process of adjusting the weights is done through the following equation 3.16:

$$\Delta w_{i+1} = \alpha . \mathrm{E} . x_i \qquad 3.16$$

When the training phase is fed to the input layer, the sum of weights from input to the j[th] node in the hidden layer is given by:

$$y = \sum W_{ij} X_i + \theta_j \qquad 3.17$$

Inside the neuron, there is a function called an activation function. In multi-layer neural networks, the sigmoid function is known to be used. The output from the j[th] neuron is:

$$Y_j = X_k = \frac{1}{1+e^{-y}} \qquad 3.16$$

δ_k: The error signal can be easily calculated on the output layer by multiplying the Δ_k with the derivative of the sigmoid function used on the activation function.

$$\delta_k = \Delta_k Y_k (1 - Y_k) \qquad 3.17$$

The weight is tuned depending on the generated error; it was decreased or increased by delta weight which is calculated using equations 3.18

$$\Delta w_{jk} = l\, \delta_k X_k \qquad 3.18$$

Δw_{jk}: The delta weight calculated depends on the weight between neuron j and k, which could be controlled using the learning rate. The weights between neurons are updated depending on the delta weight generated in the previous equation by implementing equations3.19.

$$w_{jk} = w_{jk} + \Delta w_{jk} \qquad 3.19$$

The equation 3.20 is used to calculate the partial error δ_j that used in pack propagation stage

$$\delta_j = (t_k - Y_k) Y_k \sum w_{jk}\, \delta_k \qquad 3.20$$

w_{ij}: is the weights between two neurons i,j that is updated using the equations 3.21 and 3.22 so

$$\Delta w_{ij} = l\ \delta_j X_j \qquad 3.21$$

$$w_{ij} = w_{ij} + \Delta w_{ij} \qquad 3.22$$

Forward and backward propagation is frequently used to reduce the error rate or if there is no significant improvement in forecast accuracy.

3.5 Classification Metrics Selection

One of the most important things that should be taken into consideration in measuring the effectiveness and strength of the model applied in machine learning, so it was necessary to adopt criteria for the judge and measure the feasibility of applying the machine learning model and its ability to predict the outputs with high efficiency [46]. Many criteria are applied to know the efficiency of the model applied to the data and the accuracy of the prediction. The most famous criteria here will be reviewed, which must be implemented on each machine learning model [47].

The classification has two options in binary classification processes, either a favorable option or a negative one. If the classification is correct, the classification is positive. If the classification is incorrect, the classification is negative. Still, some classifications are correct, but the model mispredicts them, and some classifications are false while the model predicts them correctly. The best mechanism for studying the classifier's performance is to apply the confusion matrix to understand those mentioned above. The diagonal line represents the correct prediction, showing the number of correctly

predicted samples. At the same time, the rest of the cells include the number of incorrectly predicted samples, whether they were positive first or negative [48].

Table 3. 1 Confusion matrix

		Predicted Class	
Actual Class		Class = Yes	Class = No
	Class = Yes	TP (N)	FN(N)
	Class = No	FP(N)	TN(N)

N: Number of data points that are predicted according to the status of the cell

TP: It represents the correct classification for the positive class classified as actual.

FP: It represents the wrong classification for the positive class that is not classified as actual.

TN: It represents the correct classification for the negative class classified as actual.

FN: It represents the wrong classification for the negative class that is not classified as actual.

The previous classifications and the number of data in each class are used in special mathematical equations to measure the different machine learning models' tools.

Accuracy: The accuracy scale is one of the most widespread measures that expresses the ratio of healthy observations expected by the learning model to the sum of all observations. It is worth noting that if the accuracy is very high, this does not mean that our model works correctly or ideally. Hence, other measuring tools to check the model's integrity must be used. The following equation is applied to measure the prediction accuracy from applying a particular learning model.

$$Accuracy = \frac{\mathrm{TP} + TN}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}} \qquad 3.\ 23$$

Precision (Positive Predictive value): This scale is used to find the ratio of the positive observations correctly predicted to the total positive observations. Hence, the accuracy is high and satisfactory if the number of false-positive observations is low. Note the following equation.

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \qquad 3.\ 24$$

$Recall$ **(Sensitivity or true positive rate):** This scale is used to find the ratio of correctly predicted positive observation to the total of all actual positive observations represented by that class

$$Recall\ = \frac{T\mathrm{P}}{T\mathrm{P} + F\mathrm{N}} \qquad 3.\ 25$$

The question here arises, which is more important, precision or recall. It is impossible to answer this question because each of them has a field in which it is used, so employing one of these measures depends on the type of field applied to it. To understand the difference, note the following questions. In the case of precision, the question is, for example, how many students have succeeded out of those who were classified as successful. As for the recall, the question is, for example, how many successful students have been labeled as successful of the total of the actual successful people.

Specificity: This is the ratio between true negative predictions to the sum of all false positive and true negative observations.

$$Specificity = \frac{TN}{TN + \mathrm{FP}} \qquad 3.\ 26$$

F1 score :This scale represents the weighted average for precision and recall. This scale considers false positives and false negatives as well. The strength of this scale appears in the unequal distribution of the category. The accuracy criterion is well suited to cases where the false positives and false negatives are equal. If false positive and false negative are not equal, it is more suitable to use these matrices to understand the efficiency of the learning model and to judge it.

$$\text{F1 Score} = \frac{2 * (\text{Recall} * \text{Precision})}{(\text{Recall} + \text{Precision})} \qquad 3.27$$

Receiver-Operating Characteristic ROC:

One of the most important things that must be studied is the mechanisms of judging our learned model and its ability to predict. Here it was necessary to address another measuring tool to have a deep understanding of the efficiency of the learned model. ROC is a graphic curve that helps understand and visualize the quality of our machine learning model. It is designed to work efficiently on binary models, and it is worth noting that it can be used to work on multi-class models. It is a curve capable of depicting and evaluating errors in binary models. It is drawn based on the true positive rate (TPR) versus the false positive rate (FPR), obtained from equations 3.28,3.29,3.30. In other words, it separates the signal to be received from the noise. The ability of a particular classifier to distinguish between classes can be measured using the area under the curve. Also, it can be used as a summary of the ROC curve [48].

$$FPR = 1 - \text{Specificity} \qquad 3.28$$

$$FPR = FP/(FP + TN) \qquad 3.29$$

$$FDR = TP/(TP + FP) \qquad 3.30$$

Suppose the system's ability to distinguish between negative and positive categories is high. In that case, the value of the area under the curve AUC is high, and this indicator can be used to know the system's ability to distinguish based on the value of the area under the curve AUC. According to the values of the area under the rocker coupling curve. These values carry logical significance as follows:

Suppose the value of the area under the curve is equal to one. In that case, the classifier can distinguish between all positive points and all negative points, which is the ideal case. But suppose the area under the curve is equal to zero. In that case, this indicates the inability of the classifier to differentiate between negative points and positive points, as shown in figure 3.10 A. But in the case in which the area under the curve is greater than half and less than the ideal position, this classifier can distinguish between positive points and negative points in a satisfactory manner and according to the value of this area as shown in figure 3.10 B. If the value of the area under the curve is equal to half, this indicates the inability of the classifier to distinguish between positive points and negative points. In this case, the classifier predicts random categories or fixed categories for all observations, as shown in figure 3.10 C.

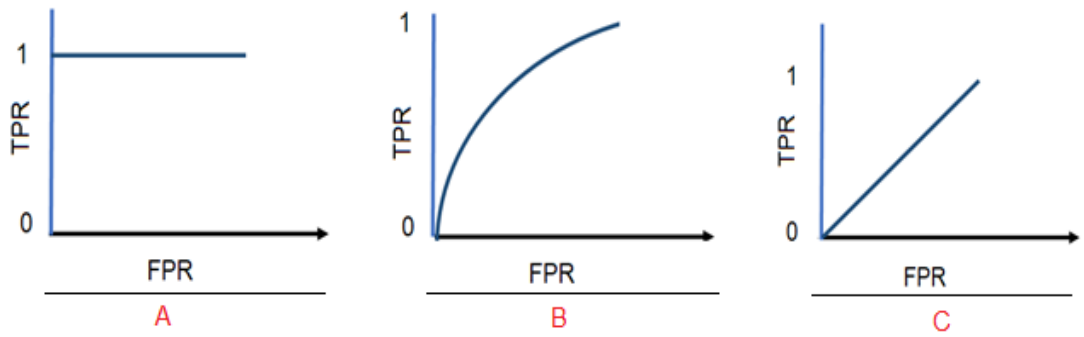


Figure 3.10 A : ROC With AUC =1, B: ROC With AUC >0.5, C:ROC With AUC =0.5

From the above, it is concluded that the value of the area under the curve, if it increases, the classifier's ability to predict and separate between positive and negative values increases. In this case, the measurement in multi-class measurement issues is done similarly, and the question is transformed from a multi-class to a binary class. The error in anticipation is only a unit group complementing the correctness in the prediction for all the sets of observations. For the sake of clarity, note the following table 3.2 and the equations that follow:

Table 3. 2 Confusion matrix multiclass

		Predicted Class			
Actual Class		Class 1	Class 2	Class 3	Class 4
	Class 1	TP	FN0	FN1	FN2
	Class 2	FP0	TN0	TN1	TN3
	Class 3	FP1	TN4	TN5	TN6
	Class 4	FP2	TN7	TN8	TN9

The following equation can be used to calculate the confusion matrix's precision.

$$\text{Precision} = \text{TP}/\text{TP} + (\text{FP0} + \text{FP1} + \text{FP2}) \qquad 3.31$$

Where the recall can be calculated by using the following equation.

$$Recall = TP\ /\ TP + (FN0 + FN1 + FN2) \qquad 3.32$$

A macro represents the arithmetic mean of all classes within the same matrices.

$$Macrro = (matrices1 + Matrices\ 2 + \cdots + Matrices\ n)/n \quad 3.33$$

While the weight represents each classes matrices multiplied by the number of its repetitions (class weight) divided by their number

$$Weighted = (W(C1) * Matrices1 + W(C2) * Matrices2 + . + W(Cn) * Matrices)/n \quad 3.34$$

Where c: class and W: weight. This strategy is called one-vs-rest, which seeks to transform the system from a multi-class problem to a binary one [49].

Chapter Four

Experiments and Result

4.1 Introduction

This chapter will review a simple statistical summary of the dataset according to students' responses, including the number of participants in each field classified by students on each branch to overview the dataset used. After the preprocessing phase, different machine learning techniques were applied for classification and applied to the collected data. The classification techniques provided by MATLAB were used, and techniques were selected according to their working mechanism to compare them and which were most appropriate for classification.

MATLAB 2018b program was run on a computer from a hp zbook 8th generation processor, 1.8 GHz, 16 GB RAM, and 512 SSD m2 hard disk. The previous device with the mentioned specifications gave the results in good time which is less than 5 minutes and is suitable for the techniques used for classification. Six techniques were applied: KNN, SVM, linear discriminant, Tree, ensembled boosted tree, and ensembled bagged tree, suitable for multi-category classifications. , The classification result of this algorithm will be discussed separately. Also, the results will be compared with each other's to determine which of these algorithms is suitable for our classification.

MLP neural networks were also used for classification, and the classification was applied by changing the number of neurons used by the neural network each time. The outputs will be displayed with accurate results. This result helps determine the

suitability, and the least neuron has to be used on classifications. A comparison will also be made between neural networks with the previous classification methods to determine who excels in the classification process, the most important results, and the criteria for excellence in each of them.

4.2 Data Analysis

The statistical data has been filled into a table. This procedure aims to review the data and important observations in them quickly. See the following table 4.1.

Table 4. 1 Students dataset overview

#	Scientific Branch				Literary Branch				Commercial Branch				Industrial Branch			
1	316	171			292	272			38	45			64	36		
2	117	370			118	446			5	78			66	34		
3	475	4	4	4	51	213	29	33	0	5	78	0	4	0	0	96
4	88	267	121	11	158	278	113	15	18	39	24	2	28	66	6	0
5	49	80	274	84	41	85	334	106	5	7	69	2	0	20	64	16
6	49	155	197	86	98	247	160	59	10	34	27	12	22	36	30	12
7	143	206	113	25	198	195	122	49	33	33	8	9	38	34	12	16
8	390	89	5	3	449	89	19	7	63	19	0	1	88	4	8	0
9	43	444			48	516			7	76			8	92		
10	349	138			491	103			47	36			68	32		
11	158	185	22	122	64	273	44	183	2	42	21	18	12	44	12	32
12	128	278	18	63	58	358	15	133	6	49	18	10	32	56	0	12
13	195	135	43	114	86	245	55	178	6	22	38	17	8	26	16	80
14	177	232	12	66	85	328	18	133	5	33	30	15	32	38	4	26
15	238	249			292	272			41	42			40	60		
16	87	400			82	482			9	74			16	84		
17	283	204			296	268			61	22			54	46		
18	63	424			66	498			12	71			50	50		
19	263	38	137		288	30	246		50	2	31		66	0	34	
20	320	107	9	51	75	375	14	100	30	19	26	8	42	4	0	54
21	94				79				79				86			
22	91				75				78				80			
23	87				67				72				71			
24	90				67				79				73			
25	89				63				69				74			
26	92				68				79				78			
27	92				72				81				78			
28	93				80				80				84			
29	93				89				92				96			
30	93				90				92				95			
31	92				79				83				78			

Where the numbers at the first columns refer to the following terms:

1	Academic level	12	Father specialization	22	Islamic education
2	Gender	13	Mother specialization	23	Arabic
3	Current major	14	Father tendency	24	English
4	The student tends to specialize:	15	Mother tendency	25	Math
5	In terms of social relations, the student is classified as	16	Father is dedicated to his children	26	Physics
6	The student is fully full time to study	17	Mother dedicated to her children	27	Chemistry
7	Number of daily study hours	18	Competition of relative	28	Life Sciences
8	Number of hours spent on online and mobile daily	19	Availability of major / specialization in his current place	26	Palestine history and geography
9	Number of hours spent on playing and physical exercise per day	20	Place of living	30	Arts and crafts
10	Father works	21	Community orientation	31	Physical education
11	Mother works			32	Technology education

After analyzing the questionnaires, it is found that most of the students in all branches have normal social relationships. Also, the percentage of voting for eleventh-grade students is greater than for twelfth-grade students which are more than 55%. Furthermore, the percentage of females voting is more than males in all branches except for the industrial branches which is more than 70%, as seen in figure 4.1.

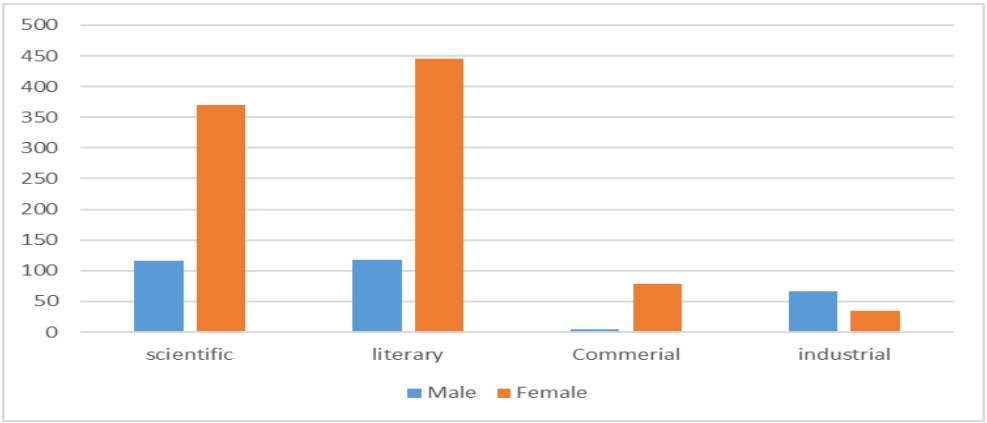


Figure 4. 1 Representation of the student's gender among the majors

As noted, too, there is a little and normal difference in the voted rate in some features. For example, the mother's preferences play an important role in selecting the students' majors. The same goes for the rivalry between relatives. On the contrary, some features differed in some branches. For instance, the highest percentage of study students between three and six hours are scientific and industrial branches students.

The questionnaire revealed important observations about the industrial branch. It is found that half of the students do not specialize in their town, as illustrated in figure 4.2. Female participants from this major are much less than males. Adding to this, the students of this major are never inclined to study the literary and commercial majors.

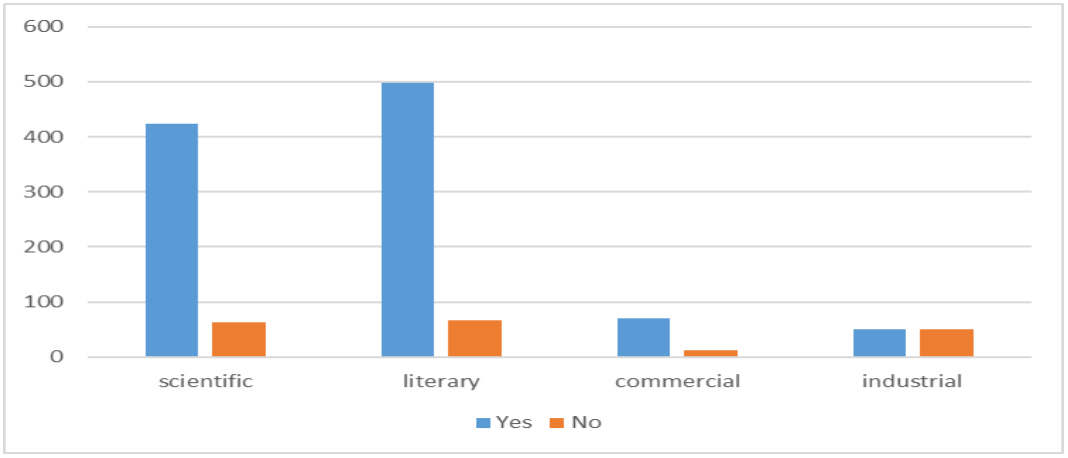


Figure 4. 2 Availability of major

Furthermore, the students' fathers in the industrial branch remarkably encouraged their children to study and noted that none of the students among the participants lived in a camp. At the same time, it is found that students of the commercial branch are never inclined to study the scientific or industrial branch, and also that the percentage of male participants in the commercial branch is much lower than females.

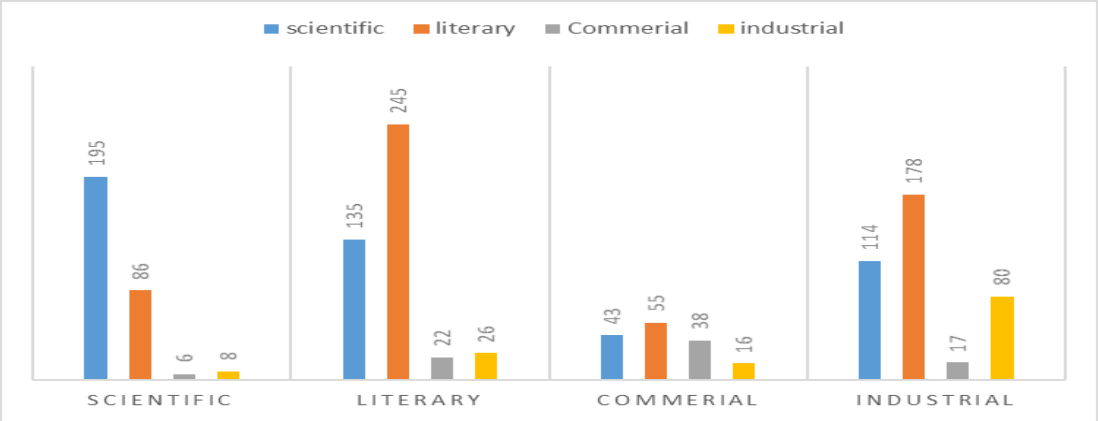


Figure 4. 3 Father tendency

4.3 Machine Learning Classification Learner Result and Comparison

In this section, the results are reviewed from the machine learning techniques used to classify students' majors. The different measurement parameters mentioned in the previous chapter are used to judge classification techniques. The judgment progress is essential for determining the technique's capability for these types of classification. Furthermore, the classification matrices are used to ensure the strength and safety of the machine learning techniques classification. Also, to help choose the suitable one for our case

4.3.1 Machine Learning Classification Learner Result

This section reviews the results produced from the application of machine learning classification learning algorithms.

4.3.1.1 KNN Experiment Result

The KNN technique is used to classify a student’s major. The number of clusters used is 10, the inputs variables are 31, and the number of observations is 1234. 5-fold cross-

validation is implemented, with several classes equaling four. After applying this technique, the result appears in the following table 4.2.

Table 4. 2 KNN confusion matrix

	Class 1	**Class 2**	**Class 3**	**Class 4**		**Recall**
	Predict				**Count**	**TP/Actual**
Actual	445	41	0	1	487	0.91
	74	485	2	3	564	0.86
	1	47	34	1	83	0.41
	0	4	0	96	100	0.96
Count	520	577	36	101	1234	

This table shows that this technique failed to predict the third classifier because the true positive is 34 and the actual is 83. This gives 0.41 TP/actual. At the same time, it succeeds strongly in the fourth classifier, which appears from TP/actual, which is 0.96. To take a deeper look at the results, different classification matrices are calculated as shown in the following table 4.3.

Table 4. 3 KNN classification matrices

	Class 1	**Class 2**	**Class 3**	**Class 4**
TP	445	485	34	96
TN	672	578	1149	1129
FP	75	92	2	5
FN	42	79	49	4
(Recall)	0.913758	0.859929	0.409639	0.96
False Negative Rate	0.086242	0.140071	0.590361	0.04
(Precision)	0.855769	0.840555	0.944444	0.950495
False Discovery Rate	0.144231	0.159445	0.055556	0.049505
Specificity	0.899598	0.862687	0.998262	0.995591
FPR = (1-Specificity)	0.100402	0.137313	0.001738	0.004409
Accuracy	85.90%			
F1 score	0.883813	0.850131	0.571429	0.955224
Macro-F1	0.815149			
Macro-(Recall)	0.785831			
Macro-(Precision)	0.897816			
weighted-f1	0.853195			
weighted-Recall	0.858995			
weighted-Precision	0.862456			

As seen from the previous table 4.3, each of the measurement matrices for each class is calculated separately. Since it is a multi-class problem, the system's accuracy is 85.9%, the F1-score reveals that class 3 has the worst result on the classification model. Also, the Macro-F1, which gets the mean of each class of F1-score, is equal to 0.81. Furthermore, the weighted-F1 measurement factor is used because it considers the weight for each class that impacts the F-score result. Because the number of observations in the third and fourth class is few compared with the first and second categories, this, in turn, makes the impact of the error in classification on the system less. This explains the high value of the weighted-f1 compared with Macro-F1 as the results shown above.

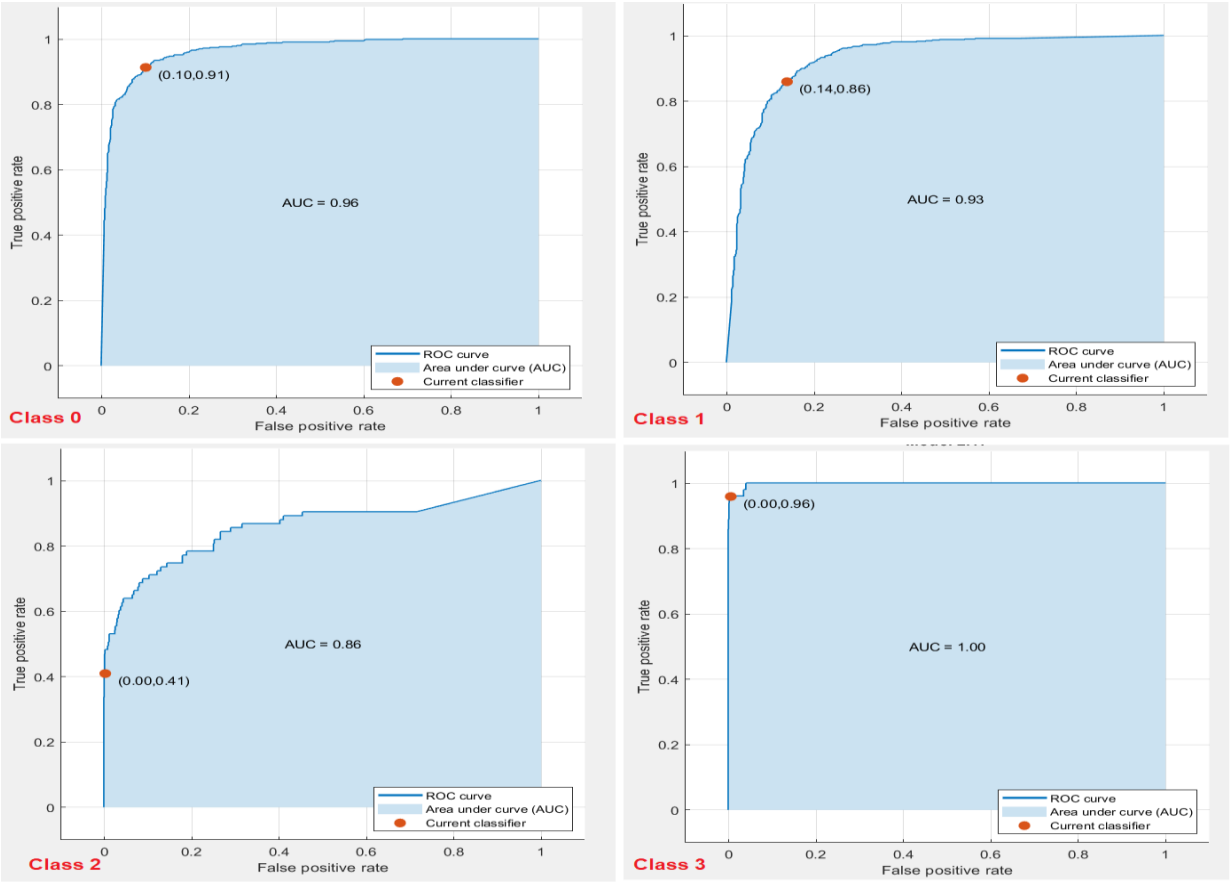


Figure 4. 4 KNN ROC curve

The ROC curve illustrated in figure 4.4, which is drawn between sensitivity vs. specificity for each class, is constructed for each class and makes it a positive and the rest of the classes as a negative, the area under the curve value is calculated and labeled using abbreviation AUC on the middle for each graph, this area appears under the dark blue line in light blue color. By looking at the graph, it is found that the value of the area under the curve in the first class is 0.96, and this indicates that the classification in the first-class falls under the excellent estimate. Also, the sensitivity of the current classifier is 0.91, and 1- specificity is 0.1. The second-class classification also showed excellent results, as the area under the curve reached 0.93, and the sensitivity and 1-specificity values for the current classification reached the values 0.86,0.14, respectively. For the third class, the value of the area under the ordeal was less valuable, it is 0.86, and the sensitivity value for this classifier is 0.4, while the 1-specificity was zero. In the fourth class, the value of the area under the curve is perfect, and this means the AUC reached 1, and the value of sensitivity and 1-specificity is 0.96,0, respectively. Hence, it is noted that the third classification had a fewer classification matrix, as the ROC graph shows. Still, this impact on the whole system is less if weighted-AUC is taken as a classification matrix due to considering the observation weight for each class on his calculations.

4.3.1.2 Linear Discrimination Experiment Result

The linear discriminant technique is used to classify students' majors using the following parameters; the inputs variables are 31, the number of observations is 1234, the validation is 5-fold cross-validation, and the number of classes is 4. After applying this technique, the result appears in the following table 4.4.

Table 4. 4 Linear discrimination confusion matrix

	Class 1	Class 2	Class 3	Class 4		Recall
	Predict				Count	TP/Actual
Actual	470	11	1	5	487	0.965092
	45	481	21	17	564	0.852837
	0	31	52	0	83	0.626506
	4	0	3	93	100	0.93
Count	519	523	77	115	1234	

This technique is better than the KNN technique because there is an enhancement in class 3, the TP is 52, and the actual is 83. This gives 0.62 TP/actual. At the same time, it succeeded strongly in the first classifier. This appears from TP/actual, which is 0.97. Different classification matrices are calculated to take a deeper look at the results, as shown in table 4.5.

Table 4. 5 Linear discrimination classification matrices

	Class 1	Class 2	Class 3	Class 4
TP	470	481	52	93
TN	698	628	1126	1112
FP	49	42	25	22
FN	17	83	31	7
(Recall)	0.965092	0.852837	0.626506	0.93
False Negative Rate	0.034908	0.147163	0.373494	0.07
(Precision)	0.905588	0.919694	0.675325	0.808696
False Discovery Rate	0.094412	0.080306	0.324675	0.191304
Specificity	0.934404	0.937313	0.97828	0.9806
FPR = (1-Specificity)	0.065596	0.062687	0.02172	0.0194
Accuracy	88.82%			
F1 score	0.934394	0.885005	0.65	0.865116
Macro-F1	0.833629			
Macro-(Recall)	0.843609			
Macro-(Precision)	0.827326			
weighted-f1	0.887078			
weighted-Recall	0.888169			
weighted-Precision	0.888695			

As seen from the previous table 4.5, each of the measurement matrices for each class is calculated separately. Since it is a multi-class problem, the system's accuracy is 88.8%. The F1-score reveals that class 3 has the worst result on the classification model. Also, the Macro-F1, which is getting the mean of each class of F1-score, is equal to 83%, which is better than the KNN technique. Furthermore, the weighted-F1 classification is 88.8%.

From figure 4.5, it is found that the value of the area under the curve in the first class is 0.99, which indicates that the classification in the first class falls under the excellent estimate. Also, the sensitivity of the current classifier is 0.97, and 1- specificity is 0.07. This class is the best classification AUC result. The second-class classification also showed excellent results. The area under the curve reached 0.96, and the sensitivity and 1-specificity values for the current classification reached 0.85 0.06, respectively. The FPR in this classification becomes better than the KNN technique value.

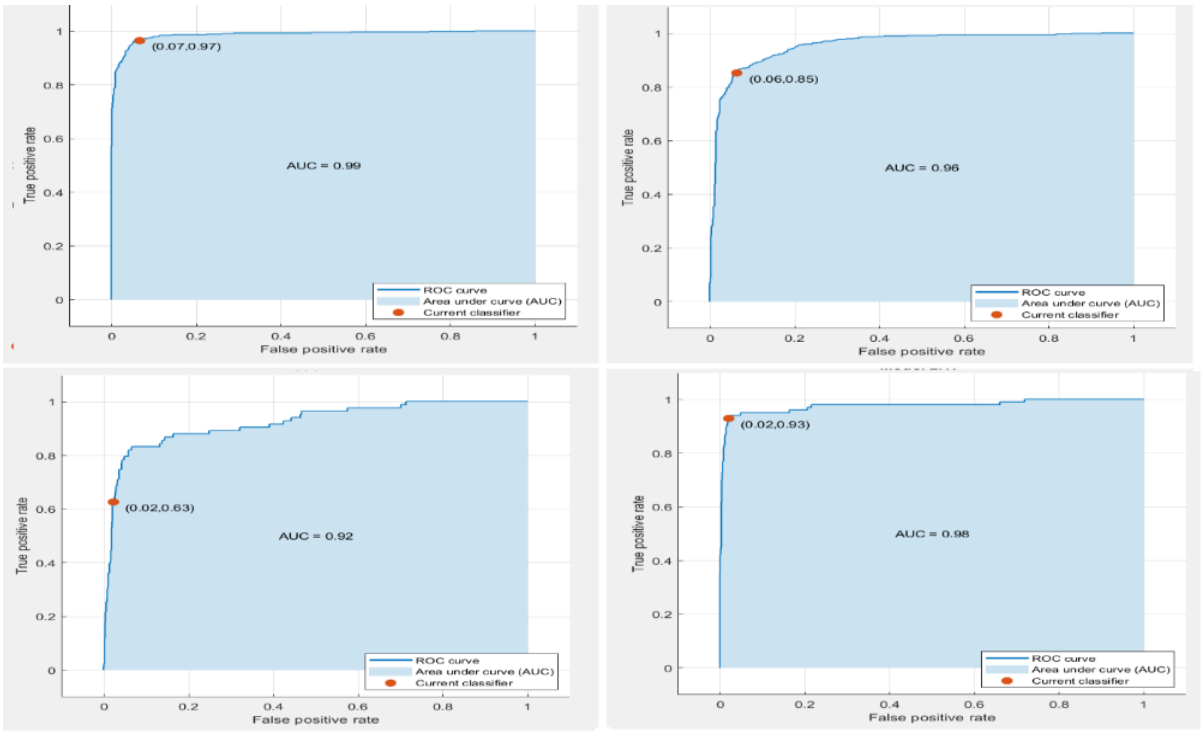


Figure 4. 5 Linear discrimination ROC curve

For the third class, the value of the area under the ordeal was less valuable. It is 0.92, this due to enhanced recall value, but the FPR increases since the number of samples classified on class 3, and it does not belong increased, and the sensitivity value for this classifier is 0.63, while the 1-specificity was 0.02. In the fourth class, the value of the area under the curve value decreases, and the value of sensitivity and 1-specificity is 0.93 0.02, respectively. Hence, it is noted that the third classification had fewer significant values, as the ROC graph shows. Still, this impact on the whole system is less if weighted-AUC is taken as classification matrices due to considering the observation weight for each class on his calculations.

4.3.1.3 SVM Experiment Result

Classifying students' specialization on an SVM algorithm is also applied to check the quality of different machine learning techniques. The inputs variables are 31 with 1234 observations, the validation is 5-fold cross-validation, and the number of classes is 4. After applying this technique, the results appear in the following table 4.6.

Table 4. 6 SVM confusion matrix

	Class 1	**Class 2**	**Class 3**	**Class 4**		**Recall**
	Predict				Count	TP/Actual
Actual	466	18	0	3	487	0.956879
	38	513	6	7	564	0.909574
	0	49	34	0	83	0.409639
	4	10	2	84	100	0.84
Count	508	590	42	94	1234	

This technique is worse than the KNN technique on class 3 classification. Based on its result, the TP is 34, and the actual is 83. This gives 0.40 TP/actual. At the same time, it succeeded strongly in the first classifier, this appears from TP/actual, which is 0.95,

which is better than KNN, and it is worse than Linear discriminant, and for more details appears in table 4.7.

Table 4. 7 SVM classification matrices

	Class 1	**Class 2**	**Class 3**	**Class 4**
TP	466	513	34	84
TN	705	593	1143	1124
FP	42	77	8	10
FN	21	51	49	16
(Recall)	0.956879	0.909574	0.409639	0.84
False Negative Rate	0.043121	0.090426	0.590361	0.16
(Precision)	0.917323	0.869492	0.809524	0.893617
False Discovery Rate	0.082677	0.130508	0.190476	0.106383
Specificity	0.943775	0.885075	0.99305	0.991182
FPR = (1-Specificity)	0.056225	0.114925	0.00695	0.008818
Accuracy	88.90%			
F1 score	0.936683	0.889081	0.544	0.865979
Macro-F1	0.808936			
Macro-(Recall)	0.779023			
Macro-(Precision)	0.872489			
weighted-f1	0.882785			
weighted-Recall	0.888979			
weighted-Precision	0.88629			

The accuracy of the SVM technique is 88.9%, the enhancement is different from previous techniques, which are 0.1, the F1-score for class 3 remains the worst result. Also, the Macro-F1 81%, which is worse than the previous techniques despite accuracy, is better, while weighted-F1 classification is still the same as a linear discriminant technique with an 88.2% value.

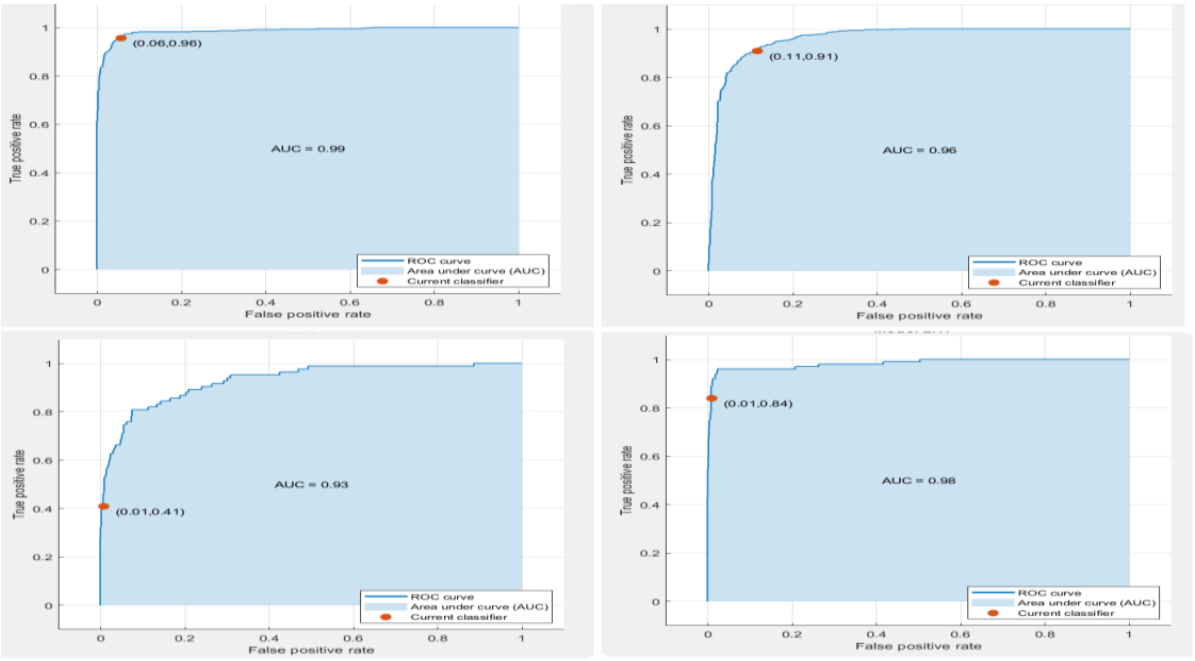


Figure 4. 6 SVM ROC curve

THE SVM ROC curve in figure 4.6 shows that the value of the area under the curve in the first class is 0.99, and this indicates that the classification in the first-class falls under the excellent estimate. Also, the sensitivity of the current classifier is 0.96, and 1-specificity is 0.06. This class is the best classification AUC result. The second-class classification also showed excellent results. The area under the curve reached 0.96, and the sensitivity and 1-specificity values for the current classification reached 0.91 0.11, respectively. The FPR in this classification becomes better than the value of the previous techniques.

For the third class, the value of the area under the ordeal was less valuable. It is 0.93 due to an enhanced recall value, but the FPR increases since the number of samples classified in class 3 and does not belong increased, and the sensitivity value for this classifier is 0.41, while the 1-specificity was 0.01. In the fourth class, the value of the area under the curve value is 0.98, and the value of sensitivity and 1-specificity is 0.84

0.01, respectively. Hence, it is noted that the third classification had a fewer classification matrix, as the ROC graph shows. Still, this impact on the whole system is less if weighted-AUC is taken as classification matrices due to considering the observation weight for each class on his calculations.

4.3.1.4 Decision Tree Experiment Result

This technique is similar in its construction to the tree model so that its nodes represent a test of the features so that their results appear on the branches, and that each end of leaf of it bears a name from the existing names of the class, it has 31 input features with 1234 observation, the validation is 5-fold cross-validation, and the number of classes is 4. after applying the algorithm, the result appears on the following table 4.8.

Table 4. 8 SVM confusion matrix

	Class 1	**Class 2**	**Class 3**	**Class 4**		**Recall**
	Predict				Count	TP/Actual
Actual	452	30	2	3	487	0.928131
	31	488	24	21	564	0.865248
	0	8	75	0	83	0.903614
	2	2	0	96	100	0.96
Count	485	528	101	120	1234	

This algorithm produces better results than all previously implemented techniques, especially for class 3. This appears on its result, the true positive is 75, and the actual is 83, this gives 0.9 TP/actual, while it succeeded strongly in the fourth classifier, this appears from TP/actual which is 0.96, for more detail see the table 4.9 below.

Table 4. 9 DT classification matrices

	Class 1	Class 2	Class 3	Class 4
TP	452	488	75	96
TN	714	630	1125	1110
FP	33	40	26	24
FN	35	76	8	4
Recall	0.928131	0.865248	0.903614	0.96
False Negative Rate	0.071869	0.134752	0.096386	0.04
Precision	0.931959	0.924242	0.742574	0.8
False Discovery Rate	0.068041	0.075758	0.257426	0.2
Specificity	0.955823	0.940299	0.977411	0.978836
FPR = (1-Specificity)	0.044177	0.059701	0.022589	0.021164
Accuracy	90.03%			
F1 score	0.930041	0.893773	0.815217	0.872727
Macro-F1	0.87794			
Macro-(Recall)	0.914249			
Macro-(Precision)	0.849694			
weighted-f1	0.901097			
weighted-Recall	0.900324			
weighted-Precision	0.905			

The accuracy of the decision tree is 90.03%, the F1-score for class 3 jumps to 0.8. also, the Macro-F1jamp to 87%, while weighted-F1 classification reached 0.9, these results of classification matrices judge that decision tree classification succeeds to classify system compared with all previous techniques, and this appears on ROC curve graph and agrees with these results, for more detail see figure 4.1.

As shown in the illustrated ROC in figure 4.7, the values of the area under the curve in the first class is 0.96, the sensitivity of the current classifier is 0.93, and 1- specificity is 0.04. For the second-class classification, the area under the curve reached 0.94, and the sensitivity and 1-specificity values for the current classification reached 0.87 0.06, respectively.

The third class shows a good enhancement in AUC value compared with all previous results. The value of the area under the curve is 0.97, which is due to enhanced recall and specificity values. The sensitivity value for this classifier is 0.90, while the 1-specificity was 0.02. In the fourth class, the value of the area under the curve value is 0.97, and the value of sensitivity and 1-specificity is 0.96 0.02, respectively

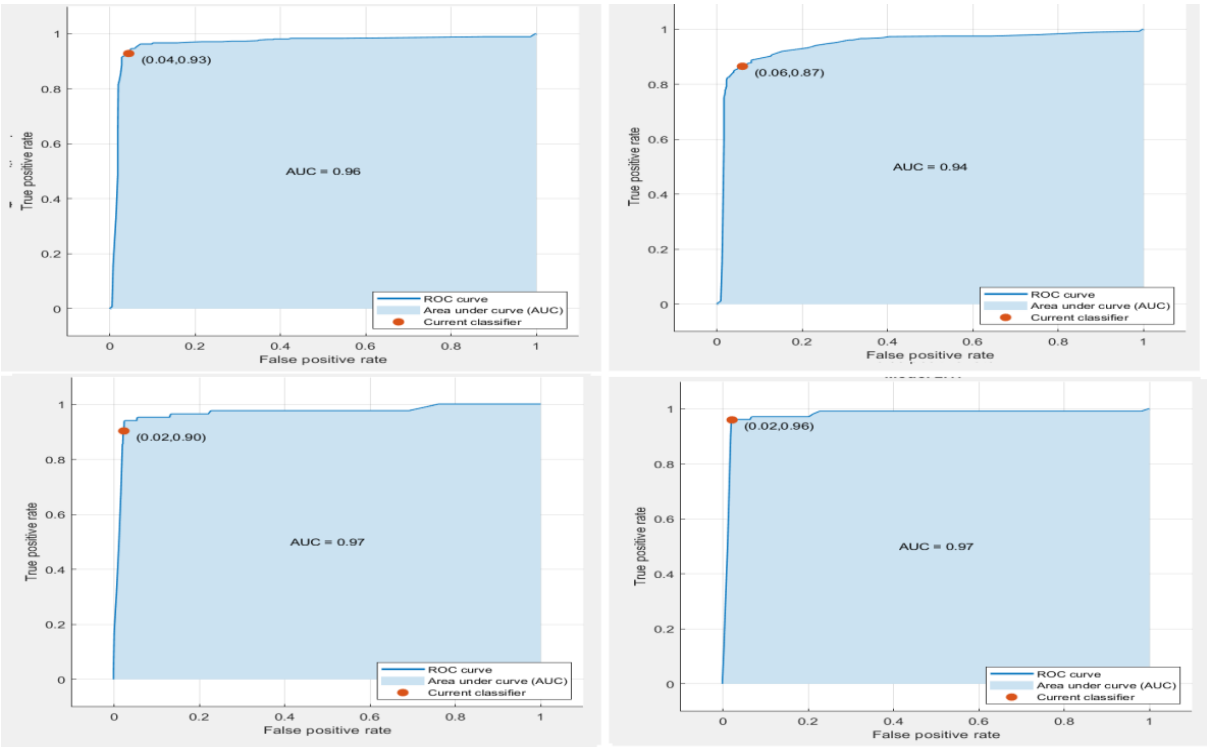


Figure 4.7 DT ROC curve

4.3.1.5 Ensembled Boosted Tree Experiment Result

Undoubtedly, using a group of combined algorithms requires more complex mathematical operations and more mathematical operations in evaluating the model and checking the integrity of the prediction. This construction improves tree algorithms by merging them to obtain high accuracy in classifications. To check the quality of this machine learning, classifying students' majors on an Ensembled boosted tree algorithm

was also applied. Its inputs are 31 with 1234 observations, the validation is 5-fold cross-validation, and the number of classes is 4. After applying the algorithm, the result appears on the following table 4.10

Table 4. 10 ES boosted tree confusion matrix

	Class 1	**Class 2**	**Class 3**	**Class 4**		**Recall**
	Predict				**Count**	**TP/Actual**
Actual	463	18	3	3	487	0.950719
	34	497	22	11	564	0.881206
	0	8	75	0	83	0.903614
	2	3	0	95	100	0.95
Count	499	526	100	109	1234	

Appears that this technique has better matrices classification values for the whole classes. Class two is the worst because the TP is 497 and the actual is 564, and this gives 0.88 TP/actual. At the same time, it succeeded strongly in the first and fourth classifiers. This appears from TP/actual, which is 0.95 for both. To take a deeper look at the results, different classification matrices are calculated, as shown in the table. 4.11

Table 4. 11 ES boosted tree classification matrices

	Class 1	**Class 2**	**Class 3**	**Class 4**
TP	463	497	75	95
TN	711	641	1126	1120
FP	36	29	25	14
FN	24	67	8	5
(Recall	0.950719	0.881206	0.903614	0.95
False Negative Rate	0.049281	0.118794	0.096386	0.05
(Precision	0.927856	0.944867	0.75	0.87156
False Discovery Rate	0.072144	0.055133	0.25	0.12844
Specificity	0.951807	0.956716	0.97828	0.987654
FPR = (1-Specificity)	0.048193	0.043284	0.02172	0.012346
Accuracy	91.57%			
F1 score	0.939148	0.911927	0.819672	0.909091
Macro-F1	0.894959			
Macro-(Recall)	0.921385			
Macro-(Precision)	0.873571			

weighted-f1	0.916235			
weighted-Recall	0.915721			
weighted-Precision	0.919106			

The accuracy of the ensembled boosted tree technique is 91.57%, the F1-score remains the worst result for class 3 .this return to miss prediction for this class. It appears on the number of predicted observations which is 100, which is too high compared to the actual, which is 83, also, the Macro-F1 89% which is better than previous techniques. In comparison, weighted-F1 classification is still the same as the previous algorithm with a 91.2% value, so from previous, it is seen that the ensembled technique enhanced classification matrices.

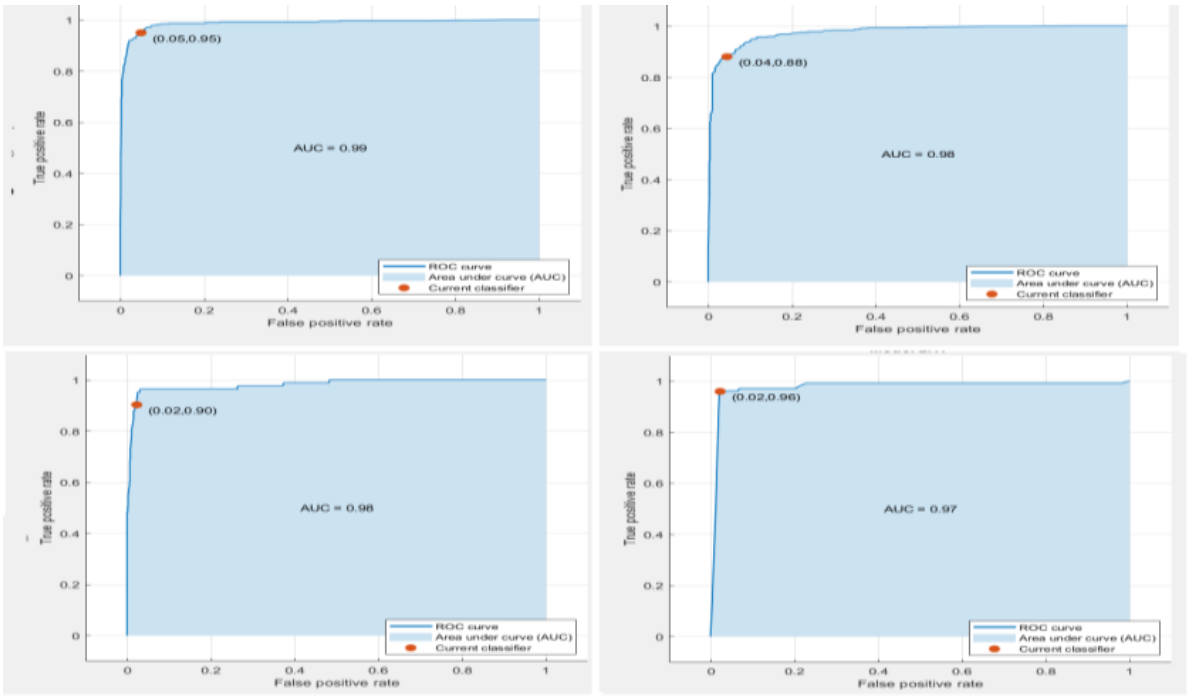


Figure 4. 8 ES boosted tree ROC curve

As shown in figure 4.8, it is found that the value of the area under the curve in the first class is 0.99, and this indicates that the classification in the first class falls under the excellent estimate. Also, the sensitivity of the current classifier is 0.95, and 1-

specificity is 0.05. This class is the best classification AUC result. The second-class classification also showed excellent results. The area under the curve reached 0.98, sensitivity and 1-specificity values for the current classification reached 0.88 0.04, respectively. The FPR in this classification becomes better than the previous algorithm results. For the third class, AUC is 0.98, and the sensitivity value is 0.90, while the 1-specificity was 0.02. In the fourth class, the area value under the curve value is 0.97, and the sensitivity and 1-specificity are 0.96 0.02, respectively. The worst AUC class is class 4, but it is accepted since the AUC value is 0.97.

4.3.1.6 Ensembled Bagged Tree Experiment Result

As mentioned before in chapter 3, bagging separates the dataset into sub-datasets with slight changes in bootstrapping. Bootstrap aggregating is carried out to understand the distribution of data. This experiment will check this algorithm's capability to classify students' majors. The inputs are 31 with 1234 observations, the validation is 5-fold cross-validation, and the number of classes is 4. After applying this technique, the result appears in the following table 4.12.

Table 4. 12 ESbagged tree confusion matrix

	Class 1	**Class 2**	**Class 3**	**Class 4**		**Recall**
	Predict				**Count**	**TP/Actual**
Actual	469	17	0	1	487	0.963039
	29	528	3	4	564	0.93617
	0	21	62	0	83	0.746988
	0	0	1	99	100	0.99
Count	498	566	66	104	1234	

This technique produces a result worse than the ensembled boosted tree on class 3 classification. Based on its result, the TP is 62, and the actual is 83. This gives 0.74

TP/actual. At the same time, it succeeded strongly in the second classifier. This appears from TP/actual, which is 0.93, better than Decision tree and boosted tree, and for more details, look to table 4.13.

The accuracy of the ensembled boosted tree is 93.84%. This result is the best, but the F1-score remains bad for class 3. Also, the Macro-F1 92%, which is worse than the previous algorithm despite accuracy, is better.Weighted-F1 is still the same as ES Bagged tree techniques with 93.7% value, this technique also has the best f1-score value for both Macro-F1 and weighted-F1.

Table 4. 13 ES bagged tree classification matrices

	Class 1	**Class 2**	**Class 3**	**Class 4**
TP	469	528	62	99
TN	718	632	1147	1129
FP	29	38	4	5
FN	18	36	21	1
Recall	0.963039	0.93617	0.746988	0.99
False Negative Rate	0.036961	0.06383	0.253012	0.01
Precision	0.941767	0.932862	0.939394	0.951923
False Discovery Rate	0.058233	0.067138	0.060606	0.048077
Specificity	0.961178	0.943284	0.996525	0.995591
FPR = (1-Specificity)	0.038822	0.056716	0.003475	0.004409
Accuracy	93.84%			
F1 score	0.952284	0.934513	0.832215	0.970588
Macro-F1	0.9224			
Macro-(Recall)	0.909049			
Macro-(Precision)	0.941487			
weighted-f1	0.937569			
weighted-Recall	0.938412			
weighted-Precision	0.93836			

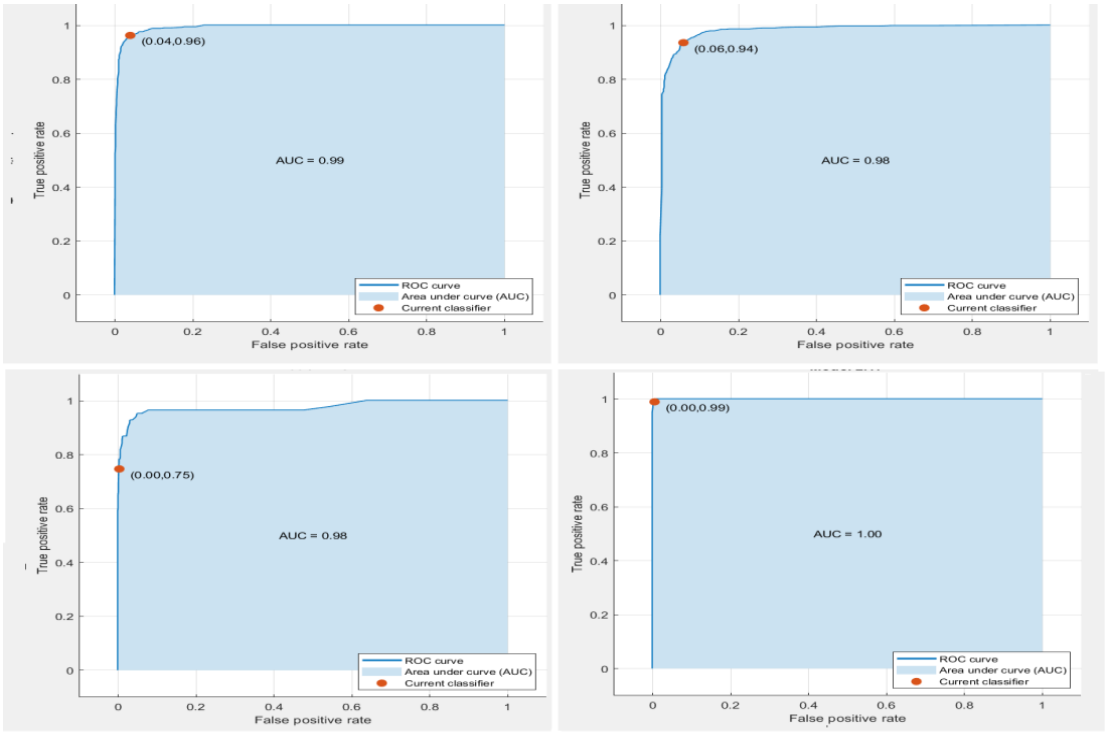


Figure 4. 9 ES bagged tree ROC curve

From the ROC figure 4.9, the FPR rate is less for whole classes, which means they miss prediction decreases, it is found that the value of the area under the curve in the first class is 0.99, and this indicates that the classification in the first-class falls under the excellent estimate. Also, the sensitivity of the current classifier is 0.96, and 1-specificity is 0.04. for the second-class classification, it also showed excellent results. The area under the curve reached 0.98, and the sensitivity and 1-specificity values for the current classification reached 0.94,0.06, respectively. The FPR in this classification becomes better than the value of the previous techniques. For the third class, it is 0.98, and the sensitivity value for this classifier is 0.75, while the 1-specificity was 0.0. In the fourth class, the area value under the curve value one and the value of sensitivity and 1-specificity is 0.99 0.0, respectively.

4.3.2 Machine Learning Classification Learner Comparison

This section compares the machine learning techniques with what is previously reviewed. A comparison is made between these techniques depending on the classification matrices, and also the important classification matrices were visualized using charts that give a clearer picture of the most important differences and what distinguishes each technique, as a result of comparing the results, the technique that is the most successful in classifying students' majors is determined.

To make a fair comparison, a table containing the most important classification matrices for each technique is created; this makes the comparison procedure easy and fair. See Table 4.14.

It is noted from the above table 4.14 that the accuracy value of the Ensembled bagged tree algorithm is the highest, as shown in the following figure 4.10.

Table 4. 14 Classification learner comparison

	KNN	Linear Deterministic	SVM	DT	Ensembled boosted tree	Ensembled bagged tree
Accuracy	85.90%	88.82%	88.90%	90.03%	91.57%	93.84%
F1 score class 1	0.88381	0.93439	0.93668	0.930041	0.939148	0.952284
F1 score class 2	0.85013	0.88501	0.88908	0.893773	0.911927	0.934513
F1 score class 3	0.57143	0.65	0.544	0.815217	0.819672	0.832215
F1 score class 4	0.95522	0.86512	0.86598	0.872727	0.909091	0.970588
Macro-F1	0.81515	0.83363	0.80894	0.87794	0.894959	0.9224
Macro-(Recall)	0.78583	0.84361	0.77902	0.914249	0.921385	0.909049
Macro-(Precision)	0.89782	0.82733	0.87249	0.849694	0.873571	0.941487
weighted-f1	0.8532	0.88708	0.88279	0.901097	0.916235	0.937569
weighted-Recall	0.859	0.88817	0.88898	0.900324	0.915721	0.938412
weighted-Precision	0.86246	0.8887	0.88629	0.905	0.919106	0.93836

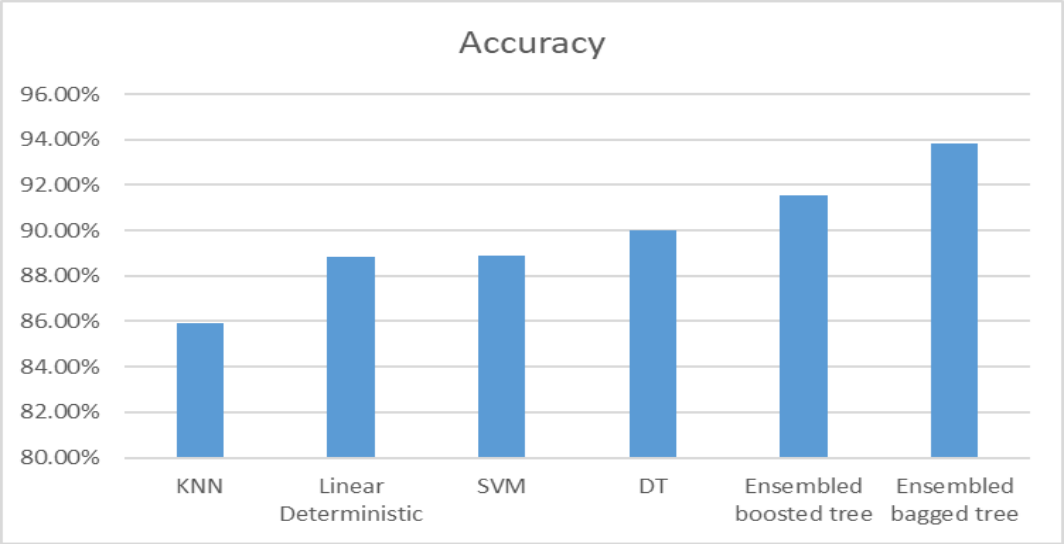


Figure 4. 10 Classification learner comparison chart

The rating values differed among the techniques, note figure 4.11. It is noted that the values of the recall parameter are improving in the weight-recall of the macro-recall. This is reflected in the values of the Weight-f1 because the weights of the observation are taken into account, see figure 4.12.

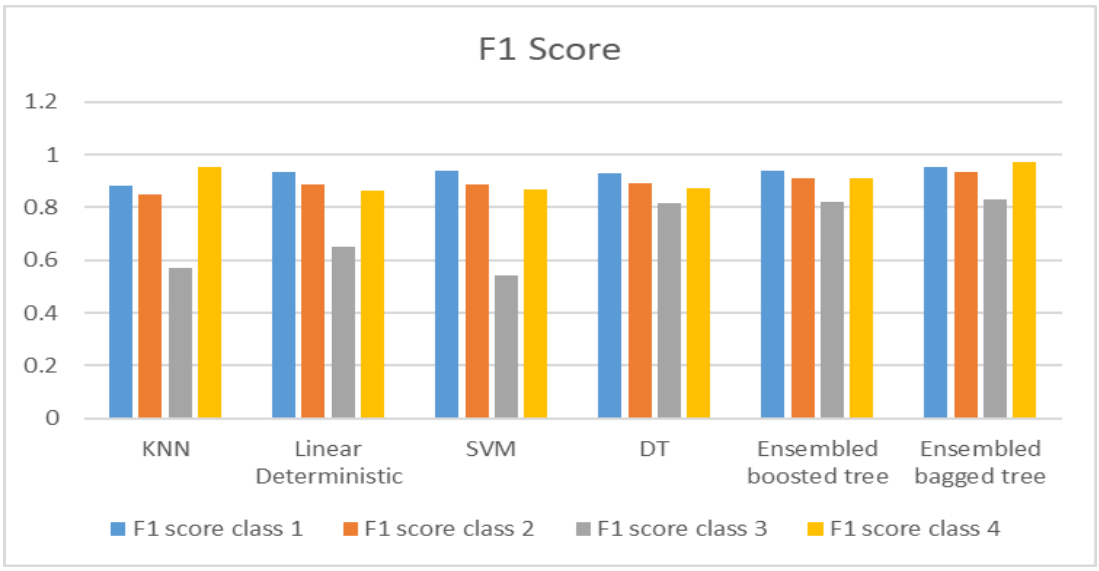


Figure 4. 11 F1-score classification learner comparison chart

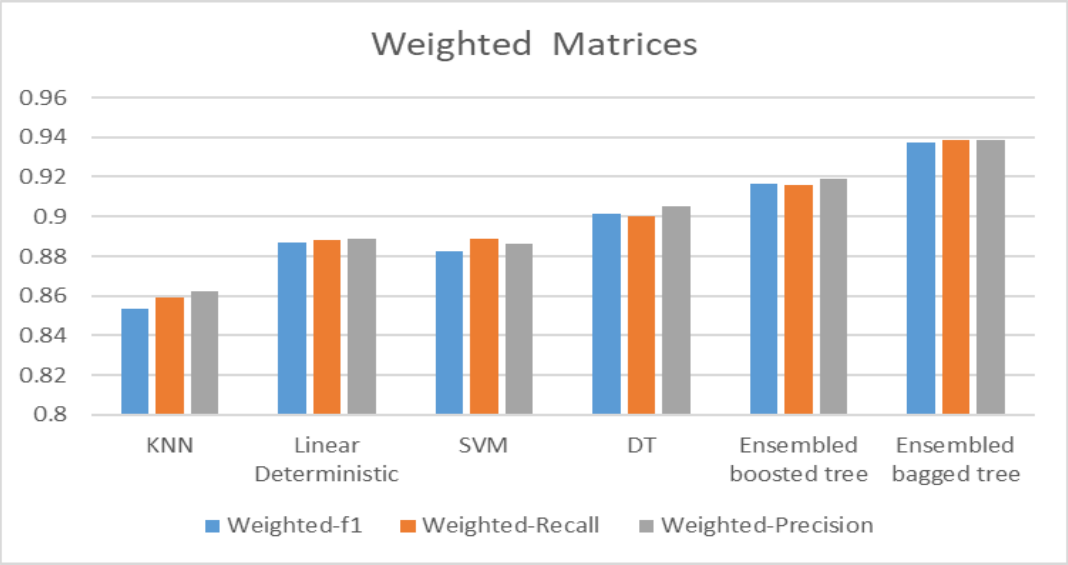


Figure 4. 12 Weight classification learner comparison chart

The difference between the weight values in the classification matrices in some classification's values has to be discussed. It is noticed that the difference between Weighted-f1 and Macro -F1 in the SVM technique is large, and this is due to the error in the classification of the third class. While the difference between Weighted-f1 Macro -F1 in the Ensembled bagged tree technique is small, this returns to the high classification matrices on whole classes, note figure 4.13.

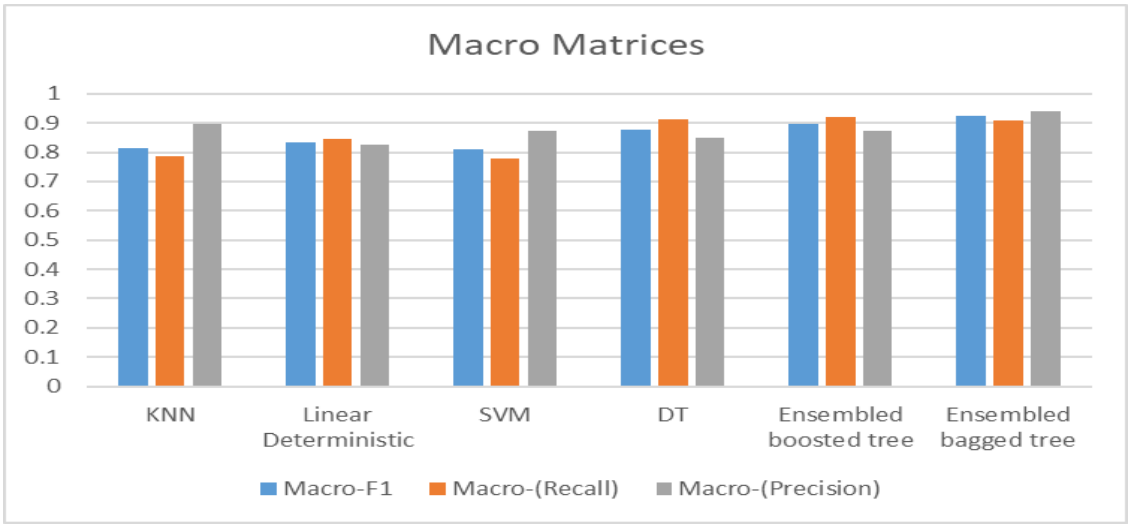


Figure 4. 13 Macro learner classification comparison

If it is assumed that the weights of the samples Neglected and the classification coefficients are the same, the macro-F1 is the most accurate. Still, if the weights of the samples are important in the classification, the weight-F1 is the closest to reality and the fairest. Note the following figure 4.14. In our case, the weight-F1 is more suitable for classification.

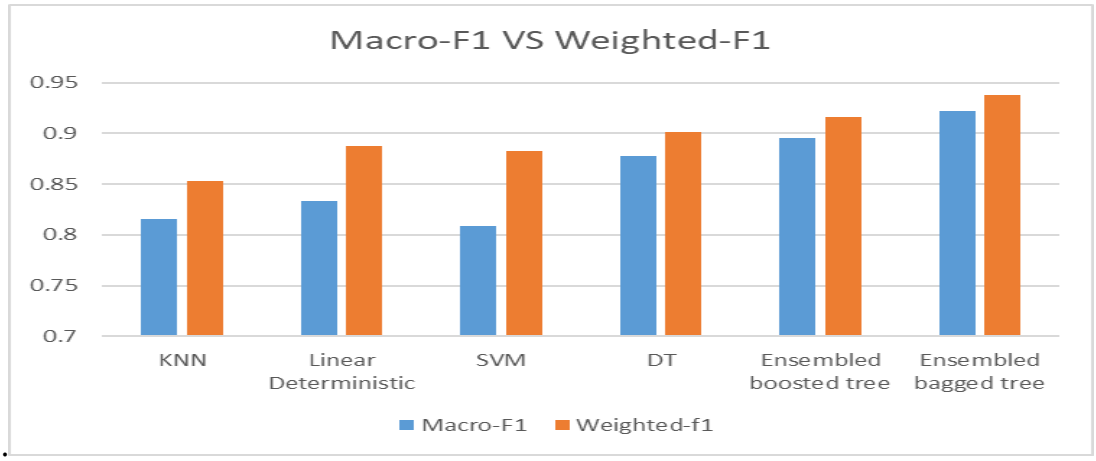


Figure 4. 14 Macro vs. weight learner classification comparison chart

4.4 Multi-Layer Perceptron Neural Networks (MLPNNs)

Neural networks are one of the virtual machine learning techniques, and they are characterized by their high ability to classify, so they are also used in classification. The classification results using NNs were the best from all classification models whose results were previously reviewed. In this section, the results are reviewed that are obtained from the application of neural networks to the different number of different NNs, which are in the order of 5, 10, 15, and 20, and then compare the results to determine the number of neurons needed to obtain the best classification to predict the students' majors. When five neurons were used, the result that appeared is unsatisfactory, and it is noted that the classification in the industrial class is terrible.

When ten neurons were used, the results are entirely different. In the third stage, when 15 neurons were used, all classification algorithms were the best. The tests are still performed to ensure that if 20 neurons are used, can the best accurate results be obtained?

Unfortunately, the system's accuracy decreased, so a comparison is made for a fair judgment and differentiation between applying the number of different neurons, as shown in the following table 4.15.

Table 4. 15 MLP 5,10,15,20 neuron.comparison

classification Matrices	5 Neuron	10 Neurons	15 neurons	20 neurons
Accuracy	88.41%	95.22%	96.92%	94.17%
Macro-F1	0.697241	0.955789	0.963855	0.937248
Macro-(Recall)	0.915272	0.960298	0.963601	0.939851
Macro-(Precision)	0.716511	0.952006	0.964474	0.935237
weighted-f1	0.919212	0.952183	0.969283	0.941728
weighted-Recall	0.884117	0.952188	0.969206	0.941653
weighted-Precision	0.959219	0.953162	0.969495	0.942656

A graph is created to compare the accuracy of changing the number of neurons in neural networks to clear the difference. The superiority of the fifteen-neuron neural network is noticed by looking closely, as illustrated in figure 4.15.

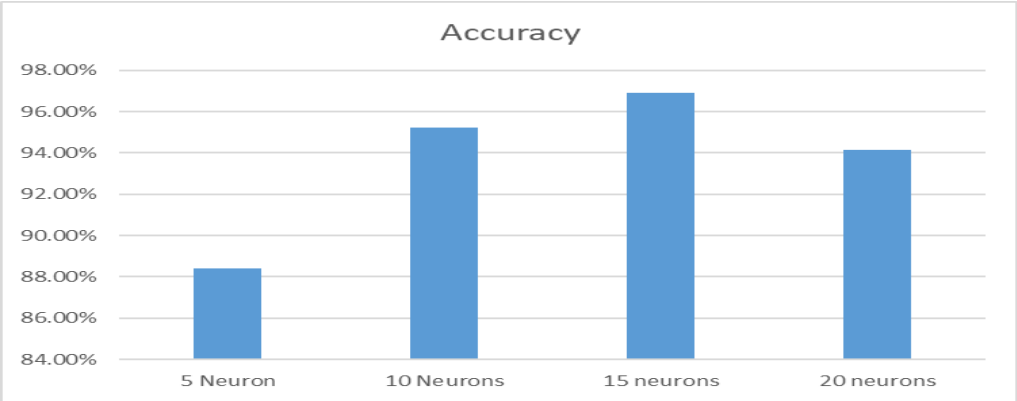


Figure 4. 15 Accuracy comparison between different MLP neuron

Also, a graph is made on weighted-F1 to compare the neural networks. The weighted-F1 is taken because it considers each class's weight separately, and as it is seen in figure 4.16, the superiority of 15neuron neural networks over other neural networks.

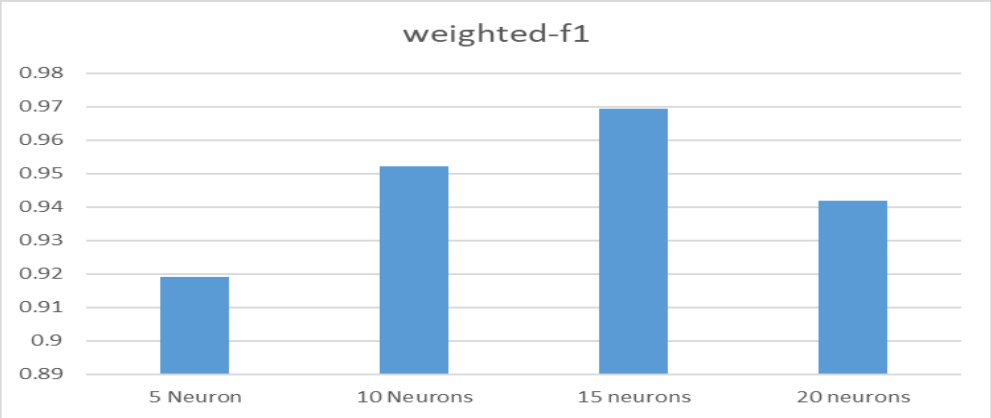


Figure 4. 16 Weight-F1 comparison between different MLP neuron

If the weight of each class in the classification is neglected, then the use of the macro-F1 is ideal in this case. Checking figure 4.17 below, it is noted the weakness of neural networks with five neurons for classification and the superiority of neural networks with fifteen neurons.

As evident below, the best values of the classification matrices are when applying a neural network with 15 neurons.

To highlight the most successful neural network in classification, the neural network consisting of 15 neurons is discussed in detail; the figure 4.18 shows the classifications for classes accurately.

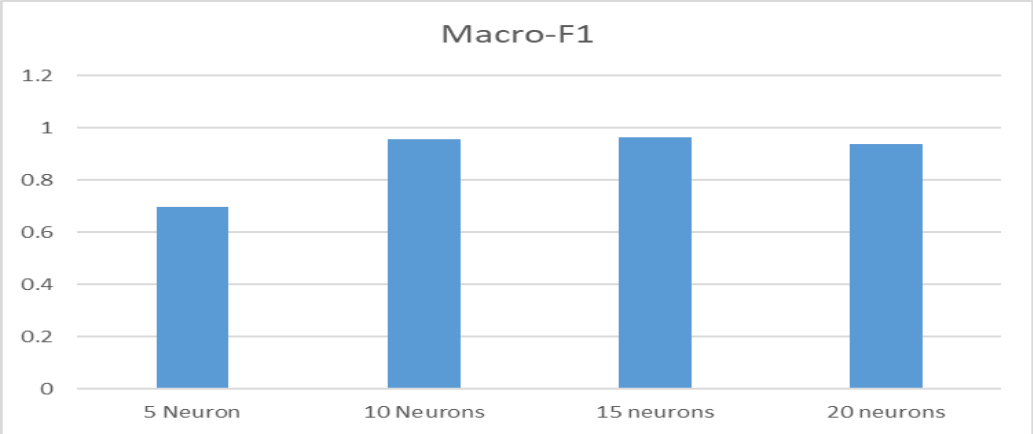


Figure 4. 17 Macro comparisons between different MLP neuron

To have a comprehensive view of the outputs of classification matrices, the total result of the confusion matrices is used from each of the training, valid, and test confusion matrix, and calculation operations are carried out. Note table 4.16.

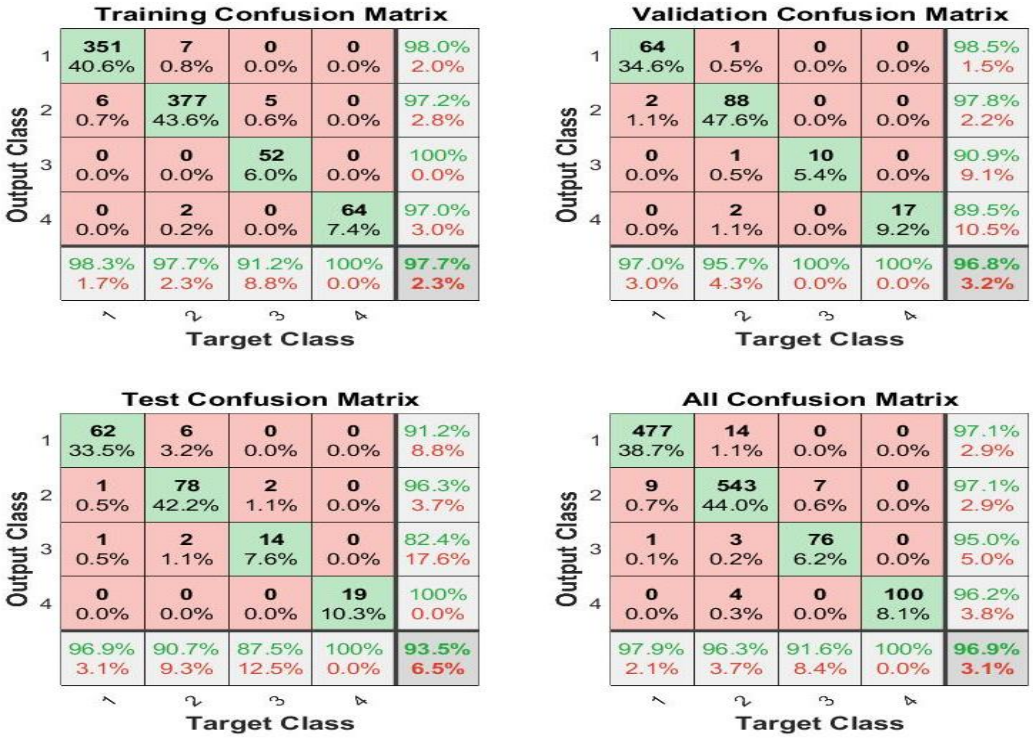


Figure 4.18 MLP 15 neuron confusion matrix

Table 4.16 MLP 15 neuron classification matrices

	Class 1	Class 2	Class 3	Class 4
TP	477	543	76	100
TN	733	654	1147	1130
FP	10	21	7	0
FN	14	16	4	4
Recall	0.971487	0.971377	0.95	0.961538
False Negative Rate	0.028513	0.028623	0.05	0.038462
Precision	0.979466	0.962766	0.915663	1
False Discovery Rate	0.020534	0.037234	0.084337	0
Specificity	0.986541	0.968889	0.993934	1
FPR = (1-Specificity)	0.013459	0.031111	0.006066	0
Accuracy	96.92%			
F1 score	0.97546	0.967053	0.932515	0.980392
Macro-F1	0.963855			
Macro-(Recall)	0.963601			
Macro-(Precision)	0.964474			
weighted-f1	0.969283			
weighted-Recall	0.969206			
weighted-Precision	0.969495			

Note that the classification accuracy is 96.92%, and when looking at other classification matrices, it is seen that the Macro-F1 value is 0.963, and the result of weighted-f1 is 0.969. it is noticed here that the classification values of weighted-f1 returned to rise from the value of classification Macro-F1. This indicates that the system succeeded in classifying the first class, representing more weight than the third and fourth class.

The following ROC curve shows the ability of this system to classify the students' majors well, as the values of the area under the curve were more than ninety-five percent. In contrast, the importance of sensitivity and 1-specificity were as follows: The value of sensitivity and 1-specificity for the first category is 0.97, 0.013, for the second class it is 0.971,0.031, for the third class, the value is 0.95,0.006, and for the fourth class, the value sensitivity and 1-specificity is 0.96,0.0

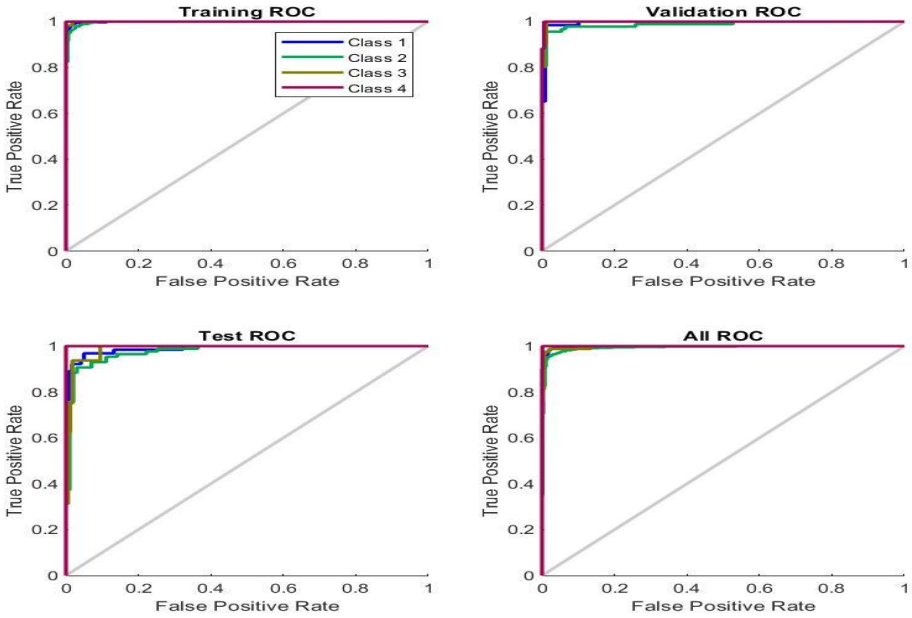


Figure 4.19 MLP 15 neuron ROC curve

4.5 MLPNNs VS Ensembled Bagged tree on Classification

This section compares the classification results produced by MLPNNs with the learner classification methods. The best neural network that gives results is selected; the best classification result is obtained when the neural networks consist of 15 neurons. Also, according to the results of the learner classification algorithm, the best one is the ensembled bagged tree.

The following table 4.17 shows the essential accuracy to classify the results accurately. It shows the superiority of neural networks in all learner classification algorithms chosen.

Table 4. 17 ES VS MLP 15 neuron

Classification Matrices	MLP 15 Neurons	ES
Accuracy	96.92%	93.84%
Macro-F1	0.963855	0.9224
Macro-(Recall)	0.963601	0.909049
Macro-(Precision)	0.964474	0.941487
weighted-f1	0.969283	0.937569
weighted-Recall	0.969206	0.938412
weighted-Precision	0.969495	0.93836

Furthermore, there is a significant difference in the values of classification coefficients, which appears clearly on weighted-F1 and macro-F1. These matrices emphasize the superiority of neural networks.

4.6 Specialist Dataset VS Marks Dataset Result and Comparison

In this section, a dataset consisting of students' marks only is adopted. The best algorithms are used in the same conditions that give a high-accuracy rating for classification when the input data is the dataset that specialists approve. The ensembled bagged tree algorithm is used to classify students' majors. The inputs are 11, the number of observations is 1234, the validation is 5-fold cross-validation, and the number of classes is 4. after applying the algorithm, the result appears in table 4.18.

Table 4. 18 ES confusion matrix for marks dataset

	Class 1	Class 2	Class 3	Class 4		Recall
	Predict				Count	TP/Actual
Actual	415	65	1	6	487	0.852156
	57	501	3	3	564	0.888298
	4	54	24	1	83	0.289157
	18	14	2	66	100	0.66
Count	494	634	30	76	1234	

This algorithm failed to predict the third classifier well from the first look because the true positive is 24 and the actual is 83. This gives 0.41 TP/actual, and to take a deeper look at the results, different classification matrices are calculated, as shown in Table 4.19 below.

As seen from the previous table, each of the measurement matrices for each class is calculated separately. Since it is a multi-class problem, the system's accuracy is 81.52%. The F1-score reveals that class 3 has the worst result on the classification model. Also, the Macro-F1, which gets the mean of each class of F1-score, is equal to 0.71. Furthermore, weighted-F1 is equal to 0.81. Because the number of observations in the third and fourth class is few compared with the first and second categories, this, in turn, makes the impact of the error in classification on the system less. Also, the area under the curve gives lower values than it is previously

Table 4. 19 ES classification matrices for marks dataset

	Class 1	**Class 2**	**Class 3**	**Class 4**
TP	415	501	24	66
TN	668	537	1145	1124
FP	79	133	6	10
FN	72	63	59	34
Recall	0.852156	0.888298	0.289157	0.66
False Negative Rate	0.147844	0.111702	0.710843	0.34
Precision	0.840081	0.790221	0.8	0.868421
False Discovery Rate	0.159919	0.209779	0.2	0.131579
Specificity	0.894244	0.801493	0.994787	0.991182
FPR = (1-Specificity)	0.105756	0.198507	0.005213	0.008818
Accuracy	81.52%			
F1 score	0.846075	0.836394	0.424779	0.75
Macro-F1	0.714312			
Macro-(Recall)	0.672403			
Macro-(Precision)	0.824681			
weighted-f1	0.805528			
weighted-Recall	0.815235			
weighted-Precision	0.816893			

By looking at figure 4.20, it is found that the valueof the area under the curve in the first class is 0.95, and this indicates that the classification in the first-class falls under the excellent estimate. Also, the sensitivity of the current classifier is 0.85, and 1-specificity is 0.1. the second-class classification also showed results, as the area under the curve reached 0.92, and the sensitivity and 1-specificity values for the current classification reached 0.89,0.2, respectively.

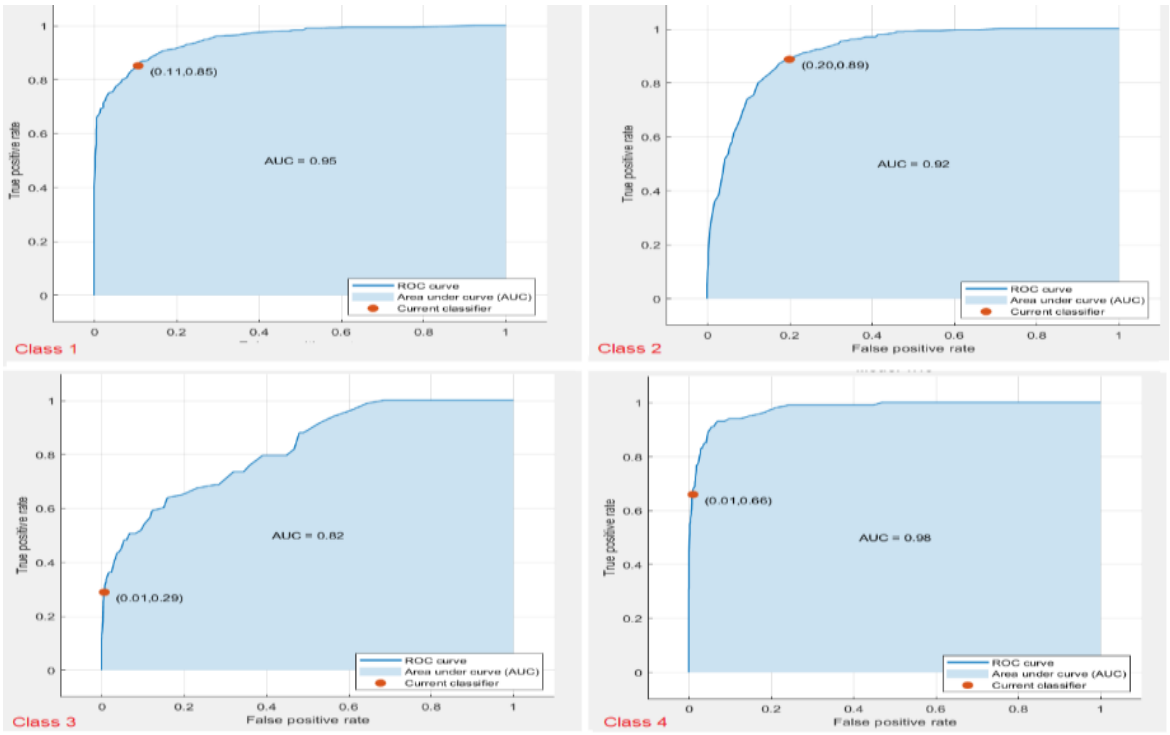


Figure 4.20 ES ROC curve for marks dataset

For the third class, the value of the area under the ordeal was less valuable, it is 0.82, and the sensitivity value for this classifier is 0.29, while the 1-specificity was zero. In the fourth class, the value of the area under the curve is 0.96, and the value of sensitivity and 1-specificity is 0.66,0.1, respectively.

Also, MLPNNs are used, 15 neurons are used, and the result is unsatisfactory. See figure 4.21.

To have a comprehensive view of the outputs of classification matrices, the total result of the confusion matrices is used from each of the training, valid, and test confusion matrix, and calculation operations are carried out. Note table 4.21.

Figure 4. 21 MLP 15 neuron confusion matrix for marks dataset

Table 4. 20 MLP 15 neuron confusion matrix for marks dataset

	Class 1	Class 2	Class 3	Class 4
TP	418	497	0	0
TN	619	479	1151	1134
FP	69	67	83	100
FN	128	191	0	0
(Recall)	0.765568	0.722384	Unknown	Unknown
False Negative Rate	0.234432	0.277616	Unknown	Unknown
Precision	0.858316	0.881206	0	0
False Discovery Rate	0.141684	0.118794	1	1
Specificity	0.899709	0.877289	0.932739	0.918963
FPR = (1-Specificity)	0.100291	0.122711	0.067261	0.081037
Accuracy	74.15%			

F1 score	0.809293	0.79393	Unknown	Unknown
Macro-F1	Unknown			
Macro-(Recall)	Unknown			
Macro-(Precision)	0.43488			
weighted-f1	Unknown			
weighted-Recall	Unknown			
weighted-Precision	0.871078			

Note that the classification accuracy is 75.41%, and also looking at the other classification factors are as follows: In terms of Macro-F1, the value is unknown, and the result of weighted-f1 is unknown. This is due to failed MLPNNs to classify the third and fourth classes, and this gives a bad result from an ensembled bagged tree on the same dataset, which is also worse compared with the dataset used in the previous test all students information.

From the ROC curve shown in figure 4.22, in which the classification of each class with a special color appeared in it as indicated in the key of the graph, the fourth class shows a weakness in the ability to classify properly, to look deeper at a result. It is noticed that the sensitivity and 1-specificity for each class is obtained as follows: The value of sensitivity and 1-specificity for the first category is 0.76, 0.1, for the second class it is 0.72,0.12, for the third class the value is unknown,0.08, and for the fourth class which is the least fortunate, the value sensitivity and 1-specificity is unknown,0.06.

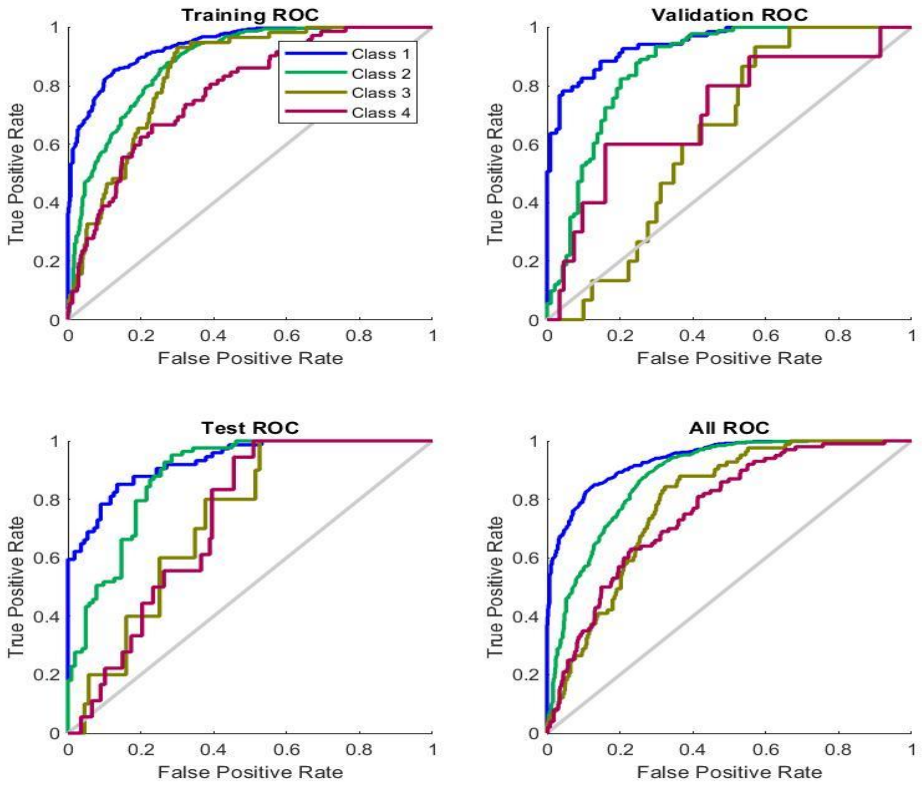


Figure 4. 22 MLPNNs 15 neuron ROC curve for marks dataset

From the above, the neural networks fail to classify the third and fourth categories and a sharp drop in the value of accuracy compared to the precision obtained when applying the dataset under the same conditions. It is concluded that relying on the grade marks alone is impossible to predict their future majors.

4.7 Challenges and Limitation

One of our goals, when this research was established, was to predict the student specialization in universities. Unfortunately, the needed features from the universities cannot be gotten because the necessary information does not exist. the following features are planned to be used: Place of residence - in terms of distance from the university, Branch - industrial science, Nationality, Date of Birth, Chronological age for

obtaining a high school diploma, Gender, Specialization cost/hour, Current Specialization, The date of issuance of the general secondary certificate (to calculate the time difference between the date of issuance of the available secondary certificate and enrollment in the university), Date of registration at the university, High school average, The country of the secondary school, and High school transcripts

The universities do not have all the required information on their database, especially the marks for each curriculum on the high school certificate. The average impact of the high school certificate was only entered into their database. These matters prevented obtaining the information for the classification of majors.

Chapter Five

Conclusion and Future Work

5.1 Conclusion

Early classification of students' enrollment in the literary, scientific, industrial, and commercial academic branches in Palestine would assist decision-makers in setting successful plans. The early category of students' enrollment in the literary, scientific, industrial, and commercial major would assist MOE decision-makers in developing successful strategies. Such as anticipating the necessary infrastructure in education and the number of human cadres from the academic staff needed to adapt to developments and also helps in developing recovery plans to increase numbers in a particular branch and also helps the student to anticipate his future specialization based on the data of his current situation.

Dataset is collected from several directorates and distributed to 3 main divisions: North, Central, and South. The voter turnout in the North was the highest, reaching 45 %, in the middle, it was 20 %, whereas, in the South, 35 %, and the same was true for the eleventh grade. The questionnaire statistics showed that females voted the most. In contrast, students in the eleventh grade voted more than the students of the twelfth grade, and the tendencies of parents and society were for the literary branch. It also showed that now most students spend their time playing and few of them in sports activities. It is noted that 13% of Students study outside their hometowns, and most have enough time to study.

In Palestine, no previous research collects datasets directly from the participants and uses machine learning techniques to classify the student major. The poor information made identifying the necessary features suitable for classification one of the essential stages implemented accurately. As is known, careful selection of the features significantly affects the machine learning technique's ability to classify with high efficiency. The features selection process based on previous studies and recommendations of professionals in this field must be referred.

Different machine learning techniques have proven their ability to categorize disciplines of student majors. Several machine learning algorithms have been used: KNN, SVM, linear discriminant, decision tree, ensembled boosted tree, and ensembled bagged tree and MLPNNs. It is concluded that the ensembled bagged tree shows an excellent classification ability and has a good result compared to the learner classification algorithm. The classification accuracy for this algorithm reached 93.8%, the weight-F1 is 0.93, the precision is 0.94, and the recall is 0.90. In comparison, the power of neural networks to classify was more robust than all learner classification algorithms, where the classification accuracy reached 96.9%, the weight-F1 is 0.96, the precision is 0.96, and the recall is 0.96. The experiments also showed a return in the classification accuracy if only the students' marks were approved, as the classification accuracy in the ensembled bagged tree algorithm reached 81.52%. In comparison, the classification accuracy in the MLPNNs algorithm reached 74.15%.

5.2 Future Work

The future view of this thesis is directed only to increase the efficiency of classification, so future efforts are directed towards establishing a hybrid system. The developed hybrid system should be able to give more accurate classifications.

Attention is also directed to developing a system capable of predicting the future grades of students in high school. Student Grade prediction depends on determining all required features, Especially since many factors affect the student's achievement. Furthermore, sufficient educational data should be collected for more accurate results to achieve the required goals before applying newer prediction algorithms.

One of the future visions is to complete the expectation of the student’s major at the university by making every effort to find mechanisms for integrating the databases in the Palestinian Ministry of Education and universities so that our research can be completed in this field

Bibliography

[1] Luan, H., & Tsai, C. C. (2021). A review of using machine learning approaches for precision education. *Educational Technology & Society*, *24*(1), 250-266.

[2] Aggarwal, D., Mittal, S., & Bali, V. (2021). Significance of Non-Academic Parameters for Predicting Student Performance Using Ensemble Learning Techniques. *International Journal of System Dynamics Applications (IJSDA)*, *10*(3), 38-49.

[3] Gajwani, J., & Chakraborty, P. (2021). Students' Performance Prediction Using Feature Selection and Supervised Machine Learning Algorithms. In *International Conference on Innovative Computing and Communications* (pp. 347-354). Springer, Singapore.

[4] Macalli, M., Navarro, M., Orri, M., Tournier, M., Thiébaut, R., Côté, S. M., & Tzourio, C. (2021). Predicting suicidal thoughts and behaviors among college students: a machine learning approach.

[5] Zeineddine, H., Braendle, U., & Farah, A. (2021). Enhancing prediction of student success: Automated machine learning approach. *Computers & Electrical Engineering*, *89*, 106903.

[6] Maniar, H. Machine Learning Based Student Performance Analysis.

[7] Arcinas, M. M., Sajja, G. S., Asif, S., Gour, S., Okoronkwo, E., & Naved, M. (2021). Role of Data Mining in Education for Improving Students Performance for Social Change. *Turkish Journal of Physiotherapy and Rehabilitation*, *32*(3), 204-226.

[8] Buraimoh, E., Ajoodha, R., & Padayachee, K. (2021, April). Application of Machine Learning Techniques to the Prediction of Student Success. In *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)* (pp. 1-6). IEEE.

[9] Jauhari, F., & Supianto, A. A. (2019). Building student's performance decision tree classifier using boosting algorithm. Indones. J. Electr. Eng. Comput. Sci, 14(3), 1298-1304.

[10] Ahammad, K., Chakraborty, P., Akter, E., Fomey, U. H., & Rahman, S. (2021). A Comparative Study of Different Machine Learning Techniques to Predict the Result of an Individual Student using Previous Performances. International Journal of Computer Science and Information Security (IJCSIS), 19(1).

[11] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Preprocessing and analyzing educational dataset using X-API for improving student's performance," in Applied Electrical Engineering and Computing Technologies (AEECT), 2015 IEEE Jordan Conference on, 2015, pp. 1-5.

[12] B. Kapur, N. Ahluwalia, and R. Sathyaraj, "Comparative Study on Marks Prediction using Data Mining and Classification Algorithms," International Journal of Advanced Research in Computer Science, vol. 8, 2017.

[13] Dutt, A., & Ismail, M. A. (2019, June). Can we predict student learning performance from LMS data? A classification approach. In 3rd International Conference on Current Issues in Education (ICCIE 2018) (pp. 24-29). Atlantis Press.

[14] Lagman, A. C., Alfonso, L. P., Goh, M. L. I., Lalata, J. A. P., Magcuyao, J. P. H., & Vicente, H. N. (2020). Classification algorithm accuracy improvement for student graduation prediction using ensemble model. International Journal of Information and Education Technology, 10(10), 723-727.

[15] Hew, K. F., Hu, X., Qiao, C., & Tang, Y. (2020). What predicts student satisfaction with MOOCs: A gradient boosting trees supervised machine learning and sentiment analysis approach. Computers & Education, 145, 103724.

[16] Fujita, H. (2019). Neural-fuzzy with representative sets for prediction of student performance. Applied Intelligence, 49(1), 172-187.

[17] Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. Computers in Human Behavior, 104, 106189.

[18] Aydoğdu, Ş. (2020). Predicting student final performance using artificial neural networks in online learning environments. Education and Information Technologies, 25(3), 1913-1927.

[19] Hussain, S., Muhsion, Z. F., Salal, Y. K., Theodorou, P., Kurtoglu, F., & Hazarika, G. C. (2019). Prediction Model on Student Performance based on Internal Assessment using Deep Learning. iJET, 14(8), 4-22.

[20] Lau, E. T., Sun, L., & Yang, Q. (2019). Modelling, prediction and classification of student academic performance using artificial neural networks. SN Applied Sciences, 1(9), 1-10.

[21] Mduma, N., Kalegele, K., & Machuve, D. (2019). A survey of machine learning approaches and techniques for student dropout prediction.

[22] Livieris, I. E., Drakopoulou, K., Tampakas, V. T., Mikropoulos, T. A., & Pintelas, P. (2019). Predicting secondary school students' performance utilizing a semi-supervised learning approach. Journal of educational computing research, 57(2), 448-470.

[23] Xu, J., Moon, K. H., & Van Der Schaar, M. (2017). A machine learning approach for tracking and predicting student performance in degree programs. IEEE Journal of Selected Topics in Signal Processing, 11(5), 742-753.

[24] Adewale, A. M., Bamidele, A. O., & Lateef, U. O. (2018). Predictive modelling and analysis of academic performance of secondary school students: Artificial Neural Network approach. International Journal of Science and Technology Education Research, 9(1), 1-8.

[25] Chen, D. (2017). Research on traffic flow prediction in the big data environment based on the improved RBF neural network. IEEE Transactions on Industrial Informatics, 13(4), 2000-2008.

[26] Hassanzadeh, Z., Kompany-Zareh, M., Ghavami, R., Gholami, S., & Malek-Khatabi, A. (2015). Combining radial basis function neural network with genetic algorithm to QSPR modeling of adsorption on multi-walled carbon nanotubes surface. Journal of Molecular Structure, 1098, 191-198.

[27] Moreno, J., Ovalle, D. A., & Vicari, R. M. (2012). A genetic algorithm approach for group formation in collaborative learning considering multiple student characteristics. Computers & Education, 58(1), 560-569

[28] Albreiki, B., Zaki, N., & Alashwal, H. (2021). A Systematic Literature Review of Student'Performance Prediction Using Machine Learning Techniques. Education Sciences, 11(9), 552.

[29] Abosamra, G., & Faloudah, A. Machine Learning Based Marks Prediction to Support Recommendation of Optimum Specialization and Study Track. *International Journal of Computer Applications*, *975*, 8887.

[30] Suykens, J. A., & Vandewalle, J. (1999, July). Multiclass least squares support vector machines. In IJCNN'99. International Joint Conference on Neural Networks. Proceedings (Cat. No. 99CH36339) (Vol. 2, pp. 900-903). IEEE.

[31] Suganda, F., Nugroho, H. W., & Asmuni, I. (2021, December). Prediction of Graduation of Students of the Lampung School of Technology Nusantara using the K-Nearest Neighbor and Naive Bayes Algorithm. In International Conference on Information Technology and Business (ICITB) (pp. 69-72).

[32] Wang, H. (2021). Analysis and Prediction of CET4 Scores Based on Data Mining Algorithm. Complexity, 2021.

[33] A. Pradhan, "Support vector machine-a survey," Int. J. Emerg. Technol. Adv. Eng., vol. 2, no. 8, pp. 82–85, 2012

[34] S. Suthaharan, "Support vector machine," in Machine learning models and algorithms for big data classification, Springer, 2016, pp. 207–235.

[35] K. S. Durgesh and B. Lekha, "Data classification using support vector machine," J. Theor. Appl. Inf. Technol., vol. 12, no. 1, pp. 1–7, 2010.

[36] M. Brijain, R. Patel, M. R. Kushik, and K. Rana, "A survey on decision tree algorithm for classification," 2014. [39] Y.-Y. Song and L. U. Ying, "Decision tree methods: applications for classification and prediction," Shanghai Arch. psychiatry, vol. 27, no. 2, p. 130, 2015.

[37] M. Pal and P. M. Mather, "An assessment of the effectiveness of decision tree methods for land cover classification," Remote Sens. Environ., vol. 86, no. 4, pp. 554–565, 2003.

[38] H. Sharma and S. Kumar, "A survey on decision tree algorithms of classification in data mining," Int. J. Sci. Res., vol. 5, no. 4, pp. 2094–2097, 2016.

[39] Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1249.

[40] Aburomman, A. A., & Reaz, M. B. I. (2017). A survey of intrusion detection systems based on ensemble and hybrid classifiers. Computers & Security, 65, 135-152.

[41] Gomes, H. M., Barddal, J. P., Enembreck, F., & Bifet, A. (2017). A survey on ensemble learning for data stream classification. ACM Computing Surveys (CSUR), 50(2), 1-36.

[42] N. Saravanan and A. Rathinam, "A comparitive study on ANN based fault location and classification technique for double circuit transmission line," in 2012 Fourth International Conference on Computational Intelligence and Communication Networks, 2012, pp. 824–830.

[43] D. A. Cirovic, "Feed-forward artificial neural networks: applications to spectroscopy," TrAC Trends Anal. Chem., vol. 16, no. 3, pp. 148–155, 1997.

[44] Zacharis, N. Z. (2016). Predicting student academic performance in blended learning using artificial neural networks. International Journal of Artificial Intelligence and Applications, 7(5), 17-29.

[45]Abu-Naser, S. S., Zaqout, I. S., Abu Ghosh, M., Atallah, R. R., & Alajrami, E. (2015). Predicting student performance using artificial neural network: In the faculty of engineering and information technology.

[46] J. Liu et al., "Prediction of rupture risk in anterior communicating artery aneurysms with a feed-forward artificial neural network," Eur. Radiol., vol. 28, no. 8, pp. 3268–3275, 2018.

[47] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," Pattern Recognit., vol. 91, pp. 216–231, 2019.

[48] D. Chicco, N. Tötsch, and G. Jurman, "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation," BioData Min., vol. 14, no. 1, pp. 1–22, 2021.

[49] Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for multi-class classification: an overview. arXiv preprint arXiv:2008.05756.

Appendix

In this part, official papers are drawn from the books of the Ministry of Education, in addition to the arbitrator questionnaire

Appendix A:

Below is an image of the official books to facilitate the researcher's tasks by the Ministry of Education and Higher Education

بسم الله الرحمن الرحيم

State of Palestine
Ministry of Education
Center for Educational Research and Development

دولة فلسطين
وزارة التربية والتعليم
مركز البحث والتطوير التربوي

الرقم: وت / ٥٦٣/١/١٣
التاريخ: 2021/9/6 م

حضرة الأخ لازر الشرافي المحترم
مستشار الوزير لشؤون التكنولوجيا

الموضوع: تسهيل مهمة بحثية

يهديكم مركز البحث والتطوير التربوي أطيب تحية، ويرجو منكم التكرم بتسهيل مهمة الباحث:
جهاد موسى امين الشيخ موسى

من **الجامعة العربية الامريكية** للحصول على المعلومات اللازمة لإعداد دراسة بعنوان:
" التنبؤ بالتخصصات الطلابية وتصنيفها باستخدام تقنيات التعلم الآلي"

وتضمن الدراسة الحصول على بيانات للطلبة من الصفوف تاسع وحتى توجيهي من موقع e-school وفق ما تقتضيه إجراءات الوزارة.

وتقبلوا فائق الاحترام والتقدير ،،،

د. محمد مطر
/مدير عام مركز البحث والتطوير التربوي

دولة فلسطين
وزارة التربية والتعليم

نسخة:
عطوفة وكيل الوزارة المحترم
عطوفة الوكلاء المساعدين المحترمين
أ.د. محمد عوض المشرف الرئيس على الدراسة المحترم -بريد الكتروني- mohammed.awad@aaup.edu

Tel (+ 970-562-501092)
E-mail (ncerd@moe.edu.ps)

بسم الله الرحمن الرحيم

State of Palestine
Ministry of Education
Center for Educational Research and Development

دولة فلسطين
وزارة التربية والتعليم
مركز البحث والتطوير التربوي

الرقم: و ت /١٣/١/ ٦١٥
التاريخ: 16 /9 /2021م

لمن يهمه الأمر

الموضوع: "تسهيل مهمّة بحثية"

يهديكم مركز البحث والتطوير التربوي أطيب تحية، ويرجو منكم التكرم بتسهيل مهمة الباحث:

"جهاد موسى امين الشيخ موسى"

من **الجامعة العربية الأمريكية** للحصول على المعلومات اللازمة لإعداد دراسة بعنوان:

"التنبؤ بالتخصصات الطلابية وتصنيفها باستخدام تقنيات التعلم الآلي"

ملاحظات:

- تتضمن الدراسة توزيع استبيان على عينة من طلبة الصفوف الحادي عشر والتوجيهي.
- ت/يتولى الباحث/ة أنشطة جمع البيانات، بتنسيق مع ملسق البحث والتطوير والجودة في المديرية.
- الاستجابة على الأدوات البحثية من قبل عينة المبحوثين طوعية.
- نظراً لظروف الجائحة يتم تطبيق أدوات البحث عبر النماذج المحوسبة دون تواصل وجاهي مع المبحوثين.

مع الاحترام،،

د. محمد مطر

/مدير عام مركز البحث والتطوير التربوي

نسخة:

عطوفة وكيل الوزارة المحترم
عطوفة الوكلاء المساعدين المحترمين
السادة مدراء عامين مديريات التربية والتعليم المحترمين
أ.د. محمد عوض المشرف الرئيس على الدراسة المحترم – بريد الكتروني– mohammed.awad@aaup.edu

Appendix B:

Here, a sample questionnaire was drawn up, which was distributed to the students

استبانة تحديد ميول الطالب

حضرة الطالب/ة المحترم/ة،

يقوم الباحث، وهو طالب في مرحلة الماجستير في الجامعة العربية الأمريكية بتخصص علم الحاسوب، بإجراء بحث تحت عنوان: **التنبؤ بالتخصصات الطلابية وتصنيفها باستخدام تقنيات التعلم الآلي.** وذلك استكمالاً لمتطلبات أطروحة الماجستير. والفئة المستهدفة من هذه الاستبانة هم طلبة الصفين الحادي عشر والثاني عشر الذين التحقوا فعليا بأحد الفروع الأكاديمية والمهنية التي توفرها وزارة التربية والتعليم، حيث تبحث هذه الدراسة التأثيرات التي دعت الطالب الى اختيار إحدى هذه الفروع ، لذا أعد الباحث هذه الاستبانة والمكونة من أربعة محاور، كل محور يحتوي على عدد من الفقرات الواجب الاجابة عليها بدقة وموضوعية، علماً بأن البيانات التي سيتم جمعها ستكون سرية، وستستخدم لأغراض البحث العلمي فقط.

وتفضلوا بقبول فائق الاحترام

الباحث: جهاد موسى أمين الشيخ موسى

يرجى تحديد ما ينطبق عليك فيما يلي:				
مديرية التربية والتعليم	1- جنين	2- قباطية	3 - طوباس	
الصف الدراسي	1- حادي عشر	2- ثاني عشر	**جنس الطالب**	1- ذكر	2- أنثى
تخصص الطالب	1- علمي	2- أدبي	3- تجاري	4- مهني	

يرجى الإجابة عن الفقرات التالية بوضع دائرة حول الإجابة المناسبة:			
كنت أميل لدراسة التخصص التالي:			
1-	علمي	2-	أدبي
3-	تجاري	4-	مهني
اعتبر نفسي في مجال العلاقات الاجتماعية:			
1-	اجتماعي جدا	2-	اجتماعي
3-	محدود العلاقات	4-	انعزالي
اعتبر نفسي متفرغاً للدراسة بشكل:			
1-	مشغول جدا	2-	مشغول
3-	متوسط الانشغال	4-	متفرغ بشكل كامل
عدد ساعات دراستي اليومية:			
1-	أقل أو يساوي ساعة	2-	أكثر من ساعة وأقل أو يساوي ثلاث ساعات
3-	أكثر من ثلاث ساعات وأقل أو يساوي ست ساعات	4-	أكثر من ست ساعات

عدد الساعات التي أقضيها على الإنترنت والمحمول يومياً:			
1-	أقل أو يساوي ساعة	2-	أكثر من ساعة وأقل أو يساوي ثلاث ساعات
3-	أكثر من ثلاث ساعات وأقل أو يساوي ست ساعات	4-	أكثر من ست ساعات
عدد الساعات التي أقضيها في اللعب والمجهود البدني يوميا			
1-	أقل أو يساوي ساعة	2-	أكثر من ساعة وأقل أو يساوي ثلاث ساعات
3-	أكثر من ثلاث ساعات وأقل أو يساوي ست ساعات	4-	أكثر من ست ساعات

أخي الطالب/ة أرجو تعبئة البنود اللاحقة بالتعاون مع ولي أمرك.							
عمل الأب	نعم			لا			
عمل الأم	نعم			لا			
تخصص الأب	علمي	أدبي	تجاري	مهني			
تخصص الأم	علمي	أدبي	تجاري	مهني			
ميول الأب	علمي	أدبي	تجاري	مهني			
ميول الأم	علمي	أدبي	تجاري	مهني			
الأب متفرغ لأبنائه الطلاب	نعم			لا			
الأم متفرغة لأبنائها الطلاب	نعم			لا			
هناك منافسة من أقاربي لدراسة نفس تخصصي	نعم			لا			

يرجى تعبئة المعلومات التالية بالتعاون مع مدير مدرستك:							
توفر التخصص / الفرع في البلد	**نعم**			**لا**			
مكان سكن الطالب	مدينة		مخيم		قرية		
توجه المجتمع	علمي	أدبي	تجاري	مهني			

يرجى كتابة علامتك أمام كل مادة دراسية فيما يلي:

المادة الدراسية	العلامة
التربية الدينية	
اللغة العربية	
اللغة الإنجليزية	
الرياضيات	
الفيزياء	

الكيمياء	
العلوم الحياتية	
تاريخ فلسطين وجغرافيتها	
الفنون والحرف	
التربية البدنية	
التربية التكنولوجية	

- ملاحظة: الرجاء الاستعانة بمنصة العلامات الرقمية التابعة لوزارة التربية والتعليم (E-School) للتأكد من دقة العلامات المدخلة.

الملخص

التنبؤ بتخصصات الطلاب هام جداً في تطوير أداء وزارة التربية والتعليم الفلسطينية، ووضع خطط الإنعاش الأولي للطلاب متدني الأداء، وتحديد الاحتياجات المستقبلية لنظام وزارة التربية والتعليم، والمساهمة في بناء هياكل لتنظيم المعلمين داخل وزارة التربية والتعليم، وغيرها من النتائج المترتبة على الترقب المبكر لتخصصات الطلاب المختلفة.

تعد هذه الدراسة الأولى من نوعها في فلسطين، وهي تختص بتحديد التخصصات المستقبلية للطلاب في الفروع العلمية والأدبية والصناعية والتجارية، بالاعتماد على تقنيات التعلم الآلي وقدرتها على التصنيف. ولمّا كان تحديد الميزات الأساسية وحجم مجموعة البيانات التي تم جمعها أمرًا بالغ الأهميّة في تعزيز دقة التصنيف، ومصفوفات التصنيف، لذلك اعتنى الباحث باختيار أفضل خوارزميات التعلم الآلي، وأكثرها دقة وكفاءة في التصنيف، وبذلك كانت المصفوفات التي تعتمد على الميزة المختارة والبيانات التي تم جمعها هي جوهر هذه الأطروحة.

الخطوة الأولى لتحقيق التصنيف بأكبر قدر ممكن من الدقة هي اختيار العوامل التي حددها المختصون في مجال التعليم بالمعلومات المطلوبة عن الطالب، وأسرته، ومجتمعه، بالإضافة إلى مستوى تحصيله الأكاديمي في الصف العاشر. تبع هذه العملية إنشاء استبيان محكّم، فتوزيعه على مديريات التربية والتعليم المختلفة بعد الحصول على إذن مسبق للقيام بذلك. وضمت مجموعة البيانات التي تم جمعها [عيّنة الدراسة] أكثر من 1200 طالبا وطالبة في مختلف الفروع. ثمّ أخضعت البيانات التي تمّ جمعها لمرحلة المعالجة المسبقة عن طريق تنظيفها وقياسها وترميزها لتكون جاهزة لتقنيات التعلم الآلي.

تمّ استخدام العديد من خوارزميات تصنيف المتعلم، بالإضافة إلى خوارزميات الشبكات العصبية. أمّا الخوارزميات الخطية المستخدمة فهي: خوارزمية k– الأقرب للجيران (KNN)، وآلات متجه الدعم (SVM)، والتمييز الخطي، وشجرة القرار، والشجرة المعززة المجمعة، والشجرة

المعبأة المجمعة. وأمّا الشبكة العصبية المستخدمة فهي شبكة عصبية متعددة الطبقات(MLPNNs).

أظهرت التجارب أن الشبكات العصبية تفوقت بشكل كبير على بقية تقنيات التعلم الآلي الأخرى؛ فقد وصلت دقة التصنيف المعتمدة عليها إلى 96.9٪، بينما كانت دقة التصنيف في معرّف الخوارزمية المجمعة 93.8٪، وهي الأفضل ضمن فئة خوارزميات متعلمي التصنيف. وأظهرت التجارب أيضًا ضعف الدقة عند استخدام العلامات في المنهج وحده للتصنيف على جميع استخدامات الخوارزمية.