www.ijaceeonline.com ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume 4, Issue 5, May 2019, pp. 1-6

Student Performance Prediction Using Educational Data Mining Techniques

Sumaiya Salim Khalfan Al Sulaimani
UG Scholar, Department of Electrical and Computer Engineering,
College of Engineering, National University, Sultanate of Oman
Email: sumaiya130134@cceoman.net

Dr. K. Vijayalakshmi
Assistant Professor, Department of Electrical and Computer Engineering,
College of Engineering, National University, Sultanate of Oman
Email: vijayalakshmi@nu.edu.om

Abstract: *One of the significant objectives of higher education institutions is to endeavor a good education quality to its students. Nowadays Educational Data Mining studies have emerged as a meaningful research field that discovers applicable knowledge from huge educational databases and repositories for various targets such as, predicting student performance, assisting for system and policy improvements in higher education. The aim of this paper is to predict the performance of students using different data mining techniques and to analyse the accuracy of data mining techniques. The paper attempts to explore the scope of data mining techniques, to enhance the quality of the educational system and improve teaching processes by evaluating student data to study the main factors that may affect their performance across various academic fields. The prediction will assist to identify the weak students and improve the performance of certain students who are predicted to fail. Various literatures are reviewed to study the existing systems and to analyse their limitations. Base on the study, the data mining techniques that are chosen for classification is clustering and for prediction the techniques that are chosen are Support vector machine, Naïve Bayes, Decision tree and Random Forest. Data collection is done from UCI machine learning repository and student performance prediction is modeled using the chosen techniques and accuracy of these models is compared. According to findings, decision tree technique is the best prediction tool in terms of data accuracy, response time and classification error.*

Keyword: *Student attributes; Data mining Prediction; Classification; Cross validation*

1. INTRODUCTION

The effectiveness of Educational Data Mining (EDM) techniques has expanded widely over different educational systems. The key goal of educational field is to maintain standards with efficient methods implemented for the development of teaching quality. This also helps to predict some trends in terms of educational standards approach. Moreover, students are the major influencers of the educational institutions; their performance plays an important role in the country's social and economic developments, by producing creative graduates, entrepreneurs, and innovators [5]. The ability to predict the performance can be a beneficial process of the current educational systems.

In this context, a flurry of attention has focused upon the concept of data mining techniques, since the development of educational information systems, which places an exceptional emphasis on discovering useful knowledge to aid students manage their deliverables and enhance their performance. Higher learning institutions, particularly, universities are the core of educational structures in which studies are performed in a competitive manner. The prerequisite mission of the universities is to collect, generate and exchange knowledge, as they commonly require information mined from current and past databases. On the above context, data mining is used widely in higher learning institutions to extract and study hidden data for subsequent procedures. These hidden data can be used in various educational procedures such as estimating dropout rates as well as predicting student performance.

It is significant to study educational data procedures, in particular students' performance, as it sets as a significant role in finding out interesting educational systems [8]. Currently, the lack of existing systematic techniques that used to monitor the student's progresses is still not addressed due to the insufficient methods for predicting the performances as well as

Cite this paper:

Sumaiya Salim Khalfan Al Sulaimani, K. Vijayalakshmi "Student Performance Prediction Using Educational Data Mining Techniques", International Journal of Advances in Computer and Electronics Engineering, Vol. 4, No. 5, pp. 1-6, May 2019.

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume 4, Issue 5, May 2019, pp. 1-6

the absence of investigation of factors affecting student's achievements within various contexts. Educational data mining considered as a new emerging field, which use data analysis techniques to identify hidden patterns in huge databases. This mining method has been successfully implemented in various arenas including the educational environment. Nowadays, educational data mining as an interesting study filed, which extract detailed, unknown patterns from different educational data sets in order to understand and improve the educational performance as well as the assessment of learning development. The main role of data mining is applying several techniques and algorithms to retrieve patterns from large data. Specifically, data can be easily taken from various kind of huge volume of databases in different formats such as scientific data, flat files, images, audios and modern kind of data formats.

The key goal of educational field is to maintain standards with efficient methods implemented for the development of teaching quality, involves some pertinent concerns to evaluate data quality. This also helps to predict some trends in terms of educational standards approach. Therefore, the usage of data mining systems is needed to follow up students' performance under a set of certain influences, which can lead to a series of teaching methods adaptation. In this respect, further studies are required to focus on student's attitude towards their performance and the affected study factors. This paper sets as a significance structure, as it is exemplifies the use of data mining in the educational filed, with the aid of data mining methods, which predicts students' performance in a professional manner.

This paper contains five sections. The first section provides a detailed introduction of the work. Followed by section two, which outlines a review of several studies on students' performance prediction using data mining. In addition, the third section addresses the different stages of data mining process and different data mining techniques that are considered. Fourth section includes a detailed description about implementation followed by results and discussion. Finally, a summary is outlined in section five.

2. BACKGROUND RESEARCH

In recent years, the significance of EDM has begun to come to forefront as a current trend of data mining that focuses in mining patterns and determining various educational information systems that are dealing with students in different educational stages.

Amjad et al led research with 270 students from different genders to recognize the relationship between students' personal and social factors and their educational performance. The study uses classification algorithms such as Naive Byes, CART decision tree, CHAID, C4.5 and ID4 decision tree to produce a qualitative prediction model. According to findings, it shows that the students' performance was not dependent only on their academic efforts, as there were several factors such as social factors, which have a greater influence on them [2].

Yassein, et al. [7] conducted the study and tested for 150 students. . The dataset was obtained from the community college, department of computer science and business administration at Najran University in the duration of two semesters, for each course, 12 attributes are considered, and for each student 7 attributes are considered. Course predictive model has split into two main stages which are data collection and preparation phase and data analysis phase. Moreover, the prediction of the student performance has been implemented using C4.5 algorithm and two clusters with the "success rate" target attribute, through "clementine" data mining tool, as well as the statistical package for social sciences which is used to arrange the data.

The study performed by Daud et al. [5] used the features and factors that are closely related to the academic performance as well as family income and assets. For instance, previous program scholarship and institution type, self-employed, father and mother income, earning hands, father status and guardian alive, as well as land value, location, house condition and number of vehicles at home. On the other Hand, some of the student personal data and family expenditures are ignored. Initially, the number of the records reaches about 3000 students, but after pre-processing and cleaning the data, the number of records is reduced to 776. Additionally, the data set are equally divided into two sets, 50% dropped students and 50% student who complete the course. The algorithm which used are, support vector machine, classification and regression tree (CART), Bayes network, naive Bayes and C4.5 algorithms. Furthermore, for the performance evaluation and comparison, 5-fold cross validation technique is used to calculate F1 score. As a result, C4.5 classification algorithm is the best predictor of the student's performance, then Bayes network and naive bays algorithms place the second and third highest F1- scores.

3. METHODOLOGY

The student performance prediction has been processed into four phases.

3.1 Data collection

The data was extracted from UCI Machine Learning Repository. Moreover, the extracted dataset comprises records of secondary education students' achievement from two public Portuguese schools that was collected during the year of 2005-2006 by using school reports and questionnaires. The dataset holds about 395 examples, 649 records, and 33 attributes [4]. The dataset contains several attributes types, which are, nominal, numerical and binary. Specifi-

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume 4, Issue 5, May 2019, pp. 1-6

cally, the attributes contain students' grades, demographic and social information as well as school-related features. Figure 1 describes each attribute name and type. G3 is the attribute that has a powerful correlation with G1 and G2, due to G3 is the final year grade. Regarding to the five level of classification system, the grade and status columns are added into the data set to classify the pass and failure students. The five level classification system is made as shown in Figure 2 by adding attributes grade and status columns to the dataset to classify the students to pass or fail.

3.2 Data preparation and pre-processing

Data must be prepared and pre-processed before any procedure or operation is implemented on it, in order to improve the data efficiency and quality before mining. This stage involves several sub stages including, data transformation, data cleaning, data blending and data filtering.

Attribute	Description (Domain)
sex	student's sex (binary: female or male)
age	student's age (numeric: from 15 to 22)
school	student's school (binary: *Gabriel Pereira* or *Mousinho da Silveira*)
address	student's home address type (binary: urban or rural)
Pstatus	parent's cohabitation status (binary: living together or apart)
Medu	mother's education (numeric: from 0 to 4[a])
Mjob	mother's job (nominal[b])
Fedu	father's education (numeric: from 0 to 4[a])
Fjob	father's job (nominal[b])
guardian	student's guardian (nominal: mother, father or other)
famsize	family size (binary: ≤ 3 or > 3)
famrel	quality of family relationships (numeric: from 1 – very bad to 5 – excellent)
reason	reason to choose this school (nominal: close to home, school reputation, course preference or other)
traveltime	home to school travel time (numeric: 1 – < 15 min., 2 – 15 to 30 min., 3 – 30 min. to 1 hour or 4 – > 1 hour).
studytime	weekly study time (numeric: 1 – < 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – > 10 hours)
failures	number of past class failures (numeric: n if $1 \leq n < 3$, else 4)
schoolsup	extra educational school support (binary: yes or no)
famsup	family educational support (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
paidclass	extra paid classes (binary: yes or no)
internet	Internet access at home (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
freetime	free time after school (numeric: from 1 – very low to 5 – very high)
goout	going out with friends (numeric: from 1 – very low to 5 – very high)
Walc	weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
Dalc	workday alcohol consumption (numeric: from 1 – very low to 5 – very high)
health	current health status (numeric: from 1 – very bad to 5 – very good)
absences	number of school absences (numeric: from 0 to 93)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20)

Figure 1. Student Performance dataset attributes (Cortez & Silva, 2008)

Country	I (excellent/very good)	II (good)	III (satisfactory)	IV (sufficient)	V (fail)
Portugal/France	16-20	14-15	12-13	10-11	0-9
Ireland	A	B	C	D	F

Figure 2. The five classification level system (Cortez & ilva, 2008)

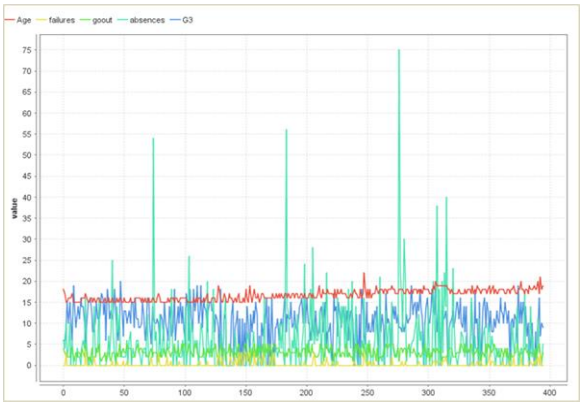

Figure 3. Data set attributes

The attributes that have more weight are used to implement the models and evaluate their performance. The most weighted attributes are age, family size, failure, higher, gout, absences, g3, grade and status. In addition, the graph in Figure 3, determines the number of students who are involved in each attribute range. After data preparation, each model must be trained and evaluated by using cross-validation.

3.3 Cross validation

It is a technique used to evaluate the predictive models, by splitting the original dataset into a training set to train the model and a test set to evaluate the model with the measures it measures the accuracy, precision and recall [6].

Accuracy, Precision and Recall are calculated by using the following equations:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision(p) = \frac{TP}{TP + FP}$$

$$Recall(r) = \frac{TP}{TP + FN}$$

$$Classification\ errors = \frac{FP + FN}{TP + TN + FP + FN} = \frac{errors}{total}$$

Measures of evaluating data mining model (Shehroz, 2016)

Where,
TP: describes the number of True Positives.
FN: False Negatives
FP: False Positives
TN: True Negatives

The accuracy is specified as the percentage of correctly classified examples. In addition, the recall is defined as the sensitivity or true positive rate, also, precision is indicated as positive predictive value (PPV).

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume 4, Issue 5, May 2019, pp. 1-6

4. DISCUSSION

4.1 Naïve Bayes

It is a high-bias, low-variance probability classifier that is able to construct a good model within a small dataset. Generally, it can be used in case of sentiment analysis, text categorization and recommender systems. This model is simple to implement and computationally inexpensive. The classifier works by giving the value of the label of any attribute that is independent on the value of the other attribute [3].

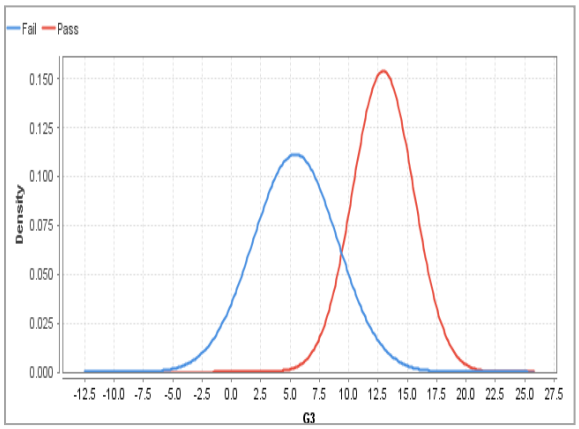

Figure 4. Naive Bayes classifier graph

TABLE 1. NAÏVE BAYES PERFORMANCE

Accuracy:98.73%			
	true fail	true pass	class precision
pred.Fail	26	1	96.30%
pred.Pass	0	52	100.00%
class recall	100%	98.11%	

As, shown in Figure 4, the graph describes the normal distribution of G3 marks. Specifically, the probability density function, predict the student who are pass or fail based on the density of the G3 attribute. The highest G3 density defines the pass students, otherwise the failures. Regarding the performance as shown in table 1, the Naïve Bayes algorithm accuracy is 98.73% with 1.27% of classification errors

4.2 Support Vector Machine

SVM is non-binary linear algorithm that classifies and separates the data into two categories, by mapping data and constructing an N-dimensional hyperplane. SVM is able to determine and define which class the new data object belongs in [9]. The Support vector machine model has the similar accuracy of the Naïve Bayes algorithm, which is 98.73%. As well as the classification error that is equal to 1.27% which is shown in table 2.

TABLE 2. SUPPORT VECTOR MACHINE PERFORMANCE

Accuracy:98.73%			
	true fail	true pass	class precision
pred.Fail	26	1	96.30%
pred.Pass	0	52	100.00%
class recall	100%	98.11%	

4.3 Decision tree

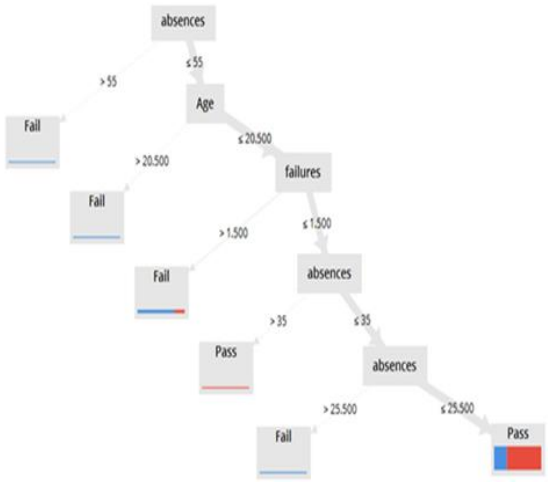

Figure 5. Decision tree for student data set

It is a collection of nodes designated to generate a decision on numerical target values. The maximum depth is the tree depth that depending on the size and dataset characteristics. Each node represents a dividing rule for one particular attribute, in order to minimize the errors in a proper way for the chosen parameter criterion. The prediction is determined depending on the majority of the examples that reached this leaf during an estimation of the numerical value which is obtained by averaging numbers in a leaf [1]. The decision tree arrived with the depth set to 25 for student data set is given in Figure 5.

TABLE 3. DECISION TREE PERFORMANCE

Accuracy:99.40% +/- 1.01%(mikro: 99.49%)			
	true fail	true pass	class precision
pred.Fail	128	0	100.00%
pred.Pass	2	265	100.00%
class recall	98.46%	100.00%	99.25%

The decision tree has 0.0% classification errors and 99.94% of accuracy, the table 3 is the accuracy table of the decision tree algorithm, where the class recall is 98.46%, and the Precision is 100.00% (positive class: Pass) as given in table 3.

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume 4, Issue 5, May 2019, pp. 1-6

4.4 Random forest

The Random forest constructs multiple trees and combines them together, in order to produce proper and accurate prediction. In addition, processes for finding the root node and dividing the feature nodes are randomly done. Depending on the inputs, the random forest algorithm has produced five decision trees, each tree delivers different prediction knowledge. For instance, in tree 1 the students are predicted to fail or pass depending on G2 and G3 marks if G2 mark ratio is >9.500, the student will pass. Otherwise, G3 mark will be checked whether it is >9.500 or <= 9.500, if it is <= 9.500 the student is predicted to fail, otherwise pass. Additionally, in tree two, the "study time" is the influence attribute that determines the student achievements. According to the performance table 4, the random forest model accuracy is 97.47%, where the classification error is 2.53%. This algorithm is easy to implement and measure the relative importance of all features on the prediction. But, it is slow compared to other models.

TABLE 4. RANDOM FOREST PERFORMANCE

Accuracy:97.47%			
	true fail	true pass	class precision
pred.Fail	24	0	100.00%
pred.Pass	2	53	96.36%
class recall	92.31%	100.00%	99.25%

4.5 Runtime and Classification errors

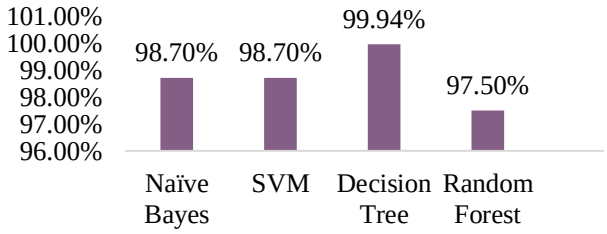

Figure 6 Models' runtime

Naïve Bayes algorithm has lowest runtime, which means it produces the output faster than SVM, decision tree and random forest algorithms. Conversely, the random forest algorithm is the slowest model, it takes about 57 seconds to run and produce the output due to the number of the trees. As well as, SVM and Decision tree algorithms approximately has run times 17 and 19 seconds which is given in Figure 6. Moreover, Naïve Bayes and SVM algorithms have the similar classification error percentage which is 1.27%, while decision tree has 0% of classification errors. In contrast, random forest algorithm has the highest classification error.

4.6 Accuracy and Performance

Naïve Bayes and SVM have the similar percentage of the accuracy which is 98.70%, where the decision tree model has the highest accuracy of 99.94% which is shown in Figure 7. In the other hand, random forest has the less accuracy of 97.50%. Table 5 describes the overall comparison of the different classification and regression algorithms. Naïve Bayes and SVM algorithms have the similar classification error percentage which is 1.27%, while decision tree has no classification errors. In contrast, random forest algorithm has the highest classification error due to imbalanced data set where the class distribution is not uniform among the classes.

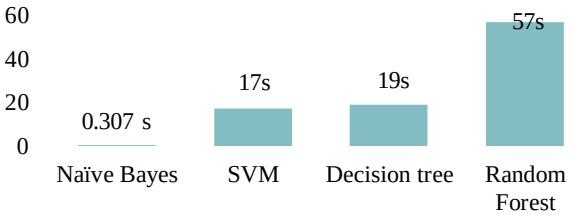

Figure 7. Models Accuracy

In the other hand, the random forest is the model which has lowest accuracy. In addition, the random forest is the slowest model comparing to other with higher classification error. More accuracy requires more trees, which makes the model slower. Furthermore, Naïve Bayes and Support Vector Machine has the same average of accuracy and classification errors. However the run time is more for Support Vector Machine due to it's over complexity. As shown in table 5, the decision tree is the model which has the highest accuracy and lowest classification errors. The results shows that decision tree is the most effective model since decision trees do not require any assumptions of linearity in the data. As the students' performance attributes are nonlinearly related, the decision tree outperforms other models.

TABLE 5. EVALUATION OF MODELS

Model	Naïve Bayes	Random Forest	Support Vector Machine	Decision Tree
Accuracy	98.7 %	97.5 %	98.7 %	99.94%
Runtime	0.307 secs	57 secs	21 secs	19 secs
Classification Errors	1.27%	2.5 %	1.27%	0.0%

ISSN: 2456 - 3935
International Journal of Advances in Computer and Electronics Engineering
Volume 4, Issue 5, May 2019, pp. 1-6

5. CONCLUSION

Educational data mining is relatively new promising area of research, which aims to increase educational experiences by assisting students, instructors and researchers to make better decisions using data available about students. This paper aims to predict the performance of students using different data mining techniques and to evaluate these techniques. Different data mining techniques such as Support Vector Machine, Decision Tree, Naïve Bayes and Random Forest are deployed and evaluated. The findings show that the decision Tree technique is the best key prediction tool in terms of accuracy and run time. According to findings, it is clearly found that students' performance is not totally reliant on their efforts, as there are a set of other factors which have equal importance that influence their academic tasks. This research can significantly motivate various educationalists to perform data mining procedures on their students' data sets to find out valuable information which could help in the development of various educational structures.

REFERENCES

[1] Ahmad, F. Hafieza, N and Abdul Aziz, A. (2015), "The Prediction of Students Academic Performance Using Classification Data Mining Techniques" *Applied Mathematical Sciences*, Vol.9, Issue.129, pp. 6415 – 6426.

[2] Amjad AbuSaa, A. (2016), "Educational Data Mining & Students' Performance Prediction" *International Journal of Advanced Computer Science and Applications*. Vol. 7, Issue.5,

[3] Cios, K.J., Pedrycz W., Swiniarski, R.W. and Kurgan, L.A. (2007), "Data Mining: A Knowledge Discovery Approach", Springer, New York, pp.27-30

[4] Cortez, P. and Silva, A, Student Performance Data Set, retrieved date:[1, March, 2018] online available at:https://archive.ics.uci.edu/ml/datasets/student+performance

[5] Daud, A., Aljohani, N.R., Abbasi, R.A., Lytras, M.D., Abbas, F. and Alowibdi, J.S.(2017), "Predicting Student Performance using Advanced Learning Analytics", *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 415-42.

[6] Nichat, A and Raut.A.(2017), "Predicting and Analysis of Student Performance Using Decision Tree Technique" *International Journal of Innovative Research in Computer and Communication Engineering*, Vol.5, Issue.4, pp. 7319-7328

[7] Nawal Ali Yassein, Rasha Gaffer M Helali and Somia B Mohomad1. (2017), "Predicting Student Academic Performance in KSA using Data Mining Techniques" *Journal of Information Technology & Software Engineering*.

[8] Osmanbegović, E and Suljić,S. (2012), "Data Mining Approach For Predicting Student Performance" *Journal of Economics and Busines*, . Vol.3, Issue.1.

[9] Sayad,S. Support Vector Machine, retrieved date:[15, April, 2018], online available at: http://www.saedsayad.com/support_vector_machine.htm

Authors Biography

Sumaiya Salim Khalfan Al Sulaimani, is a UG student of Department of Electrical and Computer Engineering, College of Engineering, National University, Sultanate of Oman. She completed her BEng in Computer Engineering. Her research interests are Programming, Data mining and Web development.

Dr. K. Vijayalakshmi, is Assistant Professor in Department of Electrical and Computer Engineering, College of Engineering, National University. She completed her BE and ME in Computer Science and Engineering in Madurai Kamaraj University. She completed her Ph.D in Anna University in the area of Component based Software Development. Her research interests are Machine Learning, Software Quality Management and Component based Software Development and Data Mining.

Cite this paper:

Sumaiya Salim Khalfan Al Sulaimani, K. Vijayalakshmi "Student Performance Prediction Using Educational Data Mining Techniques", International Journal of Advances in Computer and Electronics Engineering, Vol. 4, No. 5, pp. 1-6, May 2019.