Education Data Mining to Predict Student Exam Grades in Vocational Institutes

استخراج البيانات التعليمية للتنبؤ بدرجات امتحان الطلاب في المعاهد المهنية

By

Student Name Ashjan Ahmad Saleh Saleh

Student ID number 2013210022

Dissertation submitted in partial fulfilment of the requirements for the degree of MSc Information Technology Management

Faculty of Engineering & IT

Dissertation Supervisor

Dr. Sherief Abdullah

December-2015

DISSERTATION RELEASE FORM

Student Name	Student ID	Programme	Date
Ashjan Ahmed Saleh Saleh	**2013210022**	**ITM**	**10-May-16**

Title
Education Data Mining to Predict Student Exam Grades in Vocational Institutes

I warrant that the content of this dissertation is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

I understand that one copy of my dissertation will be deposited in the University Library for permanent retention.

I hereby agree that the material mentioned above for which I am author and copyright holder may be copied and distributed by The British University in Dubai for the purposes of research, private study or education and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

Electronic Submission Copyright Statement
Please choose one of the following two licenses and check appropriate box.

☒ I grant The British University in Dubai the non-exclusive right to reproduce and/or distribute my dissertation worldwide including the users of the repository, in any format or medium, for non-commercial, research, educational and related academic purposes only.

Public access to my dissertation in the Repository shall become effective:
☒ Immediately ☐ 24 months after my submission
☐ 12 months after my submission ☐ 48 months after my submission

☒ I grant The British University in Dubai the non-exclusive right to reproduce and/or distribute my dissertation to students, faculty, staff and walk-in users of BUiD Library, in any format or medium, for non-commercial, research, educational and related academic purposes only.

Signature
Ashajn Saleh

Dedication

Every challenging work needs self-efforts in the first step, as well as, guidance and support from people who are very close to our hearts.

I would like to dedicate my humble effort to the soul of my beloved mother, Alia Nassar, who passed on a love of reading and respect for education.

I would like to express my deepest gratitude to my husband, Oday Al Obaidi, for his endless affection, love, encouragement, support and prayers that make me able to get such success and honor.

I would like to lovingly dedicate this thesis to my respective father, sisters and brother, Ahmed Saleh, Falasteen Ahmed, Salam Ahmed, Saja Ahmed, Taqwa Ahmed, Hala Ahmed, and Saleh Ahmed, who have been my constant source of inspiration and motivation.

I would like to extend my thanks to my supervisor, Dr. Sherief Abdulla, who offered academic guidance and support throughout all stages of the dissertation.

Abstract

Data Mining (Data Knowledge Discovery) is the practice of examining large pre-existing databases in order to identify patterns, establishing relationships, and to generate new understandable and useful information. Educational Data Mining (EDM) is a new emergent field that concerns about applying data mining methods and algorithms on data residing in educational system repositories. It can be used in the purpose of discovering valuable knowledge such as learning environment, students' performance, dropout, and even students at more risk.

In this thesis, classification modeling will be used as a data mining technique to predict students' performance in the theoretical exam (Final Test) for computer science (CS) course.

The analysis shows that the best prediction accuracy is obtained when CART classification algorithm is used as a classifier. The analysis also shows that certain student assessments are ineffective in predicting the student performance in the final exam. Our analysis also shows certain student features such as the gender the campus are good predictors to student performance.

الخلاصـــــــــة

استخراج البيانات (اكتشاف دلالة البيانات) هي عملية فحص لقواعد البيانات الكبيرة والمخزنة سابقا ، و ذلك من أجل تحديد الأنماط وتاسيس العلاقات و أنتاج معلومات جديدة و مفهومة ومفيدة. استخراج البيانات التعليمية (EDM) وهو عبارة عن مجال مبتكر يعنى بتطبيق أساليب استخراج البيانات والخوارزميات على البيانات المخزنة في مستودعات النظام التعليمي و في الامكان استخدامه كوسيلة في اكتشاف معلومات قيمة مثل بيئة التعلم، ومستوى أداء الطلاب، و احتمالية اسقاط الطالب للمادة التعليمية، وكذلك تحديد الطلاب المعرضون للرسوب في المقرر.

في هذه الأطروحة، سيتم استخدام نماذج التصنيف كأسلوب لإستنباط البيانات للتنبؤ بأداء الطلاب في الامتحان النظري (الاختبار النهائي) لمادة علوم الحاسوب.

يظهر التحليل أن أدق تنبؤ يتم الحصول عليه عند استخدام خوارزمية التصنيف CART و يظهر التحليل أن أداء الطلاب لفروض معينة غير فعال في توقع نتائجهم في الامتحان النهائي.

ويوضح تحليلنا ايضاً كما يبين التحليل أيضا ان بعض المعلومات المتعلقة بالطالب كالجنس و المدرسة تعتبر كعوامل فعالة للتنبؤ بأداء الطلاب في الاختبار النهائي.

.

Table of Contents

List of Figures

List of Tables

1. INTRODUCTION

This chapter presents the motivation for applying data mining in educational systems, the problem statement, and the goals and the objectives of the thesis dissertation along with the research questions. It also includes uniqueness and significance of the study as the chapter aims to provide a proper understating of the study problem, and ensure that the need of the study is developed and communicated properly.

1.1. Overview of Educational Data Mining

Educational Data Mining (EDM) is an emerging discipline, concerns about discovering unique and valuable hidden knowledge in large-scale data that resides in educational system repositories. EDM field seeks to develop and improve data mining methods and techniques to enhance decision making process in the educational system by utilizing the discovered knowledge making the best use of it in making the decisions (Baradwaj and Pal, 2012).

Educational data mining aims to analyze the students' performance, understand learning behaviors, and highlight factors that affect learning process in a particular educational system in the purpose of increasing passing ratio for the students (Patidar et al., 2015).

Various data mining techniques can be used by Educational data mining to trim down the students failing ratio and provide recommendations to the educational system stakeholders (i.e. students, teachers, researchers and administrators), Where these recommendations might have a significant impact in improving learning process level (Patidar et al., 2015) (Baradwaj and Pal, 2012).

Classification, clustering and association rules are data mining techniques used by Educational Data Mining to predict the student's performance in a particular course. This prediction can help in predicting the students' failure earlier and take proper action to reduce or prevent this failure as much as possible (El Gamal, 2013) (Yadav et al., 2012) (Tair and El-Halees, 2012) (Namdeo & Jayakumar, 2014) (Walters-Williams and Li, 2010).

1.2. Problem Statement

Computer Science (CS) provides an asset to learn all areas of the vocational curriculum. Emerging technologies are currently centralized in modern societies where citizens play vital roles in knowledge economy growth.

The Education system such as vocational education and training schools that are followed in UAE nowadays emphasize on activity-based learning, especially computer science courses focus on the practical tutoring methodology such as portfolio tasks, projects, and lab work while the final assessments involve a lot of information on literacy questions. Students find it very difficult to pass the final assessment even though he or she does well in practical tasks.

To overcome students' low performance in CS courses, it is important to understand and identify factors that affect students' performance in the literacy exam and build a predictive model that can effectively help educators and other academic staff to predict how poorly or how well the student will perform in the final exam. This will allow educators to identify students at risk of failing and potentially intervene to improve their chances of success.

1.3. Research Goal and Objectives

This study aims to make early prediction of a student mark in the final CS exam, in order to reduce failure rate and identify the students who need more attention and help from the school. For instance, the student could pass in a certain course if s/he got 60% or more at the end of the academic year, where the student performance in the academic year is basically measured by his/her performance in the course activities during term 1, term 2 and term 3 and his/her mark in the final exam (End of year exam).

The student mark in the end of year exam is a crucial part in measuring the student's pass or failure in the course. And, because the schools is aiming to reduce the failure ratio in different courses, then it is important to develop a validated and effective set of predictive models that can be used to estimate the students' final exam mark. And give various recommendations for the students, teachers, the course developers, administrators, and other academic staff to evaluate learning process and teaching environment.

This study is aiming to find out to what extent does the vocational competency training packages, which are considered as more task-oriented and performance focused, can help students in gathering and synthesizing their academic literacy (theoretical knowledge). This can also identify about the students information that are at more risk to fail and how accurately failing students can be identified in early stages.

The three objectives of the proposed research are as follows:

1. Identify and select appropriate classification technique to identify the students that are likely to fail their final exam.
2. Identify and select what student's variables (dependent and independent) that can be used as inputs for predictive models.
3. Validate, evaluate, and compare accuracy obtained from each model.

1.4. Research Questions

This research is designed to answer the following three main questions:

- How will traditional data mining techniques perform when applied on students' information in the vocational schools?
- What students information (variables) can have a stronger effect and yield to the best prediction accuracy?

1.5. Uniqueness of This Research

EDM is an emergent research field that recently received a huge interest from the researchers due to its significant benefits in educational institutes. Many studies focused on applying data mining techniques in higher education and academic schools and researches were more interested in analyzing students' academic subject (class tests), attendance, demographic data and web usage. This study will be the first study that is aiming to analyze data and extract knowledge hidden in the vocational institute's database in the GCC where this research will focus on analyzing students' technical (Practical) skills, prior knowledge (GPA), gender, and students' location at different vocational institutes across UAE.

2. LITERATURE REVIEW

2.1. Overview about data mining and educational data mining

Data Mining is an analytic process designed with a great potential to help companies discover patterns and trends that go beyond simple analysis. Data mining uses sophisticated mathematical algorithms to explore the data, discovering patterns and hidden relationships between variables and evaluate the probability of future events.

(Baradwaj and Pal, 2012) claim that the process of extracting knowledge from big data consists of 7 sequenced steps:

1. Data integration: The process of collecting and integrating data from all different sources.
2. Data Selection: The process of selecting and identifying the data that will be useful to use in a particular case.
3. Preprocessing: Also called data cleaning; this step represents process of getting rid of data anomalies such as missing values, errors, noise, and inconsistent data.
4. Transformation: Process of converting data into forms appropriate for mining technique.
5. Data mining: Application of data mining techniques such as classification, clustering, and association rule on the transformed data for knowledge discovery purpose.
6. Evaluation and presentation: In this step, data patterns are evaluated, visualized, and transformed.
7. Decisions/Use of Discovered Knowledge: In this step, knowledge is presented and users can make use of acquired knowledge to make better decisions.

Educational data mining (EDM) is a new emerging discipline that focuses on exploring data residing in educational system repositories, and applying methods and techniques to better understand the teaching and learning processes in the educational system (Yadav et al., 2012)(Baradwaj and Pal, 2012) (Romero et al., 2013) (Shi et al., 2013).

By digging deep into the educational system, it can be found that different users and stakeholders are involved and can be benefited from EDM as follows:

- Learners: Educational data mining offers several advantages for students where it can help in improving student performance by recommending tasks, activities, and resources based on pedagogical behavior. It can also make early prediction for student performance and highlight his/her weakness points (Yadav et al., 2012).
- Educators: Analyzing students' data can help teachers understand and improve learning process by determining effective activities, identifying weaknesses in their teaching methods, classifying students and detecting students who need more support and help (Yadav et al., 2012)(Baradwaj and Pal, 2012).
- Researchers: EDM helps researchers in evaluating course structure and improving course contents, and hence, increase institutional effectiveness (Baradwaj and Pal, 2012) (Pandey and Pal, 2011).
- Administrators: EDM helps administrators to make the best utilization of institutional resources (Human and materials) and enhance decision making process (Naqvi, 2015).

2.2. Predicting student performance (predictive models)

A student's academic performance is commonly measured by examinations or continuous assessment. These assessments fall into three types:

- Formative Assessment: occurs in the short term in order to provide ongoing feedback that can be used by instructors to improve their teaching and by the students to improve their learning such as (such as quiz and homework) (Bennett, 2011).
- Interim Assessment: Takes place occasionally or at the end of course where educators use this kind of assessments to evaluate where the students are in their learning progress and determine whether they are on track to performing well on future assessments (Bekkink et al., 2012).

- Summative Assessment: The goal of summative assessment is to evaluate the student learning at the end of an instructional unit by comparing it against some standard or benchmark. (Herppich et al., 2014).

The students' final grade is basically composed of student result throughout the academic year and his/her result in the final exam where the final exam makes up the largest percentage of the final grade.

The student performance throughout the whole academic year plays an important role in predicting his/her outcome in the final assessment. Not only that, some other factors like student's demographics, past performance, prior knowledge, attendance, gender and location might also have a great impact on the prediction process (Yadav et al., 2012) (Bhise et al., 2013).

2.3. Assessment challenge facing Vocational Education and Training (VET) systems

The vocational education and training system is best known as school-based system where students are prepared for a specific work domain by developing their expertise in techniques related to technology, skill, and scientific technique to span all aspects of that specific work domain (Souto-Otero and Ure, 2012).

Vocational schools focus on procedural knowledge rather than declarative knowledge (Read & Hecker, 2013), hence VET system depends more on developing practical skills of a student in a specific work domain using hands-on, practical, and experimental activities.

The primary objectives of VET Systems are to prepare students for: (1) entry into the job market and advancement in their chosen careers. (2) Progression into international higher education. The changing nature of skills needed nowadays for achieving VET systems objectives necessitate the need of enhancing quality of the vocational education where this enhancement requires not only improving technical skills of VET students, but also enhancing their academic skills, hence VET students will have the ability to demonstrate a depth understanding of underlying theories of the work practice, meet job market requirements, and be gainfully employed after graduation. Furthermore, they can continue their education and training in universities (Eynon, 2013).

Testing and assessments are considered as the most common problems associated with VET Students. (White, 2014) explained that information literacy including writing skills, reflections, feedback, research skills, explanation, and assessment practices are important skills for VET students to have, and these skills can effectively improve their academic performance, and enhance acquisition of the skills needed for successful careers.

2.4. Need for predicting Student performance

Many research papers investigated need for predicting student performance in schools and higher education. Prediction of student academic performance is emerging as a research area with a suite of computational methods and research approaches for a number of reasons.

(1) Predicting a student performance using predictive models can help teachers in taking proactive actions to enhance the students' performance (Siemens and Baker, 2012).

Predictive models provide different tools that can help the teachers to better understand and classify students into groups, as well as detect students who need more support and who are at more risk to fail. For example, a teacher can adopt different instructional strategies such as review/check, observation, demonstration…etc. with students who are predicted by predictive model as potentially at risk. Furthermore, results from predictive model can also provide teachers with objective feedback about tasks, activities and assessments used in the course and make proper changes to the course design.

(2) Results of the predictive model can help course developers and researchers to evaluate course structure and modify course curriculum to improve learning process and enhance student achievement (Romero et al., 2013).

(Shi et al., 2013) have utilized data mining methods and visualization tools in their study to analyze learning behavior patterns for Computer Science students at the University of Warwick. This study investigated students interaction with Topolor system, which is a social adaptive personalized e-learning system, and results showed that analysis of action frequency (number of times a student perform a particular action or task) as well as action sequence (the order of performing the actions) for students while doing an online course. This will help course developers and teachers to identify common and unique learning behavior patterns, and make proper adjustments to course structure.

(3) Students themselves can also benefit from the predictive model by developing good understanding of their learning behavior. Predicted results can help the students identify their strength and weakness points in the course, recommend different activities and give them the opportunity to rethink about their learning strategy.

(Baradwaj & Pal, 2012) applied the decision tree method on student database in order to evaluate students' performance and predict their final grades in a particular course. Early prediction of student performance can help him/her identify areas of the course that need special attention and more effort to work on.

2.5. Predictive Model

This section explains important components of predictive models: Feature selection, testing and training data subsets, classification algorithms, and evaluation measures for the predictive models.

2.5.1. Features Selection:

In data mining, feature selection task is applied to identify relevant and useful attributes from dataset that contribute to the accuracy of a predictive model and in fact to increase the accuracy of the model, and remove unneeded, irrelevant, and redundant attributes from dataset. Feature selection is important for several reasons; it improves the interpretability of the classifier, increases prediction performance, and provides faster and more cost-effective predictors (Awada et al., 2012) (Lazar et al., 2012).

In classification context, most commonly used approaches for selecting optimal subset attributes for building classification model are:

- Filter approach: This approach basically measures relevance for all attributes in the data set (relevance score) and then attributes with lowest relevance score will be removed from the model (Xue et al., 2013).

- Wrapper approach: This approach measures feature usefulness by searching and evaluating the space of all possible subset features (Rodrigues et al., 2014).

Comparing filter approach with Wrapper approach; filter approach is simpler, faster and has no relation with the classification algorithm (independent) (Bermejoet al., 2012), but the most commonly drawback is filter method which may fail to select most useful feature because relationships between variables are not taken into consideration in this method. On the other hand, Wrapper method has the ability to identify most useful feature as it considers features dependencies, but overfitting risk and computation time are higher when number insufficient observations and variables are large (Cadenas et al., 2013) (Acharya and Sinha, 2014).

(Rai and Kumar, 2014) conducted a comparative study on minimizing set of features in the database for the purpose of making training and testing through a classifier more efficient and increasing the accuracy of a classification algorithm. The results from this study show that the maximum numbers of instances that are correctly classified are obtained by applying wrapper feature selection method. Filter feature selection method can be also used to reduce the overall complexity of the classification process.

2.5.2. Algorithms

In order to deliver a meaningful analysis in educational sector, many data mining techniques and methods have been used by researchers to classify students and predict their final marks. Decision Trees, Neural Networks, and Naïve Bayes are the most widely used techniques as predictive models in educational data mining.

2.5.2.1. Decision Tree:

Decision tree is a kind of predictive model in data mining, which maps observations about an item to conclusion about the item's target value. Decision Tree builds classification model in the form of a tree structure and generate rules (conditional statements) that can easily be understood by humans and easily used within a database to identify a set of records.

Most commonly Decision trees algorithms used by researcher in the educational field are:

1. **Iterative Dichotomiser 3 (ID3) Algorithm:**

   Quinlan Ross developed a core algorithm for building decision trees called ID3. ID3 is a recursive algorithm that employs a top down greedy search through the space of possible branches with no backtracking. ID3 simply uses a fixed set of examples to build the decision tree, and later the developed decision tree will be employed to classify new future samples (Quinlan, 1979).

   The ID3 constructs a decision tree based on information gain/entropy (Baradwaj and Pal, 2012). The aim is to start with splitting the original set on the attribute which provides the maximum gain or the least entropy. This will either generate leaf nodes (single class) or nodes with multiple classes. The algorithm is applied repeatedly on the new non-leaf nodes until we get leaf nodes and there is no need to split further.

   Entropy and Information Gain measures are used by ID3 to construct a decision tree. Entropy is a formula that calculates the homogeneity of a sample where the completely homogeneous sample has entropy of 0 and equally divided sample has entropy of 1. Entropy uses the formula described below to calculate the homogeneity of a training set (T) (Patidar et al., 2015)

$$\text{Entropy (T)} = -\sum_{i=1}^{n} Pi \log_2 Pi$$

Where, n: different values of the target attribute and Pi: is the probability of T belonging to class i.

Information Gain: This is the measure of the difference in entropy from before to after the set is split on an attribute. Information Gain calculates a numerical value for a given attribute, A, with respect to a set of examples, S in order to determine the best attribute to choose for a particular node in a tree.

The information gain of attribute A related to a collection of examples, T, is calculated as (Patidar et al., 2015):

$$Info\ Gain = Entropy(T) - Entropy(A, T)$$

2. **Classification And Regression Trees (CART) Algorithm:**

CART Algorithm Developed by 4 statistics professors between periods from 1974 till 1984. CART was introduced with reference to classification and regression trees, so that it can handle categorical and continuous variables to build a decision trees. Classification tree is used to identify the class that a categorical target variable would likely fall into. On the other hand, Regression tree handles prediction of continuous target variables.

CART works through recursive partitioning of the training set in order to obtain subsets that are as pure as possible to give the target class. The main steps of CART algorithm are:

1. Splitting each node in a tree: Rules for splitting data at a node based on the value of one variable.
2. Deciding when a tree is complete: In other words, deciding if a branch is terminal or it can be more split.
3. Prediction for the target variable in each terminal node by assigning each terminal node to a class outcome.

GINI Index is the default impurity measure used in CART for categorical target variables. It is essentially a measure of how well the splitting rule separates the classes contained in the parent node (El Gamal, 2013).

GINI Index at node t is defined as follows:

$$\mathrm{GINI(t)} = \sum_{j \neq i} p(j/t)\, p(i/t)$$

Where, t is a particular node, i and j are categories of the target variable, and p (j/t) and p (i/t) refers to the proportion of target category j or i present in node t.

Least Squared Deviation (LSD) and Least Absolute Deviation (LAD) are impurity measures used in CART for continues target variables.

3. C4.5 Algorithm

C4.5 was developed by Quinlan in 1993, as an extension for ID3 algorithm in order to handle problems that ID3 couldn't deal with. C4.5 generates a decision tree where each node splits the classes based on the gain of information.

C4.5 accepts both continuous and discrete features. It can also handle incomplete data points, as well as, it can solve over-fitting problem by (very clever) bottom-up technique usually known as "pruning".

(Yadav and Pal, 2012), applied ID3, CART and C4.5 decision trees algorithms on engineering student's data that includes demographic data, students' past performance, address and contact number to make early prediction of students' performance in engineering final exam. Results from this student show that on one hand decision tree algorithms helped in predicting number of students who are likely to pass, fail in the course. Comparative analysis of the study showed that 0.786 of students was truly predicted as fail using ID3 and C4.5 Classifiers. CART recorded 0.643 as a true positive rate for fail students in the course. On the other hand, shortest time required to build the predictive model is 0.00 seconds recorded for ID3 classifier.

(Yadav et al., 2012) conducted a research to study students' performance in the courses and identify low achievers students who need special attention from the teacher. ID3, C4.5 and CART algorithms are applied on students' academic database. Results from the experimental show that although time needed to build predictive model using CART classifier is the longest time among three classifiers, accuracy obtained from CART compared other methods.

(Baradwaj & Pal, 2012) applied ID3 algorithm on students' academic database of Master of Computer Applications course at Computer Applications department in VBS Purvanchal University. Experiment was conducted in the purpose of describing and evaluating students' performance in end semester examination. Result shows that ID3 classifier can help reducing fail ratio by making early prediction of student performance. Entropy and Gain Information are used in this study to measure degree of impurity in dataset and enhance prediction results.

2.5.2.2. BAYESIAN CLASSIFICATION

Bayes' theorem is a mathematical method that calculates given occurrences in prior trials; the likelihood of a target occurrence in future trials. According to Bayesian logic, the only way to quantify a situation with an uncertain outcome is through determining its probability.

The Naive Bayesian classifier is based on Bayes’ theorem with independence assumptions between predictors. It assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors.

To simplify, Bayes’ theorem states the following relationship:

$$\text{Posterior} = \frac{Prior \times likelihood}{evidence}$$

For a given a class (C) and a class variable (V) Bayes’ theorem states the following relationship (Efron, 2013):

$$\mathrm{P}(c|v) = \frac{P(v|c)P(c)}{P(v)}$$

- P(c|v) is the posterior probability of class (target) given predictor (attribute).
- P(c) is the prior probability of class.
- P(v|c) is the likelihood which is the probability of predictor given class.
- P(v) is the prior probability of predictor.

Naive Bayesian classifier is referred to as naive since it makes the assumption that each of its inputs are independent of each other and an assumption which rarely holds true, and hence the word naive.

Naive Bayesian algorithm is seen as a simple powerful tool in the world of classification and machine learning. Naive Bayesian requires small amount of training data to estimate the parameters. It can solve problems involving categorical and continues values and it can be easily implanted and used.

(Pandey and Pal, 2011) analyzed data collected from different colleagues affiliated with Dr. R. M. L. Awadh University using naïve Bayes classification algorithm. Result from this study shows that naïve Bayes classifier can help teacher and students reduce drop out ratio and improve performance level of the students and institution.

(Tair and El-Halees, 2012) recommend application of naïve Bayes classification on students database where naïve Bayes can predict a student grade and help him/her improving performance before the graduation. In this study, naïve Bayes classifier was applied on graduate students' database from college of Science and Technology in Khanyounis. Data collected includes 3360 records and 18 attributes for a period of fifteen years [1993 to 2007]. Result from this study shows that naïve Byes classifier has an acceptable accuracy of 67.5% in predicting graduate students' grades.

(Namdeo & Jayakumar, 2014) analyzed students data collected at the University of National and World Economy using rule learner, decision tree, Bayes and Nearest Neighbor classifiers. Results of the conducted research show that the Bayesian classifiers recorded the lowest accuracy among all other classifiers where average accuracy of the Bayesian classifiers is below 60%, the decision tree classifiers 66-67%, Nearest Neighbor classifiers 59-60%, and rule learners 63%.

2.5.2.3. Nearest Neighbor Classifier

The main nearest neighbor method concept is to find closest distance of the new point to predefined training samples in dataset. K nearest neighbors (KNN) is a simple algorithm that classifies and stores all available cases, and then classifies new cases based on a similarity measure (e.g., distance functions). KNN is a non-parametric technique used for classification and regression. The number of samples (K) can be a user-defined constant or vary based on the local density of points. For instances, similarities for continuous variables can be measured by different distance functions for continuous variables such as:

1. Euclidean Function: Distance between two points in Euclidean space is called Euclidean Distance (Walters-Williams and Li, 2010).

$$EUD(p,q) = \sqrt{\sum_{i=1}^{x}(p_i - q_i)^2}$$

2. Manhattan Function: The distance between two points measured along axes at right angles is called Manhattan Distance and it is calculated as follows (Walters-Williams and Li, 2010) :

$$MN(p,q) = \sum_{i=1}^{x}|p_i - q_i|$$

3. Minkowski Function: It is a metric in a normed vector space which can be considered as a generalization of both the Euclidean distance and the Manhattan distance (Amatriain et al., 2011):

$$d(x,y) = \left(\sum_{k=1}^{n}|x_k - y_k|^r\right)^{\frac{1}{r}}$$

In the instance of categorical variables, the Hamming distance must be used. The Hamming Distance is the number of digit positions in which the corresponding digits of two binary words of the same length are different (Ahmad and Dey 2007). For example, distance between "Male" and "Female" is one according to the following function (Walters-Williams and Li, 2010):

$$\mathrm{D}_H = \sum_{i=1}^{K}|x_i - y_i|$$

$$x = y \rightarrow D = 0$$
$$x \neq y \rightarrow D = 1$$

(Anuradha and Velmurugan, 2015) conducted a comparative analysis in three private colleges in Tamil Nadu state of India, with the aim of evaluating decision tree, Bayesian classifiers, K Nearest Neighbor algorithm, and two rule learner's algorithms performance in predicting students' performance at the end of semester examination. Results from this study reveal that KNN classifier correctly classified 68.32% of dataset using 10-fold cross-validation testing option. On the other hand, true positive rate was zero for predicting the class Fail.

2.6. Evaluating the Performance of Predictive Models (Comparative Analysis):

2.6.1. Splitting data to testing and training subsets:

In predictive model development process, model validation is possibly the most important step in the building sequence generating a valid, accurate, and efficient predictive model requiring assessing how the results of a predictive model will generalize to an independent data set.

Developing predictive model is usually done by training the model on a dataset of known data (training dataset), and then testing the predictive model performance on a no previously seen dataset (testing dataset) (Pinto et al., 2012) (Steyerberg et al., 2001) (Moons et al., 2012).

Model validation includes internal or external model validation. Internal validation is process of determining model performance on a new data, which is randomly selected from underlying population as a testing data .in other words. Data for model development and evaluation are both random samples from the same underlying dataset (Steyerberg et al., 2001). On the other hand, External Validity of a predictive model can be done by testing the model on related, but different dataset that is used during the model development process. External validation is also known as generalizability or transportability (Steyerberg et al., 2010) (Moons et al., 2012).

Splitting dataset into training data and testing data is the most common approach used in internal validation. In case of limited dataset where dataset cannot afford splitting process, then cross validation or bootstrapping approaches can be used (Pinto et al., 2012).

Cross validation is a model evaluation method, sometimes called rotation estimation that basically consists of two phases, training and result generation. Cross validation splits sample data during training phase into complementary subsets where all subsets except one subset are used for performing analysis, and the remaining one is used to test or validate the analysis (Hall et al., 2009). Cross validation usually runs on multiple rounds (K) and validation results average is calculated over rounds. For each K experiment, K-1 subsets (folds) are used for training and the remaining one is used for testing, and as a result, all the examples in the dataset are eventually used for both training and testing (Steyerberg et al., 2001) (Olson et al., 2012).

Bootstrapping approach is another validation technique that basically draws samples of data from a single dataset. Bootstrapping involves using some form of resampling by (uniformly) randomly selecting items and duplicating them (with replacement) (D'Haen et al., 2013) (Austin et al., 2013).

(Kabakchieva, 2013) conducted a study at Bulgarian university, which predicts the students' performance at the university using the students personal and pre-university information. In this study, two testing options are used to validate the predictive models; Cross validation (10 Folds) and the Percentage Split. A comparative analysis is conducted to evaluate the results obtained by each classifier.

(Patidar et al., 2015) proposed a student analysis system to help students improve their technical and academic skills. The proposed system is mainly developed by applying C4.5 and justified C4.5 algorithms on student's database where bootstrapping technique is used to validate the proposed model. The comparative analysis in the study shows that accuracy percentage obtained by applying enhanced C4.5 algorithm on real time dataset is 79.10%, meanwhile C4.5 algorithm recorded 73.76%.

(Yadav and Pal, 2012) applied 10-fold cross-validation technique in order to evaluate the performance of three predictive models developed using ID3, C4.5 and CART algorithms. Results show that accuracy obtained from applying three algorithms varies from (62.22-67.77) % and time required to build three models is between (0.00-0.09) seconds where highest accuracy recorded by C4.5, and lowest execution time was for ID3 model.

2.6.2. Evaluation Measures:

For predictive model, there are a couple of measures to evaluate the performance of the model. For regression problem, three measures are commonly used to measure distance between the estimated outputs from the actual output:

1. Root Mean Square Error (Dormann et al., 2013).
2. Relative Square Error
3. Coefficient of Determination

For classification problem, several measures have been proposed to evaluate models performance such as Accuracy, Specificity, Recall, or Sensitivity and Precision (Verbraken et al., 2014).

1. Accuracy = (TP + TN) / (TP + TN + FP + FN)
2. Specificity = TN / (FP + TN)
3. Recall or Sensitivity (Completeness) = TP / (TP + FN)
4. Precision (Exactness) = TP / (TP + FP)

According to Confusion Matrix below:

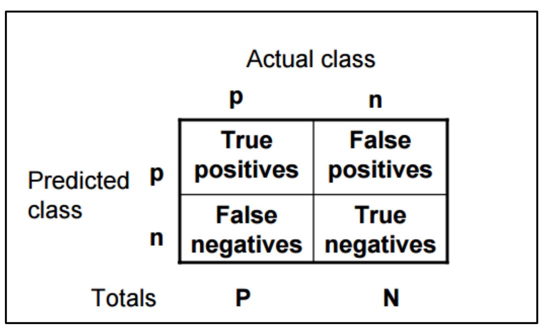


Figure 1: Confusion Matrix

- TP = the proportion of positive cases that were correctly identified. For instances, Predict as Positive when Actual Positive.
- TN = the proportion of negatives cases that were classified correctly, Predict negative when Actual positive.

- FP = the proportion of negatives cases that were incorrectly classified as positive. Predict Positive when Actual Negative.
- FN = the proportion of positives cases that were incorrectly classified as negative. Predict Negative when Actual Positive.

(Ramaswami and Bhaskaran, 2009) (Powers, 2011) Explain that the classifier accuracy alone is not sufficient to represent the quality of prediction because the cost of making a FP may be different from the cost of making a FN. F measure is another way to evaluate performance of the predictive model. F-measure can be more meaningful as it combines recall and precision into a global measure and calculated as below:

F-Measure = 2*((precision*recall) / (precision +recall)).

3. RESEARCH DESIGN

This chapter describes steps are required to develop different predictive models using 10 attributes as predictors, three modeling techniques (Decision Trees, Neural Networks and Naïve Bayes) and three different sample sizes.

The models were developed and validated on students' academic data, that was collected from 6 vocational institutes, for a period of three semesters from 2014 to 2015, and some other factors related to the students and schools. Feature selection approach is used to identify the most appropriate features as predictors. The measures used to evaluate and compare the models, are also defined in this chapter.

3.1. Predictive Model Development

Based on the extensive literature review in educational data mining field, the figure below shows the sequences of the steps used to extract knowledge and relationships were hidden in the educational database.

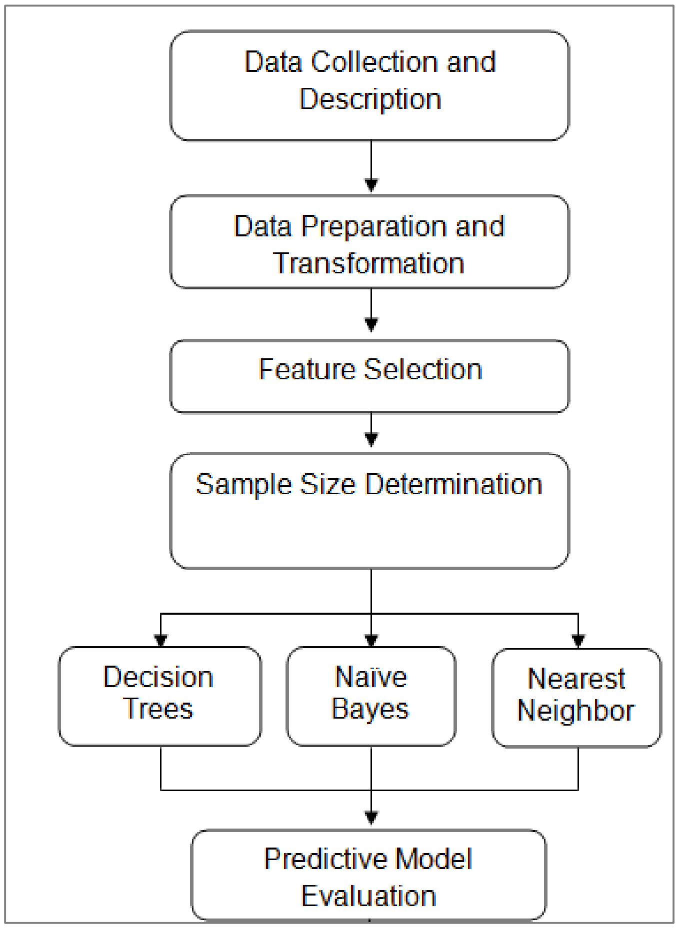


Figure 2: Modeling Framework

3.2. Data Collection and Description

Original Dataset was collected from students who were enrolled for CS course, in 6 vocational institutes across UAE, during the period 26th Aug. 2014 till 9th July 2015.

The dataset records represent the students' demographics, and information about the students' performance throughout the whole academic year. The dataset consists of a total 1056 records and 10 attributes such as Student ID, Student Name, Campus, Gender, GPA, Term 1 mark, term 2 mark, Term 3 mark and End of year assessment.

As seen in the figures below; the majority of 1056 students have their major in mechanical and electrical engineering which is approximately 402 students (38%). 54% are male students and 34% of overall students are from Ajman School.

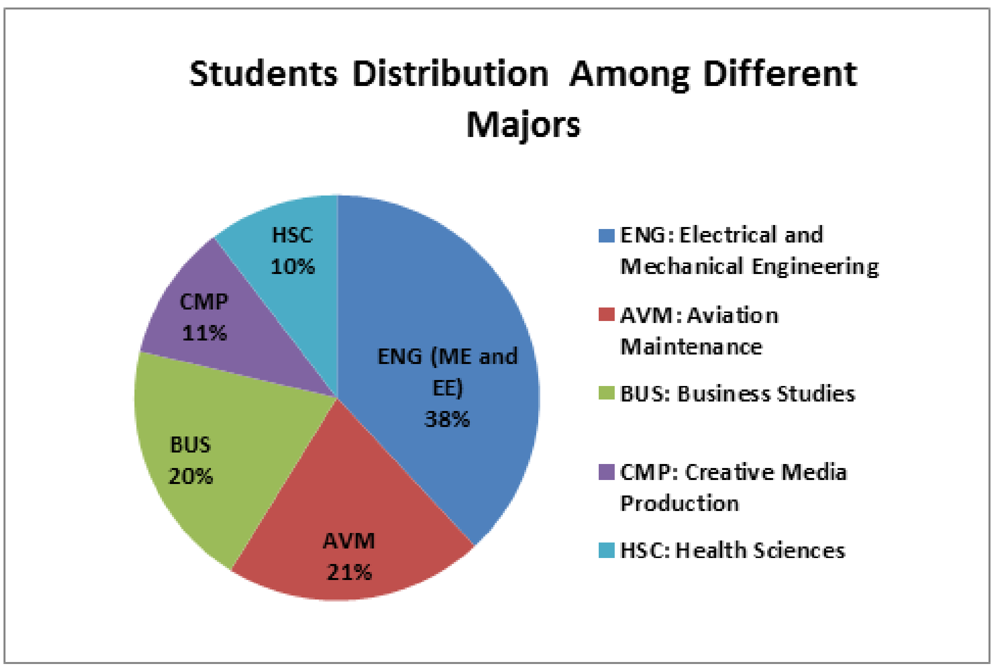


Figure3.1: Students Majors

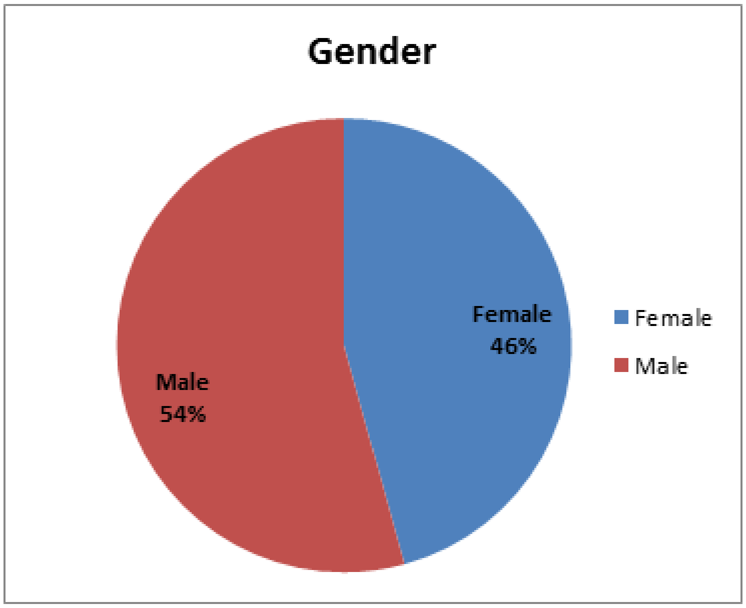


Figure3. 1: Students Gender

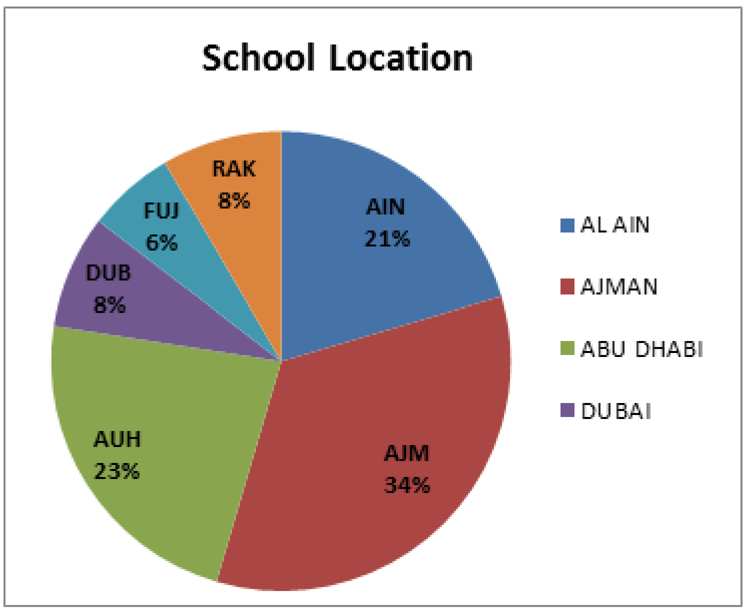


Figure3. 2: Schools Locations

The dataset is used in this paper, originally consists of 10 features. Table1 shows the dataset features, type of data and description for each feature as taken from the database:

Feature	**Data type**	**Description**
eSIS	Number	The student ID
Student Name	Text	The full name of the student
Campus	Text-Categorical	The campus location for the student: • Al Ain • Ajman • Abu Dhabi • Dubai • Fujairah • Ras Al Khaima
Gender	Text-Categorical	The gender of the student • Male • Female
Cluster	Text-Categorical	The major of the student: • ENG • AVM • BUS • CMP • HSC
GPA	Number-continuous	Accumulative GPA of the student. Ranges from 0.00- 4.00
Term 1	Integer	The student performance during the term. It ranges from (0-100) %, where : - 20% for Practical worksheets (portfolios) - 20% for labs/Projects - 20% for continual assessments - 40% for final Competency test.
Term 2	Integer	The student performance during the term. It ranges from (0-100) %, where : - 15% for practical worksheets (portfolios) - 15% for labs/Projects - 35% for continual assessments - 35% for final competency test.

Term 3	Integer	The student performance during the term. It ranges from (0-100) %, where : - 10% for Practical worksheets (portfolios) - 10% for Labs/Projects - 20% for Continual assessments - 60% for final Competency test.
EOY	Integer	End of Year mark: The student performance in the final unified written assessment among all schools. student mark ranges from 0-100

Table 1: Student's Dataset Description

In this VET system the student will fail in CS course, if s/he got 60% or less as a final mark. The student performance in the course (Final mark) is basically calculated according to the following formula:

Final Result= (30%* term 1) + (20%* term 2) + (20%* term 3) + (30%* EOY).

Based on the dataset description in table1, the student practical and academic skills (theoretical skills) are assessed as the following during the whole academic year:

- Term 1 and Term 3: 80% of the student mark is awarded for practical skills (portfolios, projects and competency test) and 20% for theoretical skills.
- Term 2: 65% for practical tasks and 35% for theoretical skills.
- EOY: 100% for theoretical skills.

Two features from the original dataset have been removed; eSIS and the student name. Reasons for removing these two features are: 1) it holds confidential information. 2) These features will have no effect, and will add no value to our analysis. And hence, the final dataset is used in this study consists of 8 features and 1056 records.

3.3. Data Preparation and Transformation

The dataset is used in this study, initially collected from different campuses and stored in different formats. So, before working on the data and applying the different analytics, the dataset is preprocessed and prepared as follows:

1. Data was combined into one unified table. Errors resulted from the joining process were removed.
2. Some attributes considered as irrelevant attributes because they only include private information about the students (i.e. Student ID and Student Name). These attributes were excluded and removed as they don't give any knowledge and don't have any importance.
3. To simplify data description, numeric attributes were discretized into categorical ones as follows:

Attribute	**Possible Values**	**Normalized value of Data**
Term 1, Term 2, Term 3 and EOY	>= 90	A
	75 - 89	B
	60 - 74	C
	< 60	D
GPA	>=3.50	Excellent
	3.00-3.49	Very Good
	2.50-2.99	Good
	2.00-2.49	Average
	0.00-1.99	Poor

Table 2: Data discretization

The figure below shows the students distribution over classes (A, B, C and D) in Term1, Term 2, Term3 and EOY:

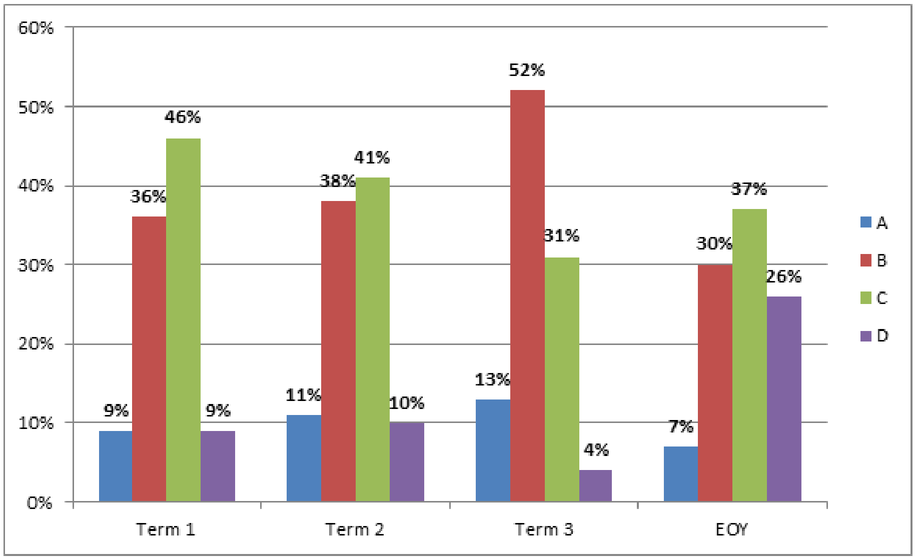


Figure 4: Students Distribution over Classes

4. Dataset contains 6 missing values; records with missing value are excluded from the dataset.
5. After the data collection, preparation and transformation, the dataset is finally composed of 8 attributes and 1050 records and saved as *StudentsInfo.arff*.

3.4. Feature Selection

Based on extensive literature review, before inducing a model, input engineering should be done first. The most useful part of this process is attribute selection (also called feature selection), where relevant attributes are identified and selected. As well as, redundant and/or irrelevant attributes are removed from the dataset. The predictive model accuracy is mainly affected by attributes used as predictors in the model.

In order to achieve a higher prediction performance from the model, *StudentsInfo.arff* dataset is fed into WEKA application which is an open source software that consists of a complete set of machine learning algorithms for data mining tasks that allow user extract and discover useful knowledge from large databases and visualize it), Wrapper feature selection method is used in this study to identify all possible relevant subsets of attributes, and hence, evaluate and select the most relevant subset, as predictors from the list of all variables for the predictive models.

Wrapper method involves an evaluation method and a search method. *WrapperSubsetEval* method is selected in this paper as an attribute evaluator in order to evaluate attribute sets using a learning scheme, and BestFirst search method is used as a search method to search list of attribute subsets using backwards direction. The backward direction mainly starts with all attributes, and then selects one to remove, and so on, so forth until it meets the search termination criteria.

3.5. Sample Size Determination

Based on the extensive literature review in educational data mining and for the purpose of training and testing the predictive model, in this paper dataset was randomly split by WEKA, into training set and testing set as follows:

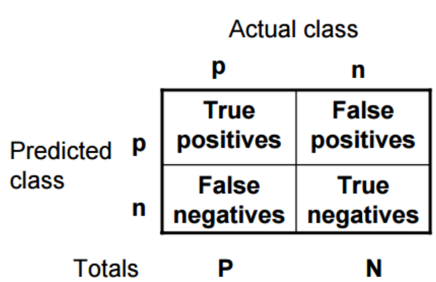
- S1=60% as training set and 40 % as testing set.
- S2=40% as training set and 60% as testing set.
- S3=Cross-validation (10 folds).

3.6. Classification

Various Data Mining techniques can be applied on the educational database, in order to discover hidden knowledge hidden and relationships between different attributes. Many classification algorithms have provided a good understanding in educational resources, and have provided an excellent performance in predicting the students' final results. In this paper k-nearest Neighbor, Naive Bayes and decision trees algorithms have been used to classify the students' results and predict their results for the final assessment.

3.7. Predictive Model Evaluation

In this paper two evaluation measures are used to assess predictive models performance from the confusion metric shown below:



1. Accuracy = (TP + TN) / (TP + TN + FP + FN), where the higher prediction accuracy is, the better the model.
2. Execution Time, where the less time needed for building the predictive model, the better the model.

Finally, a comparative analysis was conducted to compare results obtained from each predictive model, in order to decide the best classifier that can accurately and efficiently predict the student mark in the final assessment. A quantitative analysis was also conducted to explain results obtained from the predictive models.

4. Experiments Setup and Results

This chapter intends to discuss experiments conducted for this paper, explain and summarize the main findings and results of using data mining techniques in the educational sector.

All experiments in this paper are conducted using WEKA Application.

4.1. Feature Subsets:

In this paper we have discussed how the predictive model accuracy is affected by the selection of subset features that are used in the model. Wrapper feature selection method is used in this study to evaluate all attributes and to search for most relevant attribute subsets. Relevant feature subsets that resulted from applying wrapper method *on StudentsInfo.arff* are shown below:

1. Subset 1: 5 attributes; Campus, Gender, Term 1, Term 2 and Term 3.
2. Subset 2: 5 attributes; Campus, Gender, Cluster, Term 1 and Term 3
3. Subset 3: 6 attributes; Campus, Gender, GPA, Term 1, Term 2 and Term 3.

4.2. Classification Algorithms and Sample Sizes

After identifying and selecting the most appropriate and relevant attribute subsets, ID3, CART, C4.5, Naïve Bayes and KNN algorithms are trained on following sample sizes. This detail can be now used to investigate the effect of training set and testing set sizes on the prediction accuracy and also to compare prediction accuracy generated from each model:

- S1=60% as training set and 40 % as testing set.
- S2=40% as training set and 60% as testing set.
- S3=Cross-validation (10 folds).

4.3. Experiments Results

4.3.1. Experiment 1:

In the first experiment, 5 classification algorithms were executed using all the attributes and 3 different sample sizes as mentioned above. Results from first experiment shown in the table below:

Sample Size	Algorithm	Correctly identified Instances	Incorrectly identified Instances	Unclassified Instances	Accuracy	Time Seconds
S1	ID3	51.18%	40.28%	8.53%	56.10%	0
	C4.5	59.72%	40.28%	0.00	59.80%	0.06
	CART	55.69%	44.31%	0.00	55.70%	0.92
	Naïve Bayes	62.09%	37.91%	0.00	61.60%	0
	KNN	59.24%	40.76%	0.00	59.20%	0
S2	ID3	51.89%	39.43%	8.68%	56.90%	0
	C4.5	61.36%	38.64%	0.00	61.30%	0.03
	CART	61.83%	38.17%	0.00	61.80%	0.28
	Naïve Bayes	61.04%	38.96%	0.00	61.00%	0
	KNN	60.09%	39.91%	0.00	60.10%	0
S3	ID3	55.87%	36.74%	7.39%	60.40%	0.02
	C4.5	64.77%	35.23%	0.00	64.90%	0.02
	CART	62.41%	37.59%	0.00	62.50%	0.25
	Naïve Bayes	61.65%	38.35%	0.00	61.40%	0
	KNN	61.74%	38.26%	0.00	61.80%	0

Table 3: Experiment1 Results

Table 3 shows that C4.5 has the highest accuracy of 64.90% among all other classifiers, when it was executed using all attributes and under 10 folds cross-validation. Time needed to build this model is 0.02 seconds.

On the other hand, the lowest accuracy obtained in this experiment is 55.70%, while CART took 0.92 seconds to build the model using S1 sample size (60% training set and 40% testing set) and achieve the accuracy.

This experiment was conducted using all attributes in the dataset, 3 different sample sizes and 5 classifiers. The results from this experiment show the following:

ID3 classifier generated a maximum accuracy of 60.40%, and time required to build the model is 0.02 seconds.

Using all attributes, ID3 best performs when 10 folds Cross-validation is used as test option for the dataset.

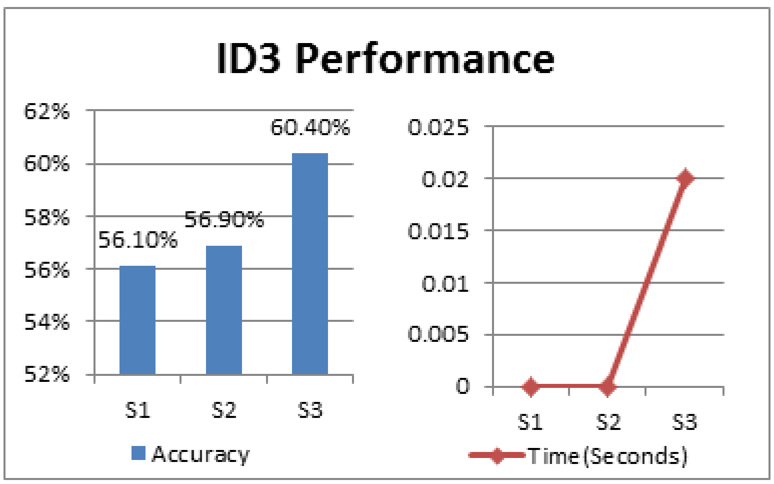


Figure 5.1: Experiment 1-ID3 Performance

The figure below shows C4.5 performance in experiment 1. The results show that the best accuracy (64.90%) was obtained when C4.5 was applied using 10 folds Cross-validation. Furthermore, time required to generate the maximum accuracy is 0.02 second.

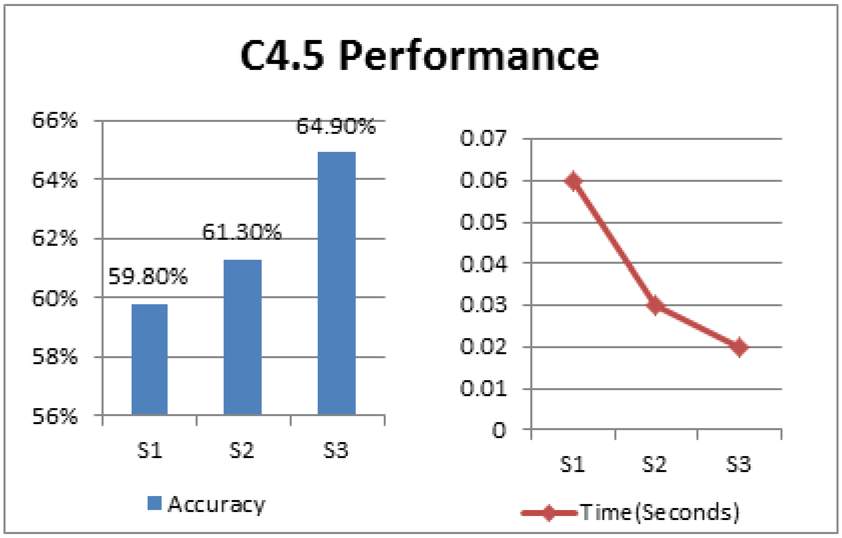


Figure 5.2: Experiment 1-C4.5 Performance

CART has the best accuracy of 62.50% when it was also applied using cross-validation (10 folds). CART execution time was almost similar to ID3 and C4.5 when was applied on the same sample. However, CART required 0.92 second; which is the longest time required by all classifiers in this study, to make the prediction using S1 sample as a test option. Figure 4.3 shows CART Performance in the first experiment.

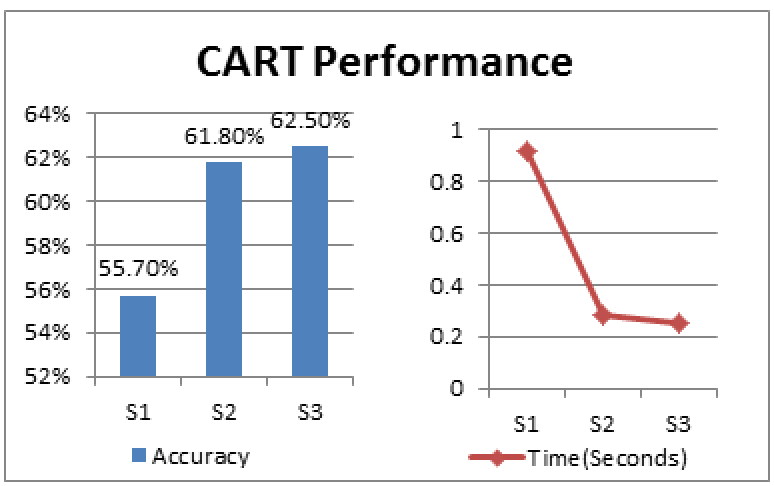


Figure 5.3: Experiment 1-CART Performance

Naïve Bayes in this experiment required 0.00 second to make classification and prediction using the various samples. The highest accuracy recorded for Naïve Bayes is 61.60% using S1 sample.

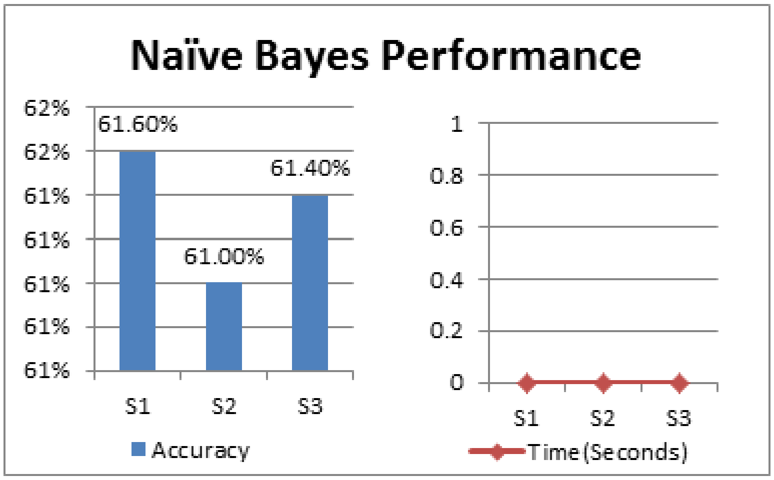


Figure 5.4: Experiment 1-Naïve Bayes Performance

Similarly, in this experiment KNN required the same amount of time as Naïve Bayes to make the prediction using different samples. Yet the highest accuracy recorded by Naïve Bayes was 61.80% using S3 sample. Figure 4.5 shows KNN prediction performance in experiment 1.

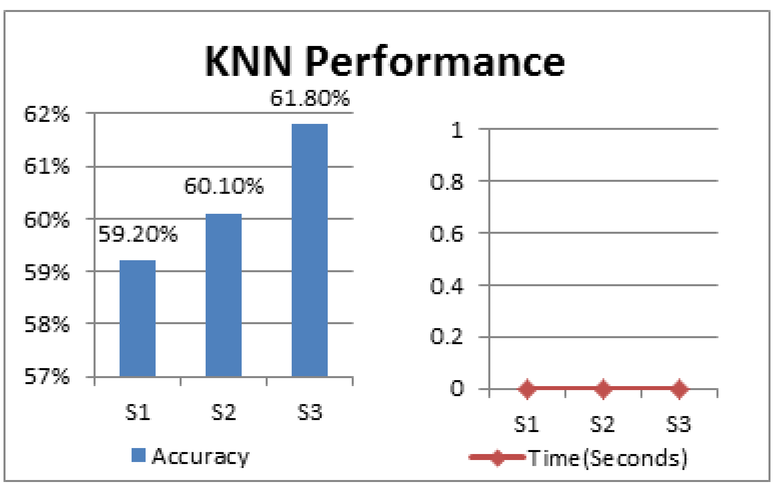


Figure 5.5: Experiment 1-KNN Performance

To summarize; using all attributes provided in that dataset, all classifiers in this experiment did better when cross-validation sample (S3) is used as a testing option, except Naïve Bayes. The best accuracy obtained using cross-validation varies from 60.40%-62.50%, and time required to generate the results range from 0.00-0.25 seconds.

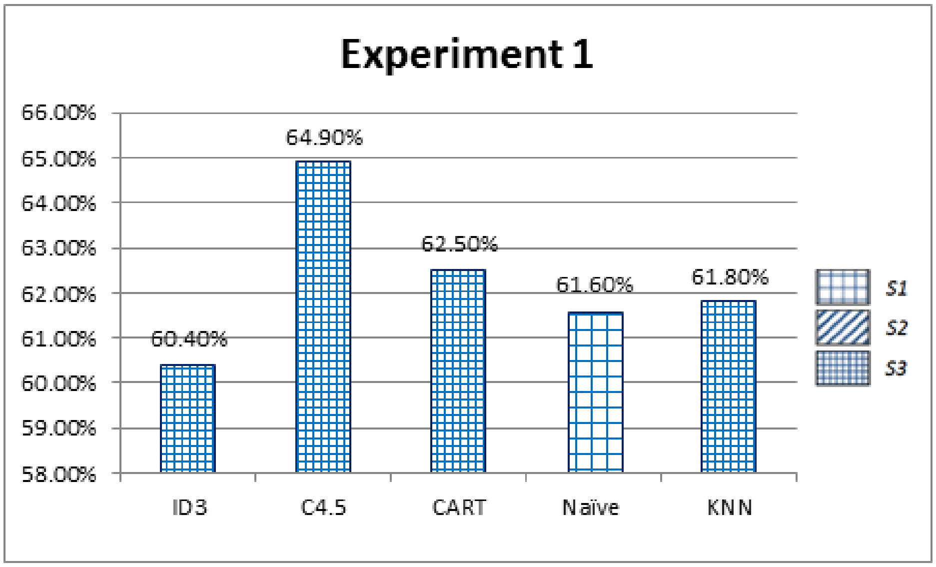


Figure 5.6: The Best Results of all Classifiers in Experiment 1

Based on the assumption that not all attributes are relevant, and can predict student's final mark, previous executions were repeated using attribute subsets generated by wrapper method (subset1, subset2 and subset3), and the three sample sizes (S1,S2 and S3).

4.3.2. Experiment 2:

The second experiment was conducted using attribute subset 1 that includes Campus, Gender, Term 1, Term 2 and Term 3) and 3 different samples (S1, S2 and S3). The results from this study are shown in table 4:

Sample Size	Algorithm	Correctly identified Instances	Incorrectly identified Instances	Unclassified Instances	Accuracy	Time seconds
S1	ID3	57.346%	39.3365 %	3.3175 %	59.30%	0.01
	C4.5	61.3744 %	38.6256 %	0%	61.30%	0.01
	CART	62.7962 %	37.2038 %	0%	62.80%	0.13
	Naïve Bayes	62.5592 %	37.4408 %	0%	62.30%	0.00
	KNN	59.7156 %	40.2844 %	0%	59.70%	0.00
S2	ID3	58.5174 %	37.8549 %	3.6278 %	60.70%	0.00
	C4.5	63.0915 %	36.9085 %	0%	63.10%	0
	CART	61.6719 %	38.3281 %	0%	61.70%	0.12
	Naïve Bayes	63.0915 %	36.9085 %	0%	63.00%	0
	KNN	61.5142 %	38.4858 %	0%	61.50%	0
S3	ID3	63.2576 %	34.9432 %	1.7992 %	64.60%	0
	C4.5	64.678 %	35.322 %	0%	64.80%	0.01
	CART	64.1098 %	35.8902 %	0%	64.30%	0.12
	Naïve Bayes	60.8902 %	39.1098 %	0	60.70%	0
	KNN	64.678 %	35.322 %	0	64.80%	0

Table 4: Experiment2 Results

Figures in the above table reveal that number of correctly identified instances varies from 57.35% to 64.68%.C4.5 and KNN techniques have the highest accuracy of 64.80% compared to the other techniques. The highest prediction accuracy is obtained by C4.5 and KNN, where C4.5 and KNN techniques were executed using cross-validation (10 folds) method, and required 0.00-0.01 second to build the models. On the contrary, the worst prediction accuracy (57.35%) is achieved when it was applied using S1 sample.

Exceptionally, Naïve Bayes best performed, when it is executed using S2 sample; the prediction accuracy is 63% and execution time is 0 second. Contrarily, all other classifiers performed better using attribute subset1 and cross-validation (10 folds) sample.

Detailed results from experiment 2 are presented in the figures below:

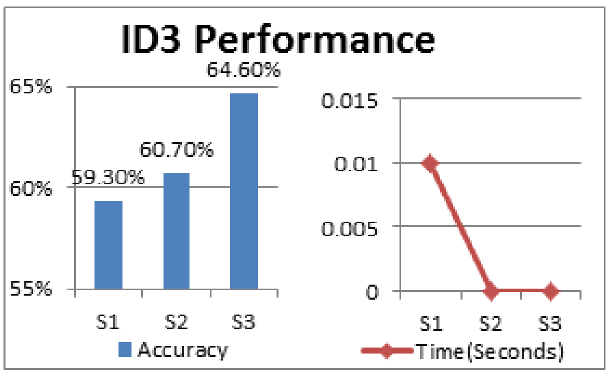


Figure 6.1: Experiment 2-ID3 Performance

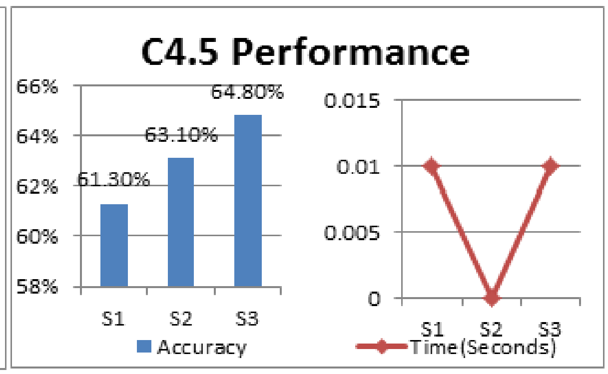


Figure 6.2: Experiment 2-C4.5 Performance

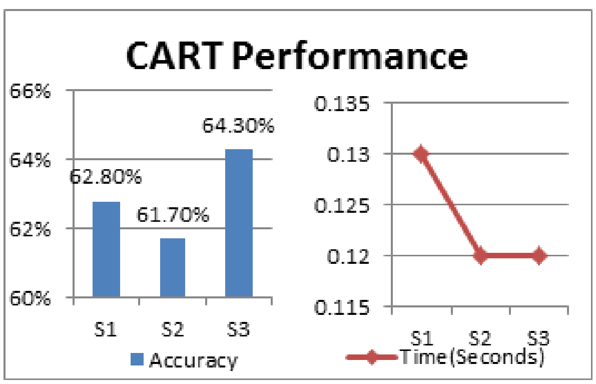


Figure 6.3: Experiment 2-CART Performance

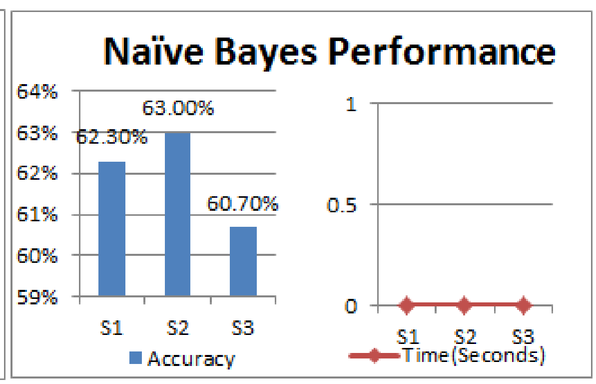


Figure 6.4: Experiment 2-Naïve Bayes Performance

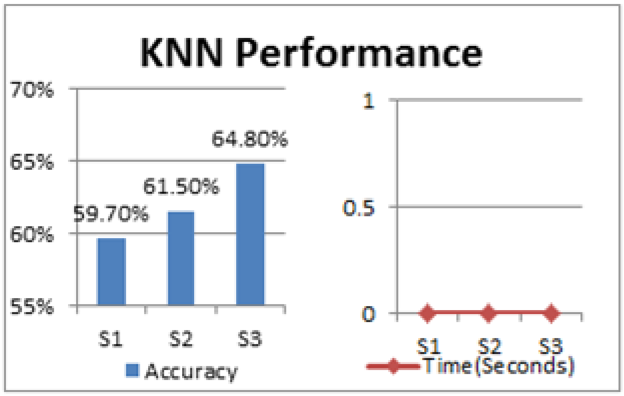


Figure 6.5: Experiment 2-KNN Performance

To summarize, the figure below shows that: using Campus, Gender, Term 1, Term 2 and Term 3 features , the best prediction performance for ID3,C4.5,CART and KNN, was achieved when each classifier was applied using S3 sample. The best accuracy obtained from mentioned classifiers varies from 64.30% to 64.80%.Naïve Bayes has its best accuracy of 63.00%, when it was applied using S2 sample.

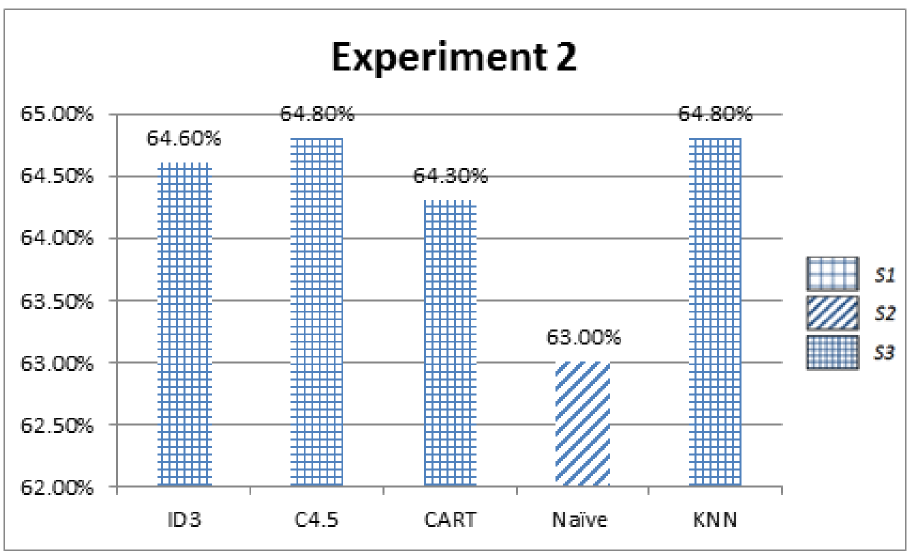


Figure 6.6: The Best Results of all Classifiers in Experiment 2

4.3.3. Experiment 3:

The second attribute subset (subset 2) which is generated by wrapper method, and consists of Campus, Gender, Cluster, Term 1 and Term 3 features, was used to run the third experiment in this paper. The experiment results are presented in table 5:

Sample Size	Algorithm	Correctly identified Instances	Incorrectly identified Instances	Unclassified Instances	Accuracy	Time Seconds
S1	ID3	59.00%	37.20%	3.79%	61.40%	0
	C4.5	60.43%	39.57%	0.00%	60.40%	0
	CART	60.66%	39.34%	0.00%	60.80%	0.16
	Naïve Bayes	55.45%	44.55%	0.00%	54.40%	0
	KNN	59.95%	40.05%	0.00%	60.10%	0
S2	ID3	58.52%	36.91%	4.57%	61.50%	0
	C4.5	63.25%	36.75%	0.00%	63.20%	0.02
	CART	62.15%	37.85%	0.00%	62.10%	0.14
	Naïve Bayes	55.52%	44.48%	0.00%	55.10%	0
	KNN	62.30%	37.70%	0.00%	62.50%	0
S3	ID3	61.08%	36.74%	2.18%	62.50%	0.02
	C4.5	65.44%	34.56%	0.00%	65.50%	0.02
	CART	65.63%	34.38%	0.00%	65.60%	0.14
	Naïve Bayes	60.42%	39.58%	0.00%	60.00%	0
	KNN	62.78%	37.22%	0.00%	62.90%	0

Table 5: Experiment 3 Results

Table 5 shows that, although Naïve Bayes took 0 second to build the model, but the prediction accuracy of this model is the lowest (54.40%), compared to the other methods. However, CART took the longest execution time among other classifiers (0.14 second), but the prediction accuracy obtained from CART is the highest (65.60%) among all classifiers.

Figure 6.1 presents CART classifier showing the poorest performance in this experiment and is generated by using S1 as sample, where accuracy equals to 60.80% and time required is 0.16 second. S2 and S3 samples will require slightly less amount of time, but accuracy varies.

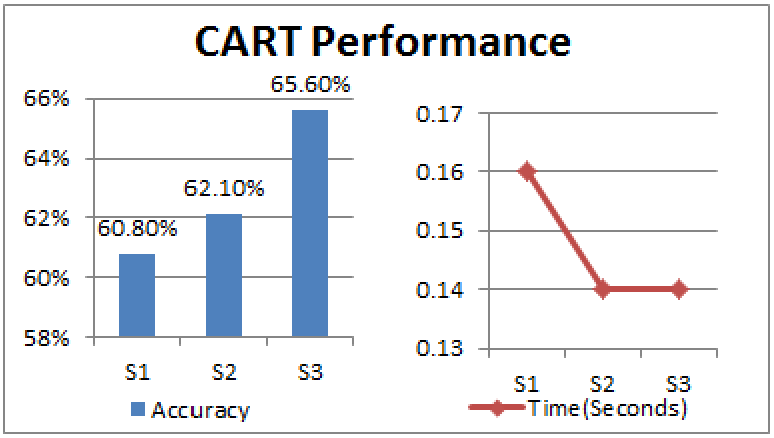


Figure 7.1: Experiment 3-CART Performance

C4.5 has almost the same accuracy as CART when using S3, but amount of time required by C4.5 (.0.02 second) is way less than the time required by CART.

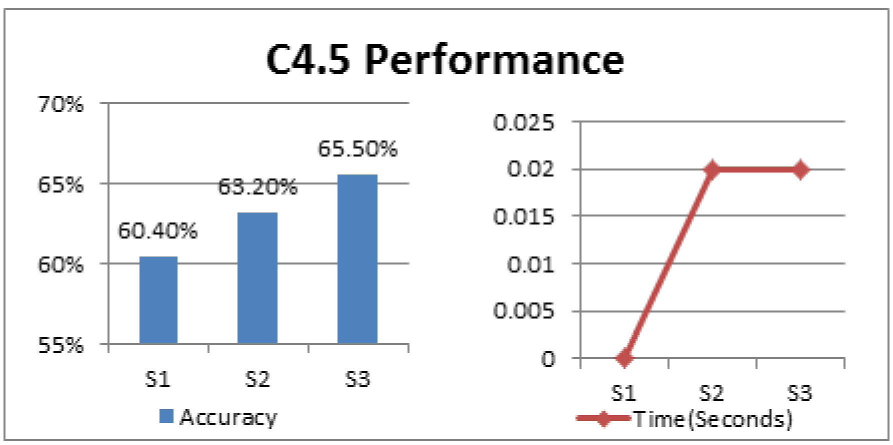


Figure 7.2: Experiment 3-C4.5 Performance

Best prediction accuracy is for ID3 and KNN, which are close to each other; ID3 has 62.50% and KNN has 62.90% but time required by ID3 is a bit longer than KNN. Figures below show the results from ID3 and KNN:

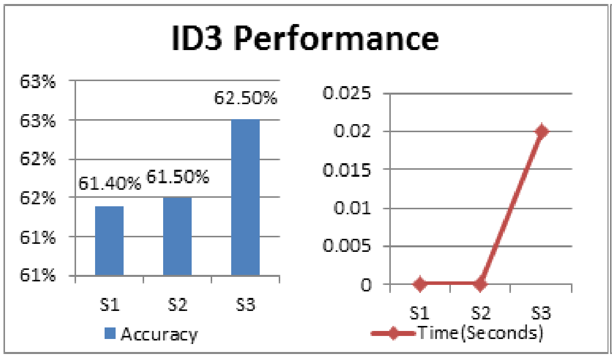


Figure 7.3: Experiment 3-ID3 Performance

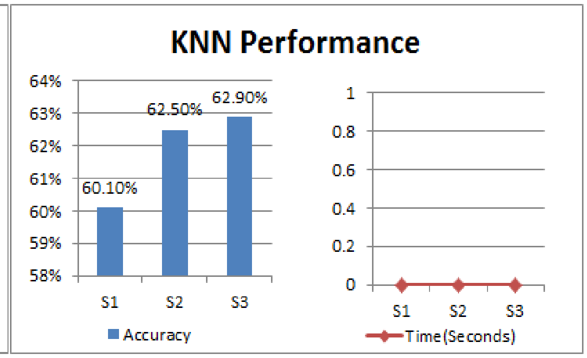


Figure 7.4: Experiment 3-KNN Performance

In this experiment, Naïve Bayes in this experiment has the lowest results among other classifiers. Although execution time required by three samples is 0 second, but the highest prediction accuracy recorded for Naïve Bayes is 60% using S3, then 55.10% using S2 and finally the lowest is 54.40% using S1.

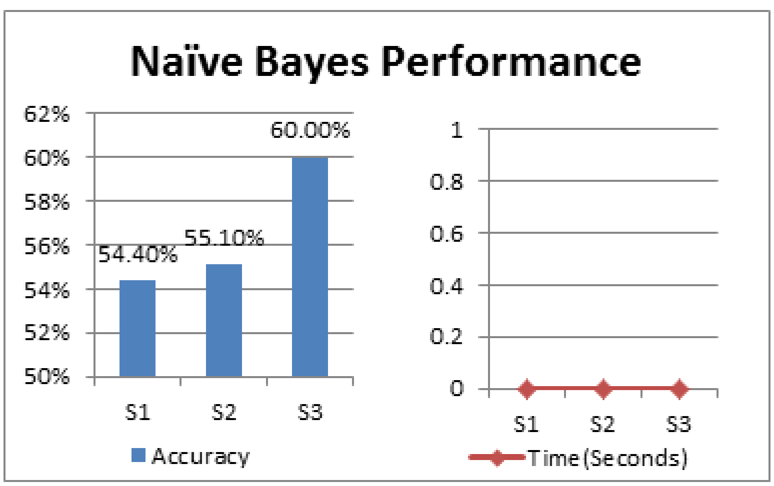


Figure 7.5: Experiment 3-Naive Bayes Performance

To summarize; using Campus, Gender, Cluster, Term 1 and Term 3 features (subset 2) , the best prediction performance for all classifiers was achieved, when each classifier was applied using S3 sample. The best accuracy obtained from all classifiers varies from 60.00%-65.60%. Figure 6.6 shows the summary of this experiment:

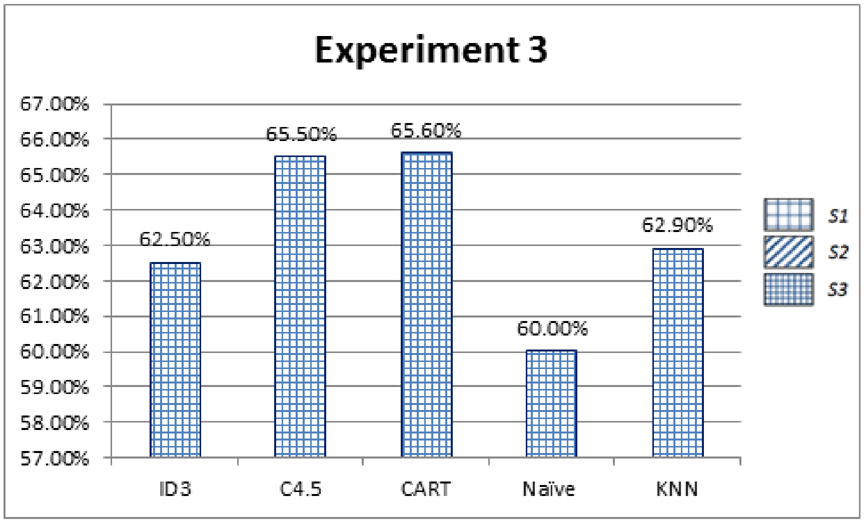


Figure 7.6 The Best Results of all Classifiers in Experiment 3

4.3.4. Experiment 4:

In this experiment 6 attributes; Campus, Gender, GPA, Term 1, Term 2 and Term 3 were selected by wrapper feature selection method as a relevant subset (subset 3) that affect prediction performance of the models. Table below shows results generated in experiment4, where the five classifiers were run using subset 3 and three samples (S1, S2 and S3):

Sample Size	Algorithm	Correctly identified Instances	Incorrectly identified Instances	Unclassified Instances	Accuracy	Time
S1	ID3	53.08%	39.57%	7.35%	57.30%	0
	C4.5	61.37%	38.63%	0.00%	61.30%	0.02
	CART	59.48%	40.52%	0.00%	59.50%	0.17
	Naïve Bayes	62.80%	37.20%	0.00%	62.60%	0
	KNN	62.09%	37.91%	0.00%	62.10%	0
S2	ID3	53.63%	36.28%	10.09%	59.60%	0.02
	C4.5	64.20%	35.80%	0.00%	64.20%	0
	CART	61.04%	38.96%	0.00%	61.20%	0.17
	Naïve Bayes	61.51%	38.49%	0.00%	61.40%	0

	KNN	62.46%	37.54%	0.00%	62.50%	0
S3	ID3	58.33%	36.55%	5.11%	61.70%	0.01
	C4.5	64.77%	35.23%	0.00%	64.90%	0
	CART	63.16%	36.84%	0.00%	63.30%	0.17
	Naïve Bayes	60.70%	39.30%	0.00%	60.50%	0
	KNN	61.84%	38.16%	0.00%	62.00%	0

Table 6: Experiment 4 Results

Based on the figures mentioned in the table above, accuracy generated by classifiers using subset 3 varies from 57.30% to 64.90%. The highest accuracy is obtained by C4.5. Where C4.5 correctly classified 64.77% of instances, and the model took 0 second to be built. ID3 has the lowest accuracy among all classifiers in this experiment. 57.30% is the lowest accuracy recorded by ID3, where 39.57% of instances were classified incorrectly and 7.35% of instances were not classified.

The accuracy generated by decision tree algorithms in this experiment was better when all classifiers used S3. C4.5 has the highest accuracy of 64.90%, ID3 and CART took more time than C4.5 to build the same models using S3.

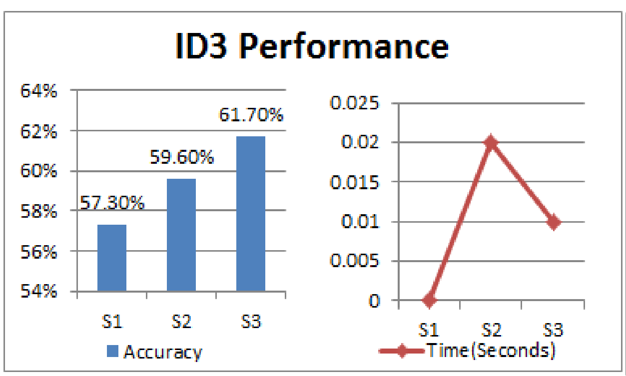


Figure 8.1: Experiment 4- ID3 Performance

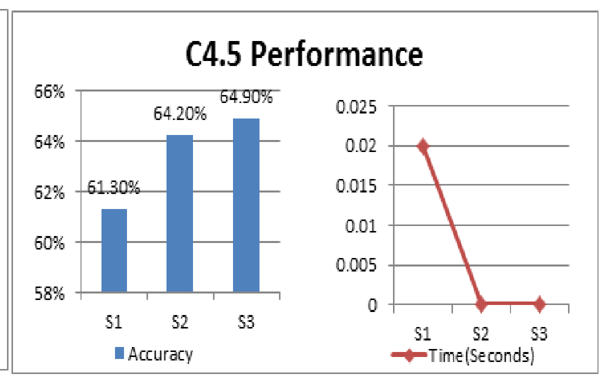


Figure 8.2: Experiment 4-C4.5 Performance

Regardless the size of the sample that is used for training and testing the model, 0.17 second was required to build CART model and predict the students' performance.

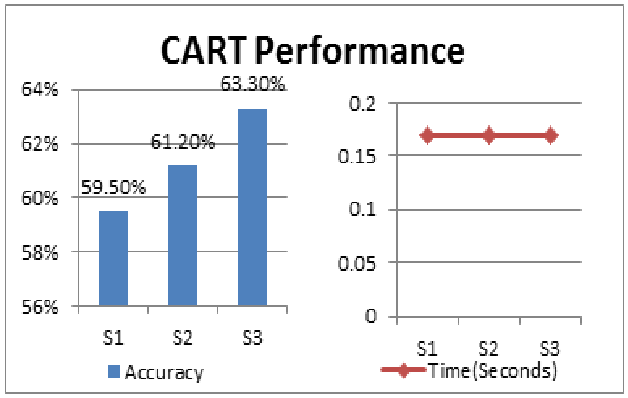


Figure 8.3: Experiment 4- CART Performance

Naïve Bayes and KNN in this experiment performed differently than decision trees. Time required by both models was 0 second for all samples. Naïve Bayes performed better when it was applied on S1, where the prediction accuracy is 62.60%. KNN performed better when it was applied using S2. The prediction accuracy for KNN using S2 is 62.50%.figures below show the experiment results for KNN and Naïve Bayes.

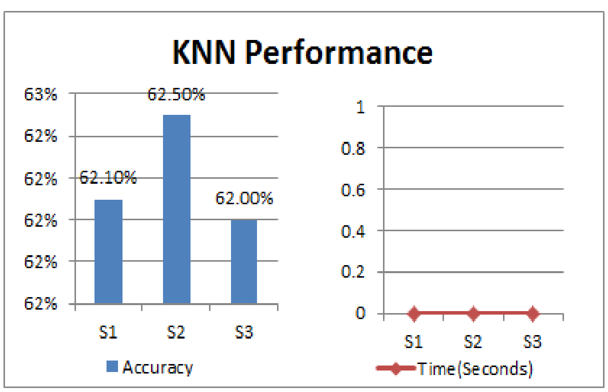


Figure 8.4: Experiment 4- KNN Performance

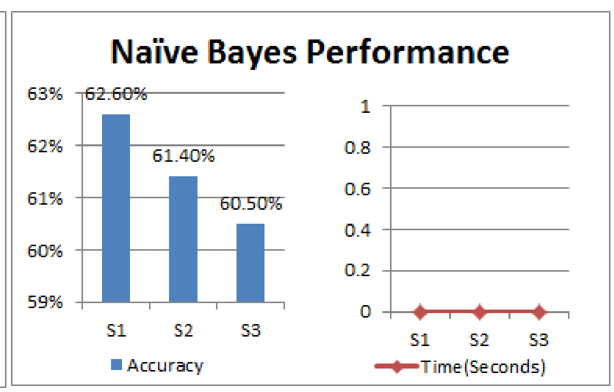


Figure 8.5: Experiment 4- Naive Bayes Performance

To summarize; using the features Campus, Gender, GPA, Term 1, Term 2 and Term 3 features (subset 3) , on one hand, the best prediction performance for Decision trees (ID3, C4.5 and CART) was achieved, when each classifier was applied using S3 sample. The best accuracy for decision trees varies from 61.70% to 64.90%.

On the other hand, Naïve Bayes and KNN performed best, when using consecutively S1 and S2 samples. The figure below shows results obtained from Experiment 4.

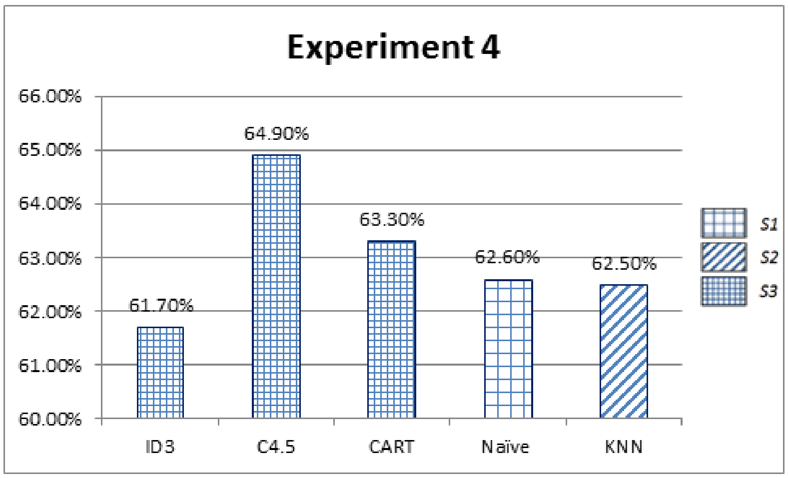


Figure 8.6: The Best Results of all Classifiers in Experiment 4

4.4. Comparison of Different Modeling Techniques

This section will compare the best accuracy achieved by each classifier on each experiment, as well as the time required to build the model:

4.4.1. ID3:

As shown in the figure below, the highest accuracy gained from ID3 algorithm is 64.60% where the model took 0 seconds to run. And thus, the student practical skills in term1, term 2 and term 3 can have a significant impact on the student performance in the final test. Not only the student technical skills, but also student's gender and location can affect the student performance. Furthermore, the major that the student is studying, and the student's achievement in the school or prior knowledge has less impact on how good or poor the students will do in the final test. The highest accuracy was gained from ID3 when it was applied using cross-validation (10 folds) sampling technique and subset1 as predictors.

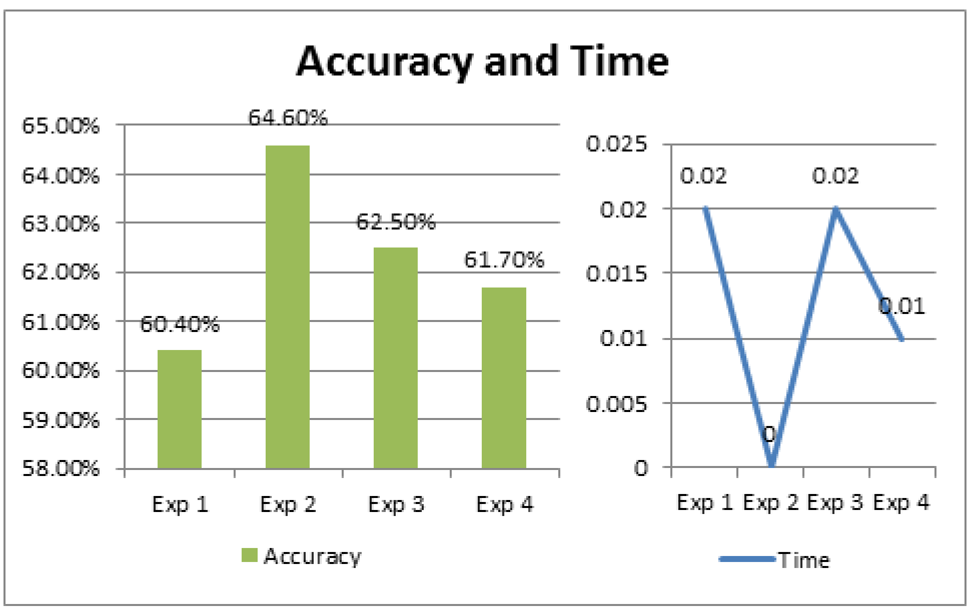


Figure 9: ID3 Best Performance in all Experiments

4.4.2. C4.5:

The Results in Figure 9 show that, the technical skills that the students learned in term 2, and a student prior knowledge has a less effects on his/her performance in the theoretical exam and don't add much to their academic skills. However, technical skills taught in term 1 and term 3 and the student major, location and gender are tightly linked to the student academic achievement. The results show that C4.5 classifier has the highest prediction accuracy of 65.50%, when term 2 marks and GPA were both excluded from the experiment. Time required to build this model is 0.02 second and by using cross-validation (10 folds) as a sampling technique.

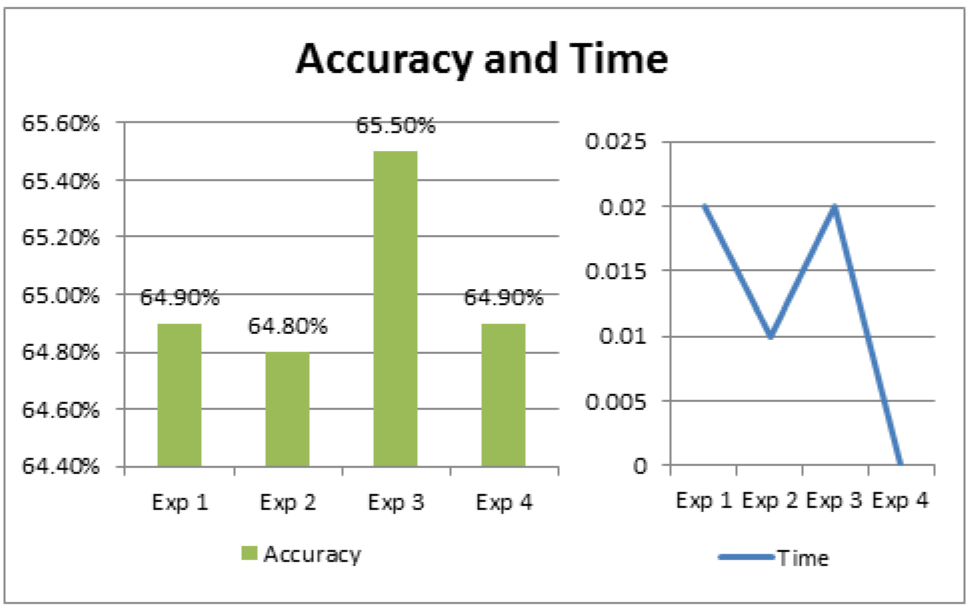


Figure 10: C4.5 Best Performance in all Experiments

4.4.3. CART

Similarly to C4.5, CART classifier best performed when it was executed using: Campus, Gender, Cluster, Term 1 and Term 3 attributes as predictors, and cross-validation(10 folds) as a sampling technique to train and test the model . The model took 0.14 second to record the highest accuracy of 65.60% using CART classifier.

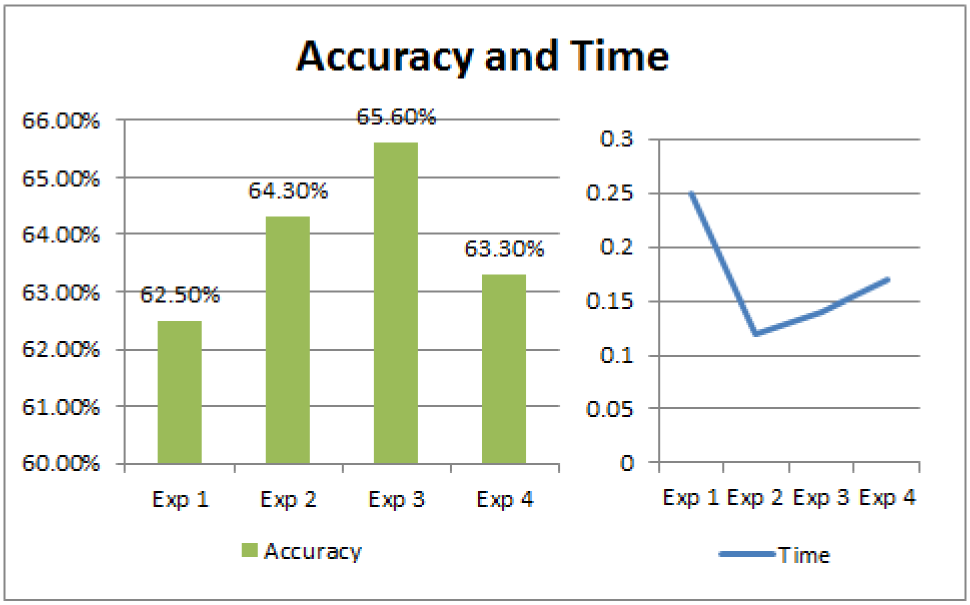


Figure 11: CART Best Performance in all Experiments

4.4.4. Naïve Bayes

The figure below represents the best performance of Naïve Bayes for the four experiments. The results show that time required to execute Naïve Bayes in different experiments is 0 second. The school location, gender and technical skills that a student learns during the academic year have a great impact on predicting the student final mark. GPA and the student major can slightly help in predicting the student performance. The highest accuracy of 63.00% is achieved when Campus, Gender, Term 1, Term 2 and Term 3 attributes are used as predictors, and when 40% of dataset is used to train the model.

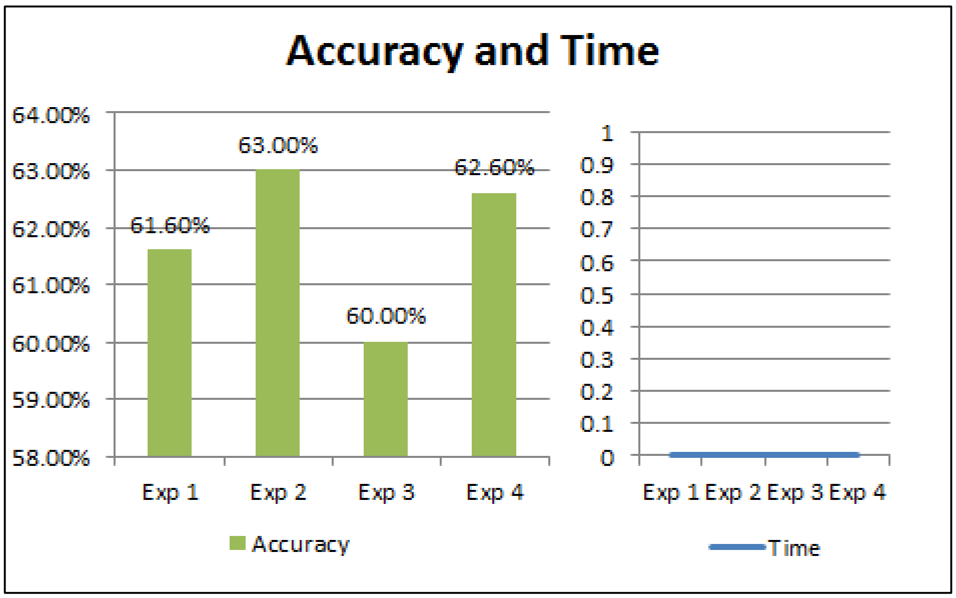


Figure 12: Naive Bayes Best Performance in all Experiments

4.4.5. KNN

The last classifier was used in this paper is KNN. KNN best performed when it was applied using student information related to his/ her technical skills during the year, his/her gender and location. The sampling method used to gain the highest prediction accuracy from KNN is cross-validation (10 folds). The figure below shows KNN best performance in each experiment:

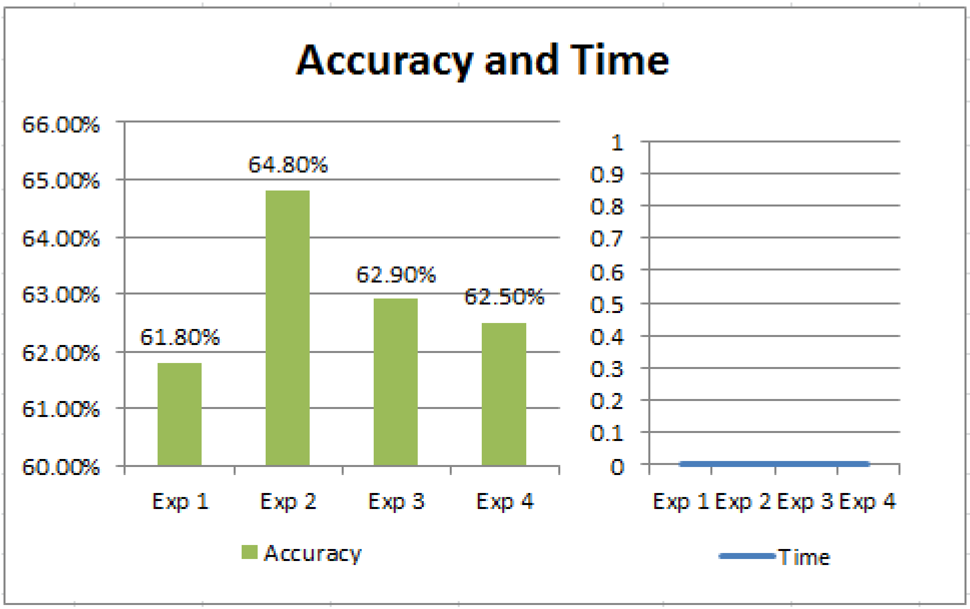


Figure 13: KNN Best Performance in all Experiments

4.5 Comparison of Best Accuracies of Modeling Techniques

This section will compare the best accuracy is achieved by each classifier, time required to build the model, combination of attributes as predictors and sampling technique used to validate each model.

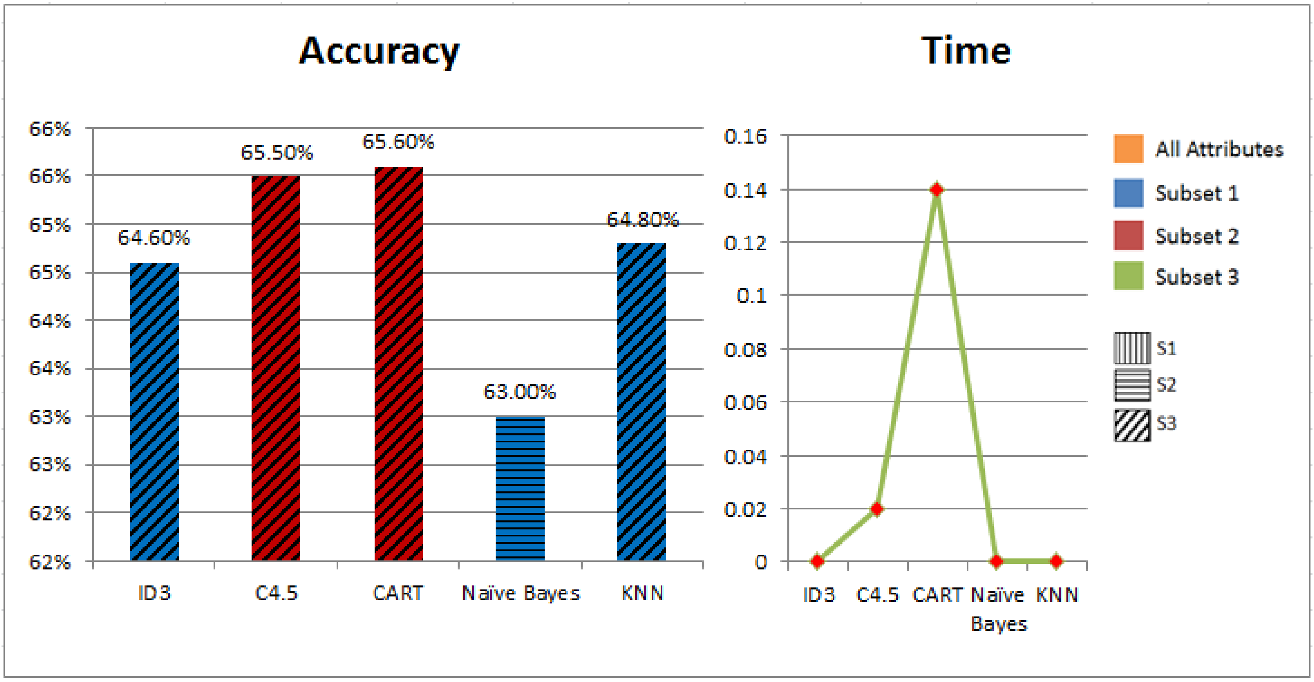


Figure 14: Best performance Comparison of Modeling Techniques

As shown in the figure above, the best accuracy is achieved by classifies ranges from 63.00% to 65.60%. Where CART has the highest among other classifiers when it was applied using S3 (cross-validation) sampling technique and Subset 2 attributes (Campus, Gender, Cluster, Term 1 and Term 3) as predictors. As well as, the longest amount of time required to build predictive model was also for CART classifier, it took 0.14 seconds to be built.

C4.5 and CART almost had the same accuracy in this study, but C4.5 took less time to build the model using the same sampling technique and predictors.

ID3 and KNN performed almost similarly to each other; campus, gender, term1, term 2 and term 3 are used to predict a student performance in the theoretical test (final assessment).

Furthermore, the same sampling technique (S3) is also used by both classifiers to generate their best prediction accuracy.

Differently, the poorest performance in this study was for Naïve Bayes. Like ID3 and KNN, Naïve Bayes used the same subset of attributes, but dissimilarly S2 sampling technique is used by Naïve Bayes to achieve the best prediction accuracy in the whole study.

5. Discussion

5.1 Results Discussion

This paper discussed how data mining techniques could be used to predict the students' performance in the end of year exam for computer science course. The following points can be deduced from the results mentioned in the section above:

1. Regardless of modeling techniques that are used to predict the students' performance, Campus, gender, Term 1 and Term 3 are always selected as relevant attributes and behaved as predictors for all models. Computer science teachers at Abu Dhabi Schools (male and female campus) explained that 40% of the materials taught in term-1 are covered in EOY exam as well as EOY exam is covering almost 50 % of term 3 materials.

2. The highest accuracy in this study was obtained when Term2 marks were excluded from the attribute subset. Although 35% of term 2 marks is dedicated for calculating the amount of academic knowledge that the student gained during the term using continual assessments (unified written assessment for all campuses), but it could be seen that this mark has no relationship with the students' performance in the EOY exam. After interviewing some computer science teachers at Abu Dhabi schools, the teachers explained that as a part of VET school's curriculum, the students must be trained in the technical, vocational and technological fields. To fulfill this requirement, the students should be sent for at least 4 weeks to work in a company that is related to the student's major in order to empower them with the competencies needed for employability and lifelong learning. And hence,

during term 2, all students are released from their schools to work placements for almost 6 weeks. And because of that, curriculum development and assessment units decided to include only 10 % of term 2 materials in the EOY exam. So topics taught in term 2 are only partially covered in EOY exam for computer science module.

3. Gender of the student also has a significant impact on the student performance in the end of year exam. The results from this study show that female students tend to perform much better than male students. Table below shows how EOY score is affected by the gender:

EOY Mark	Female	Male
A	11%	3%
B	40%	23%
C	37%	36%
D	12%	39%

Table 7: EOY Mark distribution Across Gender

The table above shows that the highest majority of Female students got grade B in the EOY exam, meanwhile the highest majority of male students got grade D.

4. The student location (campus) also has a great impact on the student performance in the EOY exam. More than 60% of the students in Fujairah and Ras-Al-Khaimah schools got grade D, and at the same time none of the students in the same schools got grade A in the EOY Exam. The highest majority of students in Al Ain, Ajman and Dubai schools got grade C. In Abu Dhabi school, the highest majority of the students got grade B.

Campus	A	B	C	D
Al Ain	4%	31%	38%	27%
Ajman	11%	30%	41%	18%
AUH	8%	48%	34%	10%
DXB	1%	25%	46%	27%
FUJ	0%	11%	28%	62%
RAK	0%	3%	18%	79%

Table 8: Students performance Across Different Campuses

The principals in schools with lower grades should pay more attention and make proper intervention programs that can help students to perform better as well as, provide variety of instructional methods that are suitable for both genders.

5. The predictive models that include the student prior knowledge (GPA) have less influence on prediction accuracy for student performance than other models. And thus, GPA does not have impact on student performance in the computer science EOY exam. This is because GPA is a more general prior achievement than the student achievement in a particular course. It also reflects that the state needs national youth trained in the technical, vocational and technological fields.

5.2 Answers to the research questions:

- First Research question: How will traditional data mining techniques perform, when applied on students' information in the vocational schools?

  ID3, C4.5 CART, KNN and Naïve Bayes techniques are used in this study to build various predictive models. On one hand, time needed to build KNN and Naïve Bayes models remains unchanged in spite of change in relevant parameters (i.e. features subsets and training and testing sample sizes). While time required by other models varies from 0.00 till 0.92 seconds.

  On the second hand, the prediction accuracies obtained by the predictive models in four experiments vary as the following:
  1. ID3: 60.4%-64.6%
  2. C4.5: 64.8%-65.5%
  3. CART: 62.5%-65.6% (the best accuracy obtained in this study).
  4. Naïve Bayes:60%-63%
  5. KNN: 61.8%-64.8%.

- Second Research question: What students information (variables) can have a stronger effect and lead to the best prediction accuracy?

According to wrapper feature selection method, three different features subsets are selected as predictors for the predictive models. Features subsets are:

1. Subset 1: Campus, Gender, Term 1, Term 2 and Term 3.
2. Subset 2: Campus, Gender, Cluster, Term 1 and Term 3
3. Subset 3: Campus, Gender, GPA, Term 1, Term 2 and Term 3.

The results in this study indicate that the best combination of features that led to the highest prediction accuracy is subset 2.

6. Conclusion

As a summary to this research, student low academic performance has been a long standing problem in Vocational Educational and Training schools. Educational Data Mining can provide educators and students with different tools to predict how poorly or how well the student will perform in a particular course. Moreover, curriculum specialist can also benefit from EDM in identifying strengths and weaknesses in materials used for the course, and based on that, proper interventions can be made to these courses. In this study, various predictive models have been developed and validated for the purpose of trimming down the student failure rate in computer science course, by making early prediction of the students' marks in final exam. Three combinations of variables have been selected by wrapper feature selection method, to be used as predictors. Decision trees, Naïve Bayes and KNN have been used as predictive techniques to predict the students' performance in the final exam. And finally three different training and testing sample sizes have been used to validate and evaluate the predictive models. The Prediction accuracy and execution time are the criterion's used to evaluate and compare the performance of the different predictive models.

ID3, C4.5 CART, KNN and Naïve Bayes techniques are used in this study to build various predictive models. Time needed to build KNN and Naïve Bayes models remains unchanged (0.00 seconds) in spite of change in relevant parameters. While time is required by ID3, CART and C4.5 varies from 0.00 till 0.92 seconds. The highest accuracy is obtained by CART (65.6%), then C4.5 (65.5%) when using Campus, Gender, Cluster, Term 1 and Term 3 attributes as predictors and Cross-Validation (10 Folds) data splitting method.

7. REFERENCES

- Acharya, A., & Sinha, D. (2014). Application of Feature Selection Methods in Educational Data Mining. *International Journal of Computer Applications*,*103*(2).

- Ahmad, A., & Dey, L. (2007). A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set. *Pattern Recognition Letters*, *28*(1), 110-118.

- Amatriain, X., Jaimes, A., Oliver, N., & Pujol, J. M. (2011). Data mining methods for recommender systems. In *Recommender Systems Handbook* (pp. 39-71). Springer US.

- Anuradha, C., & Velmurugan, T. (2015). A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Students Performance. *Indian Journal of Science and Technology*, *8*(15).

- Austin, P. C., Tu, J. V., Ho, J. E., Levy, D., & Lee, D. S. (2013). Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. *Journal of clinical epidemiology*, *66*(4), 398-407.

- Awada, W., Khoshgoftaar, T. M., Dittman, D., Wald, R., & Napolitano, A. (2012, August). A review of the stability of feature selection techniques for bioinformatics data. In *Information Reuse and Integration (IRI), 2012 IEEE 13th International Conference on* (pp. 356-363). IEEE.
- Baradwaj, B. K., & Pal, S. (2012). Mining educational data to analyze students' performance. *arXiv preprint arXiv:1201.3417*.

- Bekkink, M. O., Donders, R., van Muijen, G. N., de Waal, R. M., & Ruiter, D. J. (2012). Explicit feedback to enhance the effect of an interim assessment: a cross-over study on learning effect and gender difference. *Perspectives on medical education*,*1*(4), 180-191.

- Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, *18*(1), 5-25.

- Bermejo, P., de la Ossa, L., Gámez, J. A., & Puerta, J. M. (2012). Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking. *Knowledge-Based Systems*, *25*(1), 35-44.

- Bhise, R. B., Thorat, S. S., & Supekar, A. K. (2013). Importance of data mining in higher education system. *IOSR Journal Of Humanities And Social Science (IOSR-JHSS) ISSN*, 2279-0837.

- Cadenas, J. M., Garrido, M. C., & MartíNez, R. (2013). Feature subset selection filter–wrapper based on low quality data. *Expert Systems with Applications*, *40*(16), 6241-6252.

- D'Haen, J., Van den Poel, D., & Thorleuchter, D. (2013). Predicting customer profitability during acquisition: Finding the optimal combination of data source and data mining technique. *Expert systems with applications*, *40*(6), 2007-2012.

- Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., ... & Lautenbach, S. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 27-46.

- Efron, B. (2013). Bayes' theorem in the 21st century. *Science*, *340*(6137), 1177-1178.

- El Gamal, A. F. (2013). An Educational Data Mining Model for Predicting Student Performance in Programming Course. *International Journal of Computer Applications*, *70*(17), 22-28.

- Eynon, R. (2013). The rise of Big Data: what does it mean for education, technology, and media research?. *Learning, Media and Technology*, *38*(3), 237-240.

- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, *11*(1), 10-18.

- Herppich, S., Wittwer, J., Nückles, M., & Renkl, A. (2014). Addressing knowledge deficits in tutoring and the role of teaching experience: Benefits for learning and summative assessment. *Journal of Educational Psychology*,*106*(4), 934.

- Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. *Cybernetics and Information Technologies*, *13*(1), 61-72.

- Kumar, U., & Pal, P. S. (2011). Data Mining: A prediction of performer or underperformer using classification.
- Kwok, S. W., & Carter, C. (2013). Multiple decision trees. *arXiv preprint arXiv:1304.2363*.

- Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., ... & Nowe, A. (2012). A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, *9*(4), 1106-1119.

- Luan, J. (2002). Data mining and its applications in higher education. *New directions for institutional research*, *2002*(113), 17-36.

- Moons, K. G., Kengne, A. P., Grobbee, D. E., Royston, P., Vergouwe, Y., Altman, D. G., & Woodward, M. (2012). Risk prediction models: II. External validation, model updating, and impact assessment. *Heart*, heartjnl-2011.

- Namdeo, J., & Jayakumar, N. (2014). Predicting Students Performance Using Data Mining Technique with Rough Set Theory Concepts. *International Journal*,*2*(2).

- Naqvi, R. (2015). Data Mining in Educational Settings. *Pakistan Journal of Engineering, Technology & Science*, *4*(2).

- Olson, D. L., Delen, D., & Meng, Y. (2012). Comparative analysis of data mining methods for bankruptcy prediction. *Decision Support Systems*, *52*(2), 464-473.

- Pandey, U. K., & Pal, S. (2011). Data Mining: A prediction of performer or underperformer using classification. *arXiv preprint arXiv:1104.4163*.

- Patidar, P., Dangra, J., & Rawar, M. K. (2015). Decision Tree C4. 5 algorithm and its enhanced approach for Educational Data Mining.

- Patidar, P., Dangra, J., & Rawar, M. K. (2015). Decision Tree C4. 5 algorithm and its enhanced approach for Educational Data Mining.

- Pinto, S. K., Mansfield, R., Jacobs, M., & Rubin, D. (2012). *U.S. Patent No. 8,170,841*. Washington, DC: U.S. Patent and Trademark Office.

- Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.

- Quinlan, J. R. (1979). Discovering rules form large collections of examples: A case study. *Expert Systems in the Micro Electronic Age. Edinburgh Press: Edingburgh*.
- Rai, A., & Kumar, D. (2014). A Comparative Study on Wrapper and Filter Methods for Feature Selection in Data Mining. *International Journal Of Scientific Research And Education*, *2*(08).

- Ramaswami, M., & Bhaskaran, R. (2009). A study on feature selection techniques in educational data mining. *arXiv preprint arXiv:0912.3924*.

- Read, E. K., & Hecker, K. G. (2013). The development and delivery of a systematic veterinary clinical skills education program at the University of Calgary. *J Veterinar Sci Technol S*, *4*, 2.

- Rodrigues, D., Pereira, L. A., Nakamura, R. Y., Costa, K. A., Yang, X. S., Souza, A. N., & Papa, J. P. (2014). A wrapper approach for feature selection based on Bat Algorithm and Optimum-Path Forest. *Expert Systems with Applications*, *41*(5), 2250-2258.

- Romero, C., Espejo, P. G., Zafra, A., Romero, J. R., & Ventura, S. (2013). Web usage mining for predicting final marks of students that use Moodle courses. *Computer Applications in Engineering Education*, *21*(1), 135-146.

- Shi, L., Cristea, A., Awan, M. S., Stewart, C., & Hendrix, M. (2013). Towards understanding learning behavior patterns in social adaptive personalized e-learning systems.

- Siemens, G., & Baker, R. S. (2012, April). Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 252-254). ACM.

- Souto-Otero, M., & Ure, O. B. (2012). The coherence of vocational education and training in Norway and Spain: national traditions and the reshaping of VET governance in hybrid VET systems. *Compare: a journal of comparative and international education*, *42*(1), 91-111.

- Steyerberg, E. W., Harrell, F. E., Borsboom, G. J., Eijkemans, M. J. C., Vergouwe, Y., & Habbema, J. D. F. (2001). Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of clinical epidemiology*, *54*(8), 774-781.

- Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., ... & Kattan, M. W. (2010). Assessing the performance of prediction models: a framework for some traditional and novel measures.*Epidemiology (Cambridge, Mass.)*, *21*(1), 128.
- Tair, M. M. A., & El-Halees, A. M. (2012). Mining educational data to improve students' performance: a case study. *International Journal of Information*, *2*(2).
- Verbraken, T., Bravo, C., Weber, R., & Baesens, B. (2014). Development and application of consumer credit scoring models using profit-based classification measures. *European Journal of Operational Research*, *238*(2), 505-513.
- Walters-Williams, J., & Li, Y. (2010). Comparative study of distance functions for nearest neighbors. In *Advanced Techniques in Computing Sciences and Software Engineering* (pp. 79-84). Springer Netherlands.
- White, S. (2014). *Transitioning from vocational education and training to university: strengthening information literacy through collaboration*. National Centre for Vocational Education Research (NCVER).
- Xue, B., Zhang, M., & Browne, W. N. (2013). Particle swarm optimization for feature selection in classification: A multi-objective approach. *Cybernetics, IEEE Transactions on*, *43*(6), 1656-1671.
- Yadav, S. K., & Pal, S. (2012). Data mining: A prediction for performance improvement of engineering students using classification. *arXiv preprint arXiv:1203.3832*.
- Yadav, S. K., Bharadwaj, B., & Pal, S. (2012). Data mining applications: A comparative study for predicting student's performance. *arXiv preprint arXiv:1202.4815*.