Using Data Mining Classification Techniques to predict Students’ Performance of the Higher Institute of Medical Sciences El-Shati

Abdulmonem A. A. Ahmed¹, Aybaba Hançerlioğulları², Yasemin GÜLTEPE³
Department of Material Science and Engineering, Graduate School of Natural and Applied Kastamonu University, Turkey¹
Department of Physics, Faculty of Science and Letters, Kastamonu University, Turkey²
Engineering and Architecture Faculty, Head of Computer Engineering Department, Kastamonu University, Turkey³

Abstract:
The growth of data size in the institutes of various associations, organizations, companies, universities and higher education is very rapidly increasing. This data includes hidden knowledge. The process of data mining is the tools and techniques that can obtain such knowledge that can't be seen using conventional methods. In this paper we have applied the data mining techniques of student educational data of the Higher Institute of Medical Sciences El-Shati Libya. The data mining techniques are applied to educational data in order to assess and enhance academic success of the students. To carry out these activities we implemented a classification methodology. The data set used in this research is data from the students from fall 2007-2008 to spring 2018. The study focused on four classification algorithms (J48 pruned tree, IB1 instance-based, BayesNet and attribute Selection), we had implemented these four classifiers algorithms on the institute dataset, by comparing the three evaluate parameters (accuracy, Recall and Precision) using four algorithms; we had got the highest accuracy reached to 100 %, which means a strong relationship between features that affects the performance of students.

Keywords: WEKA; Data Mining; Classification; Education; Academic Performance; Student’s Performance.

I. INTRODUCTION

The rapid and great growth of computer and information technology has encouraged and streamlined the process of accessing public information in all fields. What we need to do is just write down some of the words in the search engine and we get huge sites to search for information about the topic. Using applications in different organizations, on the other hand leads to get huge amounts of data, which are generally stored in databases. In general, the main purpose of the systems is to save different data and calculate different processes depending on the nature of the system, these systems can perform specific functions and generally store data in the database. Data mining is used with certain patterns to derive the relevant information from massive data. It is a science that analyzes data to show in this data hidden and unknown information and patterns, serving as a tool for analyzing large amounts of data (Shammuga raijeshwari and Lawrance, 2016; Predic et al., 2018). Information mining is one of the most powerful tools for data processing and information management (Ahmed, Rizaner, and Ulusoy, 2016). The methods and techniques used in data mining include generalization, characterization, classification, clustering, association, evolution, pattern matching, and visualization of Knowledge Discovery in Databases (KDD) and Data Mining (DM) offers a platform for generalizing structures, processes, techniques, algorithms, and even automated tools that guide and promote smart, in-depth and complex research (Jothi and Husain, 2015). Educational data mining (EDM) concerned with developing methods for exploring and enhancing academic performance and establishing trends in the data of these institutions on students and graduates, all relevant to the most critical performance measures such as student performance and the quality of faculty results, then using such methods to get understanding the results and performance in better manner (M. Wook et al., 2009; Roy and Garg, 2018; Amra and Maghari, 2017; Christian and Ayub, 2014; Razaque et al., 2017). Early prediction of success of students will help enhance the educational process. Prediction of student success is very significant and can be achieved by using data mining techniques on educational data sets (Amra and Maghari, 2017; Yang and Li, 2018). These institutions may then develop policies to help students improve academic performance. This paper presents an applied data mining analysis to help decision-makers at the Higher Institute of Medical Sciences - Shati Libya by using data mining approach to detect knowledge and information from the database of the Institute. Higher Institute of Medical Sciences-Shati’s decision-makers have a database of students, courses, results and others since the institute was founded in 2000. These data have not yet been effectively exploited in the knowledge of success factors and failure factors nor have knowledge of the dominant behavior patterns of Students. Data mining techniques can be used to support decision-makers with results, showing the links between the different patterns within the institution’s structure, and showing the decision-maker’s weaknesses and strengths to do what’s useful in assessing performance within the institution. In this paper we will using data mining classification techniques on dataset selected of the institute, therefore in this section we will give an overview of data mining and classification techniques.

II. RELATED WORKS

There are many studies in predicting academic failure of students through the use of data extraction techniques. However,
we are trying to explore some of the most important patterns of student achievement at the Higher Institute of Medical Sciences - Shati, and try to find out relationships between these patterns, submit them to decision makers at the Institute to take advantage of them. In this section we will present some previous work in this area from different regions of the world. Although these works differ in some respects, they are similar in other respects. Here we will highlight some of these actions. (Alqahtani and John 2017) presented two experiments to predict the final mark for the student. The first experiment compared classification algorithms using three datasets, in which results were obtained with higher precision, taking into account all available data versus filtering. They concluded in the second experiment that the highest precision can be attained by applying the classification model to numerical as well as categorical data. (Marquez-Vera et al., 2013) evaluate the use of Educational Data Mining (EDM) techniques, induction rules, and decision trees to predict academic failures of students in middle school. The results showed evidence that such techniques are a promising way to execute predictions of this kind with relevant accuracy. However, to obtain the results, the study had to analyze several different variables from various data sources, including non-academic data, such as family and personal data; finally, they don't determine whether a student fails early enough. (Ahmad, Ismail, & Aziz, 2015) introduced a predictive approach for first year bachelor students’ academic success in computer science course. It uses a framework which supports the classification techniques of Decision Tree, Naive Bayes, and Rule Based, to be applied to the data of the students to generate the academic performance prediction model of the best students. During 8 semesters the used data were collected, containing demographic details of the students, previous academic records, and background information of the family. The results show that by obtaining the highest accuracy value of 71.3%, the Rule Based classifier is a best choice among other techniques. (Watson and Li, 2014), proposed a programming course strategy, called “Watwin”, designed to predict student results. In general, when a student compiles his software on a university Computer, the method takes a snapshot of the source code of the software and it collects details about the program’s success or failure, the execution time, error messages and code line number. The “Watwin” approach uses that information to predict student failure. The results show that the approach can achieve efficiency of up to 75 %.(Shahiri and Husain, 2015), presented a description of the data mining techniques used to predict the achievement of students. They focus on how to use the prediction algorithm to identify the most important attributes in data for a student. They recommend making more efficient use of educational data mining techniques to boost the performance and progress of the students. The benefits and impacts they could bring to students, educators and academic institutions.

### III. METHODOLOGIES AND METHODS

Data mining techniques have many achievements in many areas, such as engineering, education, marketing, medical, financial, and sport. Data mining normally follows a general framework starting with Dataset preparation and if any data processing in the dataset is required then Preprocessing will be applied to the dataset after the data mining algorithms and techniques such as classification, clustering will be applied to the dataset, finally analyzing the results. In this paper although there are a variety of techniques used in data mining, we will apply the classification method that suits this work, there are also several algorithms in the classification process, the output data set is evaluated and analyzed using four Classification algorithms (J48 pruned tree, IB1 instance-based, BayesNet and attributeSelection). We used WEKA workbench for the implementation of all of these classification tasks. WEKA (Knowledge Research Climate, Waikato). It is a popular Java-written data mining tool developed at 'Waikato University, New Zealand.' It (Ahmad, Ismail, & Aziz, 2015; Kaur and Singh, 2015; Ahmed, Rizaner, and Ulusoy, 2016; Goga and Kuyoro, 2015; Ahmed, Rizaner, and Ulusoy, 2016) can be known as the WEKA workbench for the implementation of all of these classification techniques of Decision Tree, Naive Bayes, and Rule Based. The main objective of this work to discover the patterns in academic student data at the Higher Institute of Medical Sciences - Al-Shati, and to produce results that will help decide on the academic aspect of the Institute. The research was applied to the data of students at the institute, the data which will use is academic data without using of other data such as financial or personal data, the research will apply to students’ data from fall 2007-2008 to Spring 2018, finally, the classification is the only techniques of data mining will apply on the selected dataset. 

#### 3.1. Data Set

The dataset which was used in this work included students’ data with (34600) records, although the institute was founded in (2000), at the beginning of study in institute it was studying the system of the year, and in (2007) the system was changed to the semester system by six semesters, for this reason we begin from 2007. The database used in the system of the Institute is from type of Microsoft SQL Server and consists of a set of tables related together with primaries keys and forging keys. The data was converted to a Microsoft Access file, the reason is Microsoft Access can works independently of the database management system. For some reason we cannot deal directly with database with its original status, and it change daily. This is why a group of fields or attributes was selected from several different tables and a suitable data category was created for this work as dataset. The dataset was then exported to a text file type. Finally, the text file was modified to a file type (.arff) and this type is used with the (Weka). The main objective of this work to discover the patterns in academic student data at the Higher Institute of Medical Sciences - Al-Shati, and to produce results that will help decide on the academic aspect of the Institute. The research was applied to the data of students at the institute, the data which will use is academic data without using of other data such as financial or personal data, the research will apply to students’ data from fall 2007-2008 to Spring 2018, finally, the classification is the only techniques of data mining will apply on the selected dataset.

#### 3.2. Data Mining

Data mining is the method or process of extracting knowledge from huge amounts of data contained in databases, these knowledge can be patterns, correlations or any other information repositories (Rahman, 2011). It also can be known as the extracting hidden useful information from a dataset through scientific analysis and methods that identify data, the hidden patterns within the given dataset, therefore, data mining can be called knowledge discovery. Hidden information is discovered by running data mining algorithms combining statistics with computer science to mine valuable information from a seemingly meaningless jumble of data. These statistics have developed and become smart about understanding the relationships. The results will be valuable and useful in the decision making process (M. Wook et al., 2009). The principles and methods of data mining can be applied in different fields.
such as marketing, medicine, and all fields that used the data (Bunkar et al., 2012). Throughout this section we will provide an overview of some aspects relevant to the topics in this paper, such as the database and dataset, the pre-processing operation defined in detail as it is important and takes time and effort, and it is essential to do other operations and apply data mining techniques, the collection of attributes that can affect the result is not easy and it needs many works.

3.3. Preprocessing

The dataset which was used in this work taken from the main database used in the system of the Institute. Since we cannot deal directly with the database and use it directly with the application used in this study, the program 'Weka', and also because the data is usually stored in the database in the form of tables linked to each other with relations, primary keys and foreign keys, we will give an overview of the database used and table relationships among them as in Figure (5.1) below.

A large number of attributes can hinder predictive algorithms in a short time to reach interesting results (Costa et al., 2017; Crone, Lessmann, & Stahlbock, 2006), we had remove some of the attributes not required for this work and we had select just needed attributes. The selected attributes from different tables in the same database known as ‘View’, then we convert that view to text file, modified the text file and added the attributes names, relation modified the text file, finally save this text file as type 'arff' which is the type that most appropriate for the program 'Weka'. This phase was applied to the data before being transferred to the data store which is the most effortless and longest time in this work. We faced some of the mistakes and errors in the stored data, and we had treat with it as follows:

Birth date problem: We found that the date of birth of some students are (2619, 1199, and 0619). We was correct such errors to fit the logical situation, or even excluded some of these data.

Problem of registration number: We also found that some data is not logical for the student profile stored in the database. For example, we found numbers of some students (‘00000’, ‘000222’, ‘000555’, ‘0082002’, ‘0092307’, and ‘0092322’). After editing and deleting some records, we have 32,259 records from 34,600 and represent 1161 students, for which data mining techniques will be applied and explore the different patterns of student performance in the Institute in the above period that may be useful in the future.

![Figure 3.1. the Institute Entity Relations Database (ERD).](image)
3.5. Classification Model

Classification process is to classifying data objects with the specified class labels, it is the process of identifying models that define and differentiate data classes, such that an object whose class label is unknown can be predicted using the model. The model is based on a training data analysis set within a data object whose class is known (Imdad et al., 2017). Data mining techniques were used to predict students’ academic performance based on preceding academic performances (Mishra et al., 2014). By defining the distribution patterns based on available data, classification can be used for prediction. Classification is one of the algorithms overseen for machine learning. The pre-processed dataset was divided into two sets, for example training set and test set. The classifier model was developed using training set. The training set was applied and will be classified on the classification model. Classifier that is constructed from a set of objectives to define an abstract model of a test data. The classification model is then used for proper classification of new data for which the class label is unknown. Different approaches to building accurate classifiers have been proposed, such as naive Bayes classification, decision trees and SVMs. The association rules have recently become a useful technique for classification tasks (Shanmugarajeshwari and Lawrance, 2016; Baralis, Chiusano, & Garza, 2008). In data mining, classification of data is classified as a predictive method that is used to classify unseen data. There are two main steps in classifying data, namely the step of learning and the step of classification. In the learning stage, a model of classification is constructed on a training set using an algorithm. The training set used for the learning step must contain class labels for the data given. It is used for the prediction of class labels for unseen data after a classifier model is developed (Christian and Ayub, 2014). Classification is an important Data Mining. A variety of classification modeling algorithms, such as decision tree, neural network, support vector machines, nearest neighbor, were well developed and applied successfully in many fields (Gu et al., 2008). Although, Weka has six groups of classification (Bayes, function, lazy, Meta, misc, rules and trees), each of each has many algorithms, in this paper we will use just four algorithms (J48 pruned tree, IB1 instance-based, BayesNet and attribute Selection) and we compare the results of these algorithms.

IV. RESULTS AND DISCUSSIONS

In this section, we will discuss, analyze and compare the results of outcomes from the output of four classification algorithms. Classification matrix of four classifiers is shown in Table (4.1), while Table (4.2) shows the comparison of the classifiers on the classified instances and errors.

Table 4.1. classification Algorithms Accuracy

<table>
<thead>
<tr>
<th>Classifiers Algorithm</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>MCC</th>
<th>ROC Area</th>
<th>PRC Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48</td>
<td>0.99</td>
<td>0.01</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>1.00</td>
<td>0.99</td>
</tr>
<tr>
<td>IBk</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>BayesNet</td>
<td>0.95</td>
<td>0.02</td>
<td>0.95</td>
<td>0.95</td>
<td>0.95</td>
<td>0.93</td>
<td>1.00</td>
<td>0.982</td>
</tr>
<tr>
<td>Attribute Selection</td>
<td>0.90</td>
<td>0.03</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.88</td>
<td>0.98</td>
<td>0.91</td>
</tr>
</tbody>
</table>
In table (4.1) only the weighted average of all classes taken, since we need here the accuracy comparison between all four algorithms. Performance of any algorithm is evaluated on the basis of precision and recall. In our study, by comparing the three evaluate parameters (accuracy, Recall and Precision) for four algorithms had high accuracy which means a strong relationship between features that affects the performance of students, and that will help for prediction of student’s performance (Alqahtani and John 2017). As we can see in table (4.2), the algorithm (IBk) has the highest accuracy between all other algorithms reach to (100%), while the (Attribute Selection) algorithm has lowest accuracy which is (90%) which is also considered as strong. The total number of instances in dataset is (32259), and Incorrectly Classified Instances in (Attribute Selection) algorithm (3173), which is the biggest number of Incorrectly Classified Instances in all of our study.

<table>
<thead>
<tr>
<th>Classifiers Algorithm</th>
<th>Correctly Classified Instances</th>
<th>Incorrectly Classified Instances</th>
<th>Kappa statistic</th>
<th>Mean absolute error</th>
<th>Root mean squared error</th>
<th>Relative absolute error</th>
<th>Root relative squared error</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48</td>
<td>32100</td>
<td>159</td>
<td>0.99</td>
<td>0.01</td>
<td>0.04</td>
<td>1.19 %</td>
<td>10.90 %</td>
</tr>
<tr>
<td>IBk</td>
<td>32259</td>
<td>0</td>
<td>1</td>
<td>0.00</td>
<td>0.00</td>
<td>0.02 %</td>
<td>0.02 %</td>
</tr>
<tr>
<td>BayesNet</td>
<td>30530</td>
<td>1729</td>
<td>0.93</td>
<td>0.04</td>
<td>0.12</td>
<td>13.37 %</td>
<td>33.30 %</td>
</tr>
<tr>
<td>Attribute Selection</td>
<td>29086</td>
<td>3173</td>
<td>0.87</td>
<td>0.05</td>
<td>0.17</td>
<td>20.84 %</td>
<td>45.65 %</td>
</tr>
</tbody>
</table>

V. CONCLUSIONS

Educational database contains fast-growing amount of data. The classification algorithms are applied to the educational datasets to get the knowledge about student performance. The study focused on four classification algorithms (J48 pruned tree, IB1 based instance, BayesNet, and Selection attribute), we implement these four classifier algorithms on dataset selected from database of student performance in Higher Institute of medical Science Shati. In this paper, by comparing the three evaluate parameters (accuracy, Recall and Precision) for four algorithms, we had get the highest accuracy 100% which means a strong relationship between features that affects the performance of students, and that will help for prediction of student’s performance. The algorithm (IBk) has the highest accuracy between all other algorithms reach to (100%), while the Attribute Selection has lowest accuracy which is (90%), the (J48) also get high performance with ratio (99.5%) and the BayesNet get the value (94.6%). All the value obtained are high score and that reflect the strong relationship between features that affects the performance of students.

VI. REFERENCES


