Palestine Technical University Research Journal, 2020, 8(2), 23-39

Predicting Students Performance Based on their Academic Profile

التنبؤ بأداء الطلاب بناء على ملف الطالب الأكاديمي

Hadi Khalilia[1*], Thaer Sammar[2], Yazeed Sleet[2]

هادي خليلية[1*]، ثائر سمار[2]، يزيد سليط[2]

[1]Applied Computing Department, Palestine Technical University-Kadoorie, Tulkarm, Palestine, [2]Computer Engineering Department, Palestine Technical University-Kadoorie, Tulkarm, Palestine

[1]قسم الحوسبة التطبيقية، جامعة فلسطين التقنية خضوري، طولكرم، فلسطين، [2]قسم هندسة الحاسوب، جامعة فلسطين التقنية خضوري، طولكرم، فلسطين

Received: 23/06/2020 **Accepted: 22/11/2020** **Published: 01/12/2020**

Abstract: Data mining is an important field; it has been widely used in different domains. One of the fields that make use of data mining is Educational Data Mining. In this study, we apply machine learning models on data obtained from Palestine Technical University-Kadoorie (PTUK) in Tulkarm for students in the department of computer engineering and applied computing. Students in both fields study the same major courses; C++ and Java. Therefore, we focused on these courses to predict student's performance. The goal of our study is predicting students' performance measured by (GPA) in the major. There are many techniques that are used in the educational data mining field. We applied three models on the obtained data which have been commonly used in the educational data mining field; the decision tree with information gain measure, the decision tree with Gini index measure, and the naive Bayes model. We used these models in our work because they are efficient and they have a high speed in data classification, and prediction. The results suggest that the decision tree with information gain measure outperforms other models with 0.66 accuracy. We had a deeper look on key features that we train our models; precisely, their branch of study at school, field of study in the university, and whether or not the students have a scholarship. These features have an influence on the pre-diction. For example, the accuracy of the decision tree with information gain measure increases to 0.71 when applied on the subset of students who studied in the scientific branch at high school. This study is important for both the students and the higher management of PTUK. The university will be able to do some predictions on the performance of the students. In the carried experiments, the prediction of the model was in line with the actual expectation.

Keywords: Machine Learning, Data Mining, Decision Tree, Gini Index, Naive Bayes, Prediction.

المستخلص: يعد التنقيب عن البيانات مجالًا مهمًا حيث تم استخدامه على نطاق واسع في مجالات يومية مختلفة مثل التعليم، كما يعد التنقيب في البيانات التعليمية من أهم استخدامات التنقيب عن البيانات. في هذه الدرسة: قمنا بتطبيق بعض النماذج من نماذج التعلم الآلي على بيانات لطلاب قسم هندسة الحاسوب والحوسبة التطبيقية في جامعة فلسطين التقنية - خضوري في طولكرم. حيث يدرس الطلاب في كلا التخصصين نفس المواد التخصص الاجبارية مثل مساق لغة السي بلس بلس ومساق لغة الجافا. لذلك، اعتمدنا على مثل هذه المساقات للتنبؤ بأداء الطالب. الهدف من هذه الدراسة هو توقع أداء الطلاب في تخصصه حيث يتم قياسه من خلال المعدل التركمي. هناك

* Corresponding author: h.khalilia@ptuk.edu.ps

العديد من التقنيات المستخدمة في مجال التنقيب عن البيانات في المجالات التعليمية. تم تطبيق ثلاثة نماذج على بيانات الدراسة والتي كانت شائعة الاستخدام في مجال التنقيب عن البيانات وهي شجرة اتخاذ القرار مع مقياس اكتساب المعلومات ، وشجرة القرار بمقياس مؤشر "جيني" ، ونموذج بيز. تم استخدامنا هذه النماذج لأنها فعالة ولديها سرعة عالية في تصنيف البيانات والتنبؤ بها. تشير النتائج إلى أن شجرة اتخاذ القرار مع مقياس اكتاب المعلومات تتفوق على النماذج الأخرى بدقة 0.66. تم دراسة وتحديد حقول البيانات الرئيسة المتضمنة في بيانات الدراسة والتي تم استخدامها في تدريب وتطبيق النموذج المختار، على وجه التحديد تم اختيار حقل تخصص الطالب في الدراسية الثانوية (علمي , ادبي ..)، تخصص الطالب في الجامعة، وما إذا كان الطالب حاصل على منحة دراسية أم لا. حيث ان هذه الحقول لها تأثير على التنبؤ باداء الطالب. على سبيل المثال، تزيد دقة شجرة القرار مع مقياس اكتساب المعلومات إلى 0.71 عند تطبيقها على مجموعة فرعية من الطلاب الذين درسوا في الفرع العلمي في المرحلة الثانوية. هذه الدراسة مهمة لكل من طلبة وادارة جامعة فلسطين التقنية – خضوري، حيث ستكون الجامعة قادرة على استخدام النموذج للتنبؤ باداء الطلبة. هذا وقد كانت نتائج التجارب بالتنبؤ باداء الطلبة تتماشى مع اداء الطلبة الحقيقية الموجودة في عينة البيانات التي تم تطبيق الدراسة عليها.

الكلمات المفتاحية: تعلم الالة، التنقيب عن البيانات، شجرة القرار، مؤشر جيني، تصنيف بيز، تنبؤ.

INTRODUCTION:

Data mining techniques are used to extract knowledge from raw data. It has been used in different domains; one of the domains is the education domain. Data mining educational data is related to extracting knowledge from educational data. This knowledge will help in improving education and improving student's performance (Baker, 2010; Algarni, 2016).

In this research, we apply Data mining techniques on the university data of undergraduate students from two fields of study; Computer Engineering and Applied Computing. Our data set was obtained from the deanship of admission and registration in our university. In total, we have information about 422 students. These students' study either Computer Engineering or Applied Computing. We focused on these two fields because the main goal was to predict students' performance in the major. Students from the two fields study C++ and Java under the same conditions. These two courses form the main major courses in the first semester and the second semester .

In order to predict students' performance, we use the naive Bayes and decision tree models. The inputs of each model are the selected attributes from the data warehouse such as: student's Tawjihi branch, his/her mark in Tawjihi, his/her GPA in first year, if he/she has a scholarship in first semester, if he/she has a scholarship in second semester, the number of hours that he/she registers in first semester, his/her GPA in first semester, the number of hours that he/she registers in second semester and his/her GPA in second semester. By applying the models on the relevant features, we predict the student's major GPA at the end of his/her first year as the output of the model which indicates the student's performance in the major.

In our research, we try to answer the following research questions:

RQ1 Can we predict student performance from past results?

We apply machine learning models on students' data sets taken from the deanship of admission and registration in our university.

RQ2 Does the fact that the student has a scholarship affects her/his performance?

Some students in the university get scholarship when they register based on their GPA in the high school. The criteria to keep the scholarship for the next semester are to have a high GPA in the university. Therefore, we would to like to explore the impact of the scholarship on the performance in this research question.

RQ3 What is the impact of the field of study?

Our data set consists of students studying Computer Engineering or Applied Computing. After school students with high GPA (usually above 89/100) go to the Computer Engineering. We explore the behavior of the prediction models on these two fields.

RQ4 What is the impact of the high school background on the performance in the university?

Finally, in this research question we try to investigate the impact of the high school background on students' performance. Students who got accepted in the Computer Engineering or Applied Computing came from two different branches at high school; either Scientific or Vocational.

The rest of the paper is organized as follows. After discussing related work (Section 2) we describe our approach to answer the research questions introduced in this section (Section 3). We discuss the experimental setup in detail in (Section 4) and answer research questions RQ1-4 in Sections 5, 5.1, 5.2, 5.3 respectively. Finally, we discuss conclusions drawn from our findings (Section 6) .

RELATED WORK:

Data mining techniques are used to extract useful information from raw data. Extracted information can help in making decisions. Data Mining techniques have been used to extract knowledge from educational data, the field known as Educational Data Mining (Baker, 2010; Algarni, 2016; Romero and Ventura, 2007).

Educational Data Mining can be used in the education field to improve our understanding of learning process to focus on identifying, extracting, testing and evaluating attributes related to the learning process of students as shown in El-Halees's (2009).

Data mining models have been applied to study the accomplishment of students, focusing on two steps of their contribution (Asif et al., 2017). Two important groups of students have been taken: high performance and low performance. The results point out that by focusing on a lesser number of topics that are indicators of performance level, it is possible to arrange for suitable comments and give chances to high achieving students, and guidance and support to low achieving students.

Classification techniques such as Decision Trees and Bayesian Network were applied on engineering's student's data with the goal to predict their performance in the final exam (Yadav and Pal, 2012). The socio-demographic variables such as age, gender, and work status and study environment were used to predict the success and unsuccess of students (Kovacic, 2010). The decision tree method was used to extract knowledge that describes students' performance in end semester examination (Baradwaj and

Pal, 2012). The main goal of this work was to help in identifying the dropouts and students who need special attention.

Pandey and Pal conducted two researches on the interestingness of student in class teaching language and his/her performance based on the means of Bayes Classifier and association rules. They predict whether new comer students will perform well or not using his language, category and background qualification (Pandey and Pal, 2011a, b).

Baradwaj and Pal (2012) extract knowledge that describes students' performance and classify the students at VBS Purvanchal university based on information's like attendance, class test, seminar and assignment marks using the decision tree classification approach with information gain and splitInfo measures (Baradwaj and Pal, 2012).

Amazona and Hernandez (2019) applied three data mining classification models (Naïve Bayes, Decision Tree and Deep Learning in Neural Network) using RapidMiner framework to predict students' performance at the Aurora State College of Technology. Results show that Deep Learning classifier exceeds other two models by gaining the overall forecast accuracy of (0,95).

Hijazi, et al. (2006) worked on the student's performance study by selecting a sample of 300 students (75 females, 225 females) from Punjab University in Pakistan. The assumption that was as "Students' mother's education and mother's age are related with student performance, students' family income, attendance in class and daily hours spent in study at the university" was taken. Using the linear regression model and its analysis, it was found that the factors as student's family income and mother's education were highly associated with the student academic performance.

Astha Soni, et al. (2018) predicted student's performance from different universities during the period (2017 - 2018) using the decision tree, the naïve Bayes and the support vector machine data mining techniques. The dataset is represented in academic, behavior, extra-curricular and placement categories. The authors got that academic information, personal information and family details have strong impact on the students' performance due to instinctive reasons using SVM (Soni et al., 2018).

Wahbeh et al. (2011) applied the decision tree model to predict the final result of students who studied the C++ course at Yarmouk University. Three classification methods namely C4.5, ID3, and the naïve Bayes were used. As an output, they were able to predict the grade of students in C++ course and even can improve his performance by taking training lessons in C++, and the results indicated that Decision Tree model is the best in terms of accuracy (Wahbeh et al., 2011).

In our university, we have computer engineering and applied computing. Students from both fields study some similar courses that represent major requirement. In the first year, for example, students from the two fields study C++ in the first semester and Java in the second semester. In our study, we apply machine learning to predict student's performance in major courses by applying the models on students who came from different school background and studying in two different fields but taking same major courser in the first year.

Data mining models have been applied in order to predict the performance of students at Palestine Technical University. Focusing on two categories of relevant features the first is: student's academic information like his/her high school branch and grade, his/her major, his/her number of registered and passed hours in first and second semesters, student's GPA in first and second semesters, and his/her major GPA. The second category is: student's scholarship and financial information using the naive bayes and decision tree models. The data warehouse consists of two datasets which are: computer engineering dataset and applied computing dataset. We got that three potential features have strong impact on the students' performance such as: the field of study, high school branch and a scholarship.

OUR APPROACH:

In this section, we describe our approach to answer the main research question of our work. In order to predict students' performance with the help of machine learning, we decided to use Naive Bayes and Decision Tree Classifiers. Regarding to the decision tree classifier, we used two measure variants; Information Gain and Gini Index. In the following, we describe these models .

Decision Tree Classifier:

Decision trees are famous and powerful methods for classification and prediction commonly used in machine learning and data mining. Decision tree induction is a learning method used to create a model from class-labeled training tuples that predicts the value of a class label based on a number of input variables in testing data (Baker, 2010; Yadav et al., 2012). A decision tree (DT) is a flowchart-like tree structure, where each non-leaf node (internal node) is denoted by rectangles (bounded or not) and it represents a test on an attribute. Each branch acts an outcome of the test, and each terminal node (or leaf node) is denoted by oval and holds a class label. DT starts with a root (topmost) node on which it is for users to take actions. From the root, users split source set into subsets according to DT learning algorithm recursively. The recursion is completed all subsets of a node have the same value of the target variable (Han et al., 2011). In data mining, DTs represent the combination of mathematical rules, which help in the categorization and description of a given set of data and can be understood by humans and used in knowledge base (Yadav and Pal, 2012). In general, decision tree model has good accuracy. DT learning algorithms have been used for classification in many areas, such as financial analysis, biology, medicine, marketing and astronomy. The construction of DT classifier does not require any parameter setting or domain knowledge for knowledge extraction. The key issues to do knowledge extraction using decision trees are: attribute-value explanation, determine target attribute values (predefined classes), discrete classes' specification and training cases of data must be sufficient usually hundreds or thousands of training cases (Han et al., 2011). In tree construction, attribute selection is based on two measures: information gain measure and Gini index.

Information Gain:

Information gain is an attribute selection measure that is used by ID3 algorithm and it is based on C. Shannon work, which studied the information content. Let node N represents the tuple of partition D (Yadav et al., 2012). The attribute with the highest information gain is chosen as the splitting attribute for the node N. The information gain is calculated by the following formula:

$$info(D) = -\sum_{i=1}^{m} pi \ \log_2(pi) \qquad (1)$$

Info(D) (the entropy of D): is the average amount of information to recognize the class label of a tuple in D. Pi is the probability of a tuple in D belongs to class Ci and is estimated by |Ci, D|/|D|. The information required to classify a tuple from D, based on the partitioning by attribute A is given by:

$$info_A(D) = -\sum_{j}^{v} \frac{|Dj|}{|D|} \times info(Dj) \qquad (2)$$

In brief, the difference between the original data information and the new data information (after partitioning on A) is called information gain as in equation (3):

$$\text{Gain(D)} = \text{info(D)} - \text{info}_\text{A}\text{(D)} \qquad (3)$$

Gini Index:

The Gini index is used in CART algorithm in case of the binary partitioning for each attribute. "The Gini index measures the impurity of D, a data partition or set of training tuples, as:"

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2 \qquad (4)$$

Where Pi: is the probability that a tuple in D belongs to class Ci and is calculated by |Ci,D|/|D.|

A weighted sum of the impurity of each partition (Gini index of D) that resulted by a binary split on attribute A is given in equation (5):

$$Gini_2(D) = \frac{|D1|}{|D|} \ Gini(D1) + \frac{|D2|}{|D|} \ Gini(D2) \qquad (5)$$

The binary split of each attribute is considered. For a discrete-valued attribute, "the minimum Gini-index for that attribute is selected as its splitting subset" (Baker, 2010).

If an attribute is continuous-valued then a split-point for attribute (e.g.: A) will be considered, the impurity reduction of a binary split on a discrete- or continuous-valued attribute A is given in equation (6):

$$\Delta Gini(A) = Gini(D) - Gini_A(D1) \qquad (6)$$

Naive Bayes Classifier:

Naïve Bayes (NB) classifier is a statistical classifier. It can predict class membership probabilities, such as the probability that a given record belongs to a specific class. NB classifier has a class conditional independence; that assumes each attribute value is independent of the values of the other attributes. The NB method has various advantages: only one scan of the training data is required; it is easy to use; easily handle mining value by simply omitting that probability. NB classifier requires a small amount of training data to calculate the parameters that are important for classification as (mean and variance) (Pandey and Pal, 2011a). The Naïve Bayes (NB) classifier works as: Let D be a training data set and

associated with a number of class labels. And each tuple in the data set is denoted by X (an n-dimensional attribute vector), the NB classifier predicts that each tuple X in the data set belongs to a class C in the associated M classes (Pandey and Pal, 2011a). NB classifier works based on Bayes' theorem (see equation (7)):

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \qquad (7)$$

As P(X) is constant for all classes, only P (X | Ci) P(Ci) need be computed. And by referencing to NB assumption of class conditional independence; P(X | Ci) can be shown as in eq (8):

$$P(X|C_i) = \prod_{k=1}^{n} P(X_k|C_i) \qquad (8)$$

$$= P(X_1|C_i) \times P(X_2|C_i) \times \dots\dots\dots \times P(X_k|C_i)$$

When the probabilities P(X1 | Ci) × P(X2 | Ci) × … × P(Xn | Ci) for training tuples are estimated, we can easily predict the class label of X, by evaluating the P(X | Ci)P(Ci) and compare it with P(X | Cj)P(Cj). NB predicts that the class label of tuple X is the class Ci if and only if: P (X| Ci) P(Ci) > P (X | Cj) P(Cj) for $1 \le j \le m, j \neq i$.

If the comparison result is true then NB predicts that the class label of tuple X is the class Ci else it predicts the class label is Cj. Feature selection is an important phase to select the relevant variables on the basis of the values of probabilities; feature selection is the ancestor phase of NB classification technique usage as a data mining tool to design the student's major GPA prediction model.

EXPERIMENTAL SETUP:

In this section, we describe the dataset, present some statistical information derived from the dataset and discuss features that we used to train our models.

Data Set:

The data set used in this study was obtained from the deanship of admission and registration at the Palestine Technical University - Kadoorie. The data set consists of 422 students' records. This includes all students who are enrolled in computer engineering and applied computing fields since 2017 to 2019. The number of students in the dataset contains all students from the two fields in the target year of study, this dataset is enough for our analysis. The dataset has a variety of student; two fields of study, different school background, different level of performance in university study and school .

For each student, the following information was provided to us by the deanship of admission and registration at the Palestine Technical University – Kadoorie based on our request :

School Information:

Tawjihi Branch which represent the field of study at school before the university and the GPA in the national school exam (Tawjihi).

Our goal of the study is to predict students' performance in the major courses of the first year; which are C++ in the first semester and Java in the second semester. Therefore, we got student information in the first semester and second semester

First Semester Information:

Which includes the total number of credit hours taken by the student in his/her first semester of study, total number of credit hours that the students pass, the GPA of the first semester, and whether or not the student has a scholarship

Second Semester Information:

Total number of credit hours in the second semester, total number of passed credit hours, the GPA in the second semester, and whether or not the student has a scholarship

For each student, we got the grade in all courses studied in the first semester and the second semester. However, we excluded students' grades in the courses other than the computer programming (C++) course and the object-oriented programming (JAVA) course.

The Major GPA:

Which we computed based on C++ grade and Java grade

Gender of Student:

We did not include this information during the training of the model

First Year GPA

The GPA for each student was available to use, based on this we derived the first year GPA evaluation by classification each GPA into one of the following categories based on university system (Excellent, Very Good, Good, Satisfactory, Fail)

STATISTICAL INFORMATION ON THE DATASET:

Here, we give some statistical information but our dataset. As mentioned above, our data set contains 422 records, 86% of students come from the scientific high schools, and the rest come from the vocational high schools. 66% of student's studies at the computer engineering department and the rest studies at the applied computing department. 79% of the students had a scholarship at the first academic semester of their studies, while 82% of the students had a scholarship at the second academic semester of their studies.

In **Table 2**, we present some statistical information about the dataset. 364 students out of 422 studying computer engineering and applied computing came from the scientific tawjihi branch. The rest (58 students) finished high school in vocational branch. The average of tawjihi marks of students from scientific branch is higher than the average for students from vocational branch, which is 87.6 and 84.7 respectively. If we look at fields of study. The average of tawjihi marks of students in computer engineering who studies in scientific high school is higher compared to student in applied computing

who studied in vocational high school. The average of first year GPA f show same pattern like the average of tawjihi marks .

TRAINING FEATURES:

As mentioned above, our data set contains 422 records, but we have some details about the data set. 86% of the students come from the scientific high schools, and the rest come from the vocational high schools. 66% of student's studies at the computer engineering department and the rest studies at the applied computing department. 79% of the students had a scholarship at the first academic semester of their studies, while 82% of the students had a scholarship at the second academic semester of their studies.

Table (1) describes the selected fields in terms of the field name, the field general description, the field value domain and the domain description. In the following we describe features that were included in training the models for the GPA prediction:

- Tawjihi Branch: Students in our dataset are from the either the computer engineering field or the applied computing field, students from both fields came from the scientific or vocational major at high school.
- Tawjihi Mark: in training our models, we include student's GPA at high school, marks are in the 0 – 100 scale.
- Major: this feature represents the major of students at university; this feature has two possible values; either computer engineering or applied computing .
- Scholarship Sem 1: this feature in a Boolean feature. True indicates that the student got a scholarship at admission time to the university based on his high school (Tawjihi) GPA .
- Scholarship Sem 2: this feature is a Boolean feature; has two possible values {True, False} indicates whether the student got a scholarship or not for the second semester based on his performance in the first semester.
- Num of hours in Sem 1: this feature represents the number of credit hours taken by the student in the first semester. The possible value of this feature ranges between 12 and 18.
- Num of pass hours Sem 1: this feature shows how many credit hours the student finished successfully in the first semester. The value of this feature in a number that ranges between 12 and 18.
- First Sem GPA: this feature represents the GPA of the student in the first semester. The value of this feature ranges between 40 and 100 .
- Features in row 10, 11 and 12 in Table 1 are the same as the previous three features but for the second semester.

Table (1): Data Set Fields Description

No.	Field Name	Field Description	Value Domain	Domain Description
	Tawjihi Branch	The students major in the high school.	{scientific | vocational}	There are many majors in high school. But since our target

			group is the students from the faculty of Applied Sciences and the faculty of Engineering and Technology, we have only two majors due the admission and registration rules.
Tawjihi Grade	The student achievement in the high school.	Any decimal value that is greater than or equal to 0, and it is less than or equal to 100.	In general, the minimum grade is 0. The maximum grade is 100, but the allowed-minimum grade in Computer Engineering major is 89 and 70 in Applied Computing.
Gender	Students' gender.	{Male | Female}	
Major	The student's major in the university.	{AppliedComp compEng }	AppliedComp means that the student study at the applied computing department at the faculty of Applied Sciences. CompEng means that the student study at the computer engineering department at the faculty of Engineering and Technology.
Scholars-hip Sem 1	Determines whether the student has a scholarship at the end of the first semester of his first study year or not.	{TRUE| FALSE}	Based on the PTUK university rules and criteria in scholarship system including the student's academic achievement in Tawjihi examination, the student may deserve a scholarship. *True* means that the student has a scholarship and *False* which mean that the student does not have a scholarship.
Scholars-hip Sem2	Determines whether the student has a scholarship at the end of the second semester of his first study year or not.	{TRUE| FALSE}	Based on the PTUK university rules and criteria in scholarship system including the student's academic achievement in the first semester, the student may deserve a scholarship. *True* means that the student has a scholarship and *False* which means that the student does not have a scholarship.
Num of Hours Sem1	Determines the number of the credit hours that the student has registered at the first semester of the first study year in his/her program.	Any integer value that is greater than or equal to 12 and less than or equal to 18.	The minimum number of credit hours that the student can register at the first semester of the first year of study is 12 credit hours and the maximum is 18 credit hours.

Num of Pass Hours Sem1	Determines the number of the credit hours that the student has passed successfully at the end of the first semester of the first study year.	Any integer value that is greater than or equal to 0 and less than or equal to 18.	The minimum number of credit hours that the student has passed successfully at the end of the first semester of his first study year is 0 credit hours and the maximum is 18 credit hours.
First Sem GPA	The grade point average that the student has achieved in the first semester of his first study year.	Any decimal value that is greater than or equal to 40 and less than or equal to 100.	The minimum grade is 40. The maximum grade is 100.
Num of Hours Sem2	Determines the number of the credit hours that the student has registered at the second semester of the first study year.	Any integer value that is greater than or equal to 12 and less than or equal to 18.	The minimum number of credit hours that the student can register at the second semester of his first study year is 12 credit hours and the maximum is 18 credit hours.
Num of Pass Hours Sem2	Determines the number of the credit hours that the student has passed successfully at the end of the second semester of the first study year.	Any integer value that is greater than or equal to 12 and less than or equal to 18.	The minimum number of credit hours that the student has passed successfully at the end of the second semester of his first study year is 0 credit hours and the maximum is 18 credit hours.
Sec Sem GPA	The grade point average that the student has achieved in the second semester of his first study year.	Any decimal value that is greater than or equal to 40 and less than or equal to 100.	The minimum grade is 40. The maximum grade is 100.
First Year GPA	The grade point average that the student has achieved at the end of his first study year.	Any decimal value that is greater than or equal to 40 and less than or equal to 100.	The minimum grade is 40. The maximum grade is 100.
First Year GPA Eval	The student's level in the grading system at the end of the first year of study year.	{Excellent| Very Good | Good | Satisfactory| Fail}	The grade level depends on the first year GPA as (Excellent: 100>=GPA>=90. Very Good: 90>GPA>=80. Good: 80>GPA>=70. Satisfactory: 70>GPA>=60. Fail: 60>=GPA>=0).
Major GPA Eval	The student's level in the grading system at the end of his first study year for only two courses the JAVA and C++. Major GPA = ("JAVA" GPA + "C++" GPA)/2.	{Excellent| Very Good | Good | Satisfactory| Fail}	The level depends on the GPA of the major courses as (Excellent: 100>=GPA>=90. Very Good: 90>GPA>=80. Good: 80>GPA>=70. Satisfactory: 70>GPA>=60. Fail: 60>=GPA>=0).

Table (2) below shows some statistical information derived from the dataset:

Table (2): Statistical Information

	Count of Students	Average of Tawjihi Mark	Std. Dev. of Tawjihi Mark	Average of First Year GPA	Std. Dev. of First Year GPA	Count of Scholarships / First Semester	Count of Scholarships /Second Semester
Scientific	**364**	**87.61**	**8.50**	**74.89**	**8.02**	**309**	**318**
Female	**242**	**88.93**	**7.65**	**75.91**	**7.96**	**212**	**217**
Applied Comp.	64	77.60	5.08	69.78	6.02	45	46
Comp. Eng.	178	93.00	2.72	78.12	7.41	167	171
Male	**122**	**84.98**	**9.48**	**72.85**	**7.78**	**97**	**101**
Applied Comp.	50	74.54	3.82	67.52	5.08	35	35
Comp. Eng.	72	92.23	3.60	76.56	7.16	62	66
Vocational	**58**	**84.68**	**8.29**	**64.86**	**7.98**	**28**	**29**
Female	**13**	**90.47**	**4.97**	**67.82**	**7.61**	**10**	**9**
Applied Comp.	2	80.60	2.55	53.85	7.71	1	0
Comp. Eng.	11	92.26	2.44	70.35	4.18	9	9
Male	**45**	**83.01**	**8.34**	**64.00**	**7.96**	**18**	**20**
Applied Comp.	25	76.33	4.20	59.04	5.38	5	4
Comp. Eng.	20	91.35	2.64	70.21	6.09	13	16
Grand Total	**422**	**87.20**	**8.52**	**73.51**	**8.72**	**337**	**347**

RESULTS AND DISCUSSION:

In this section, we study the use of Machine Learning to predict student's performance (RQ1). We explore the accuracy of the prediction models discussed in Section 3. Precisely, we applied three prediction models; Decision Tree using the information gain measure, Decision Tree using Gini index measure, and the naive Bayes Classifier. The main goal of the study is to explore the ability of

Machine learning prediction models to predict students' performance by predicting their GPA in the major. Our data set includes students from two study fields; the computer engineering and the applied computing. In Both fields, students are required to take two major courses; C++ in the first semester and Java in the second semester .

In Section 4, we described the fields that we train our models on. Based on the given information about the student, the task of the model is to predict student's major GPA. The prediction indicates one of the following classes; Fail, Satisfactory, Good, Very Good, or Excellent. These classes are based on the mark classification in our university; **Table (3)** summarizes the classes and the corresponding

GPA range. For example, fail means GPA is less than 60.

Table (3): GPA classes

Prediction GPA	Class GPA Range

Fail	GPA < 60
Satisfactory	60 ≤ GPA < 70
Good	70 ≤ GPA < 80
Very Good	80 ≤ GPA < 90
Excellent	90 ≤ GPA ≤ 100

The three models were trained and tested on 422 students from Computer Engineering and Applied Computing fields. The results show that the decision tree with information gain measure outperforms other models with accuracy = 0.66, see the accuracy of the prediction models in **Table (4**).

Table (4): Prediction Models Accuracy

Prediction Model	Accuracy
Decision Tree (Gini Index)	0.61
Decision Tree (Information Gain)	0.66
Naive Bayes	0.50

Prediction on Subsets:

On the following, we apply the same models on subsets derived from the original dataset. We used all features as discussed before; the only difference is that we took out one feature from the training which the one is representing the target of sub setting .

First Subset Decision:

We look at the scholarship, some students get scholarships when they enter the university, based on this, we got two subsets; one that contains all students from the original dataset who have scholarship, the second subset contains all students from the original dataset who do not have scholarships. Thus, in the prediction on these two subsets, we use all features except the scholarship .

Second Subset Decision:

The focus is on the field of study; whether students are studying Computer Engineering or Applied Computing. The criterion to accept students in either of the two fields is the GPA in the high school. As a result, we got two subsets one for the computer engineering students and the other for the applied computing students. We apply the model on the two, all features are used except the major field because it is the same for all in the same subset .

Third Subset Decision:

Here, we explore the high school branch (Tawjihi). Students at school after the tenth grade, choose a discipline to specialize in such the Scientific and the Vocational. Students who got accepted in Computer Engineering or Applied Computing finish high school from either the scientific branch or the vocational branch. From the original dataset, we generate two subsets; the first subset contains all students who came from the scientific branch at high school, the second subset contains all students who studied at vocational Brach at high school regardless of their field of study at university. For these two subsets we apply the same models using all features except the Tawjihi branch because all students in the same subset have the same value; either scientific or vocational.

In the following subsections, we present the results of the prediction after being applied on the subsets.

Scholarship:

At the beginning of this section, we showed how to use three Machine Learning models to predict students' performance. In this section, we study the impact of the scholarship on students' performance (RQ2). Our data set consists of 422 students from Computer Engineering and Applied Computing. Some students might have scholarship based on their GPA in the High school when they entered the university. The student who has a scholarship must have a high GPA in the university to keep the scholarship. Therefore, we think that students who have scholarships will try to perform well in the university in order to not lose their scholarship. Previously, the result suggests that the decision tree model with information gain measure outperforms NB and the decision tree with Gini index measure. Therefore, we apply the decision tree model with information gain measure on the subset of students who have scholarship. The accuracy of the model increases to 0.68 on this subset of the data. We look deeper on the results. Specifically, we examined the students' records where the model did not predict the correct GPA class. We found that in all cases in which there are a mismatch between the predicted GPA class and the actual GPA class; the model predicts higher than the actual value. More precisely, the miss predicted class is one class higher than the actual class. For example, when the actual class is Fail, the predicted class was Satisfactory. Most of the mismatch cases occur in the low GPA classes; Fail and Satisfactory. The prediction of the model is inline of what is expected.

STUDY FIELD:

In this section we investigate the impact of the study field on the student's performance (RQ3). As mentioned earlier, our data set consists of students from Computer Engineering and Applied Computing. The main selection criterion at the admission time is the GPA in the High School. Students admitted in the Computer Engineering has a GPA in the High school (called Tawjihi) higher than 89, while students admitted in the Applied Computing have a GPA higher than 70 in the High School Exam (Tawjihi). We split the students' data set into two subsets. Namely, the first subset consists of students studying Computer Engineering. The second subset consists of students studying Applied Computing. After applying the decision tree model with information gain measure on the two subsets, we found that the accuracy of the model was higher when applied on students in the Computer Engineering; with accuracy equal to 0.69.

High School Major:

In this section, we study the impact of high school background on the students' performance in Computer Engineering and Applied Computing fields (RQ4). Another feature that we look into deeply is the students major in the high school. In our university, the admission in the computer engineering and the applied computing fields requires that students finish their high school study from either the scientific branch or the vocational branch. Students from the scientific branch have stronger background specially in mathematics and physics compared to students from the vocational branch. We created two subsets of our students' data set; the first subset consists of students who came from the scientific high school branch. The second subset contains students who came from the vocational branch. We applied the decision tree with gain information on the two subsets, separately. The

performance of the model was higher on the subset consisting of students from the scientific branch; the accuracy of the model was 0.70, while the accuracy of model was 0.66 when applied on the subset consisting of students from the vocational branch. As we mentioned earlier in this section, students from the scientific branch have stronger background in mathematics and physics compared to students who came from the vocational branch. Figure 1 shows the decision tree for the vocational subset.

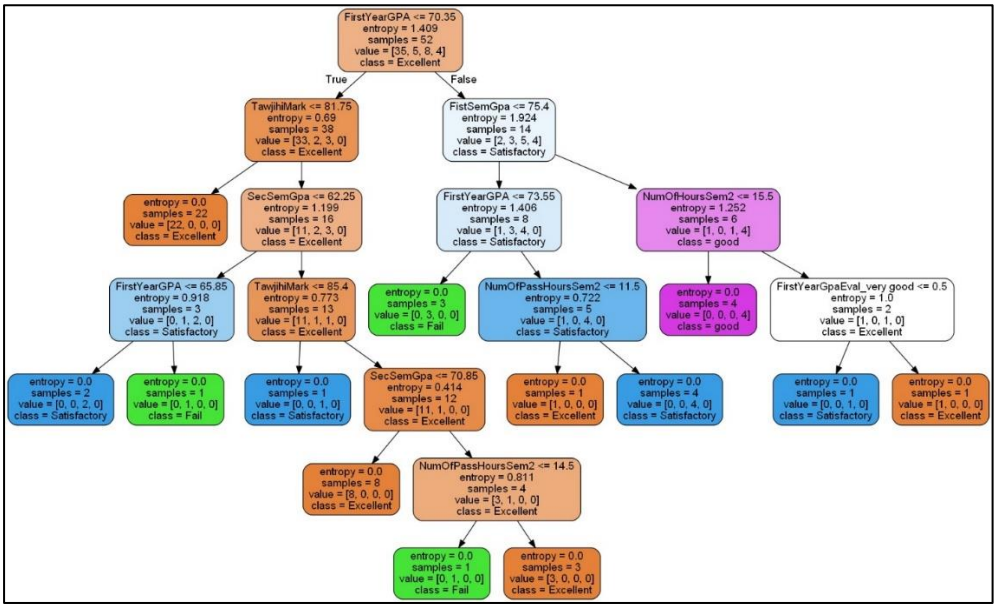

Figure (1): Decision Tree of the Vocational subset

CONCLUSIONS:

In our research, we explore the use of Machine Learning for predicting student's performance. Therefore, we applied three Machine Learning and Data Mining models to analyze the educational data for students who study at Computer Engineering and Applied Computing departments at Palestine Technical University (PTUK). The main goal of our study is to help the decision makers to improve student's performance and get better alumnus quality .

In Section 4, we discussed how features used to train our models were filtered from student educational data and how we constructed the task-relevant data set. We applied three models on the constructed data set; the decision tree with information gain measure, the decision tree with Gini index measure and NB.

The potential features that we trained our models on, include: high school (Tawjihi) branch and mark, gender, male, scholarship, major, number of registered and passed hours in first and second semesters, student GPA in first and second semesters, and major GPA. The results show that the decision tree model with information gain measure outperformed the Naive Bayes Classifier and the decision tree with Gini index measure.

Moreover, we explore three features that we think play an important role when predicting student's major GPA. We shed the light on three features: scholarship, high school branch and the field of study. It can be easily derived from the discussion and results we conclude that the three features have very good impact on the students' academic throughput. We found that the accuracy of the decision tree

with information gain model was higher when applied on students' subset who have a scholarship, Computer Engineering students, or students whose high school branch is scientific. The decision makers and higher administrations at the university can use our study results to improve major courses plans such as C++/Java and enhance their policies and strategies based on the extracted knowledge. For example, the accuracy of the decision tree with information gain measure was 0.7 when applied on the students who came from the scientific branch at high school. Students enrolled in Computer Engineering or Applied Computing, students studied either in the scientific branch or the vocational branch at high school for two years. Students from the scientific branch have better background in Mathematics and Physics. So even if they have same level of GPA at school, the scientific branch students have stronger background.

Future Work:

Data Mining and Machine Learning models are the most attractive techniques to extract knowledge from haystack of data, but various other classification algorithms can also be applied to test the most appropriate model that suit the texture of the student data and give better prediction accuracy.

ACKNOWLEDGEMENTS:

The Authors Would Like to Thank Palestine Technical University Kadoorie for Providing us with the Historical Data

REFERENCES:

Algarni, A. (2016). Data mining in education. International Journal of Advanced Computer Science and Applications, 7(6), 456–461.

Amazona, M. V., and Hernandez, A. A. (2019). Modelling Student Performance Using Data Mining Techniques: Inputs for Academic Program Development. Proceedings of the 2019 5th International Conference on Computing and Data Engineering, 36–40.

Asif, R., Merceron, A., Ali, S. A., and Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. Computers and Education, 113, 177–194.

Baker, R. (2010). Data mining for education. International Encyclopedia of Education, 7(3), 112–118. Baradwaj, B. K., and Pal, S. (2012). Mining educational data to analyze students' performance. ArXiv Preprint ArXiv:1201.3417.

El-Halees, A. M. (2009). Mining student's data to analyze e-Learning behavior: A Case Study. Mining Students Data to Analyze E-Learning Behavior: A Case Study, 29.

Han, J., Pei, J., and Kamber, M. (2011). Data mining: Concepts and techniques. Elsevier.

Hijazi, S. T., and Naqvi, S. M. M. (2006). Factors Affecting Students' Performance. Bangladesh E-Journal of Sociology, 3(1).

Kovacic, Z. (2010). Early prediction of student success: Mining students' enrolment data.

Pandey, U. K., and Pal, S. (2011a). A Data mining view on class room teaching language. ArXiv Preprint ArXiv:1104.4164.

Pandey, U. K., and Pal, S. (2011b). Data Mining: A prediction of performer or underperformer using classification. ArXiv Preprint ArXiv:1104.4163.

Romero, C., and Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. Expert Systems with Applications, 33(1), 135–146.

Soni, A., Kumar, V., Kaur, R., and Hemavath, D. (2018). Predicting student performance using data mining techniques. International Journal of Pure and Applied Mathematics, 119(12), 221–227.

Wahbeh, A. H., Al-Radaideh, Q. A., Al-Kabi, M. N., and Al-Shawakfa, E. M. (2011). A comparison study between data mining tools over some classification methods. International Journal of Advanced Computer Science and Applications, 8(2), 18–26.

Yadav, S. K., Bharadwaj, B., and Pal, S. (2012). Data mining applications: A comparative study for predicting student's performance. ArXiv Preprint ArXiv:1202.4815.

Yadav, S. K., and Pal, S. (2012). Data mining: A prediction for performance improvement of engineering students using classification. ArXiv Preprint ArXiv:1203.3832.