Tracking Student Performance in Introductory Programming by Means of Machine Learning

Ijaz Khan[1,2]
[1]*Information Technology Dept. Buraimi University College,* Al-Buraimi, Oman
[2]*Dept. of Systems and Networking, University of Technology(UniTen)* Kajang, Malaysia
ijaz@buc.edu.om

Abir Al Sadiri[1,2]
[1]*Information Technology Dept. Buraimi University College,* Al-Buraimi, Oman
[2]*Dept. of Systems and Networking, University of Technology(UniTen)* Kajang, Malaysia
abir@buc.edu.om

Abdul Rahim Ahmad
Dept. of Systems and Networking, University of Technology(UniTen)
Kajang, Malaysia
abdrahim@uniten.edu.my

Nafaa Jabeur
Computer Science Dept.
German University of Technology,
Muscat, Oman
nafaa.jabeur@gutech.edu.om

Abstract*— large amount of digital data is being generated across a wide variety of fields and Data Mining (DM) techniques are used transform it into useful information so as to identify hidden patterns. One of the key areas of the application of Education Data Mining (EDM) is the development of student performance prediction models that would predict the student's performance in educational institutions. We build a model which can notify students (in introductory programming course) about their probable outcomes at an early stage of the semester (when evaluated for 15% grades). We applied 11 Machine Learning algorithms (from 5 categories) over a data source using WEKA and concluded that Decision Tree (J48) is giving higher accuracy in terms of correctly identified instances, F-Measure rate and true positive detections. This study will help to the students to identify their probable final grades and modify their academic behavior accordingly to achieve higher grades.

Keywords—educational data mining, machine learning, decision tree, Weka

1. INTRODUCTION

Numerous data is being produced and collected across a wide variety of fields. The big data produced needs to be collected, organized and processed to extract valuable information [1]. The real-world domains and industries needs to analyze large amount of data generated to obtain useful information. In order to do so, the Data Mining techniques are used to build model which analyses the provided dataset and identify useful pattern in the data [3]. Data Mining (DM), comprises of computational techniques to process data and identify useful information [2].

Educational Data Mining (EDM) is an emerging discipline of Data Mining. The key objective of EDM is to collect data about learner and their learning environment, and propose novel methods to identify useful patterns within this data which appear beneficial in apprehending learner, learning environment and factors associating both of them [4]. The Education institutions always show keen interest in collecting data about their students. The significant processing of this data can highlight the areas where the institutions need improvement. This interest in data collection and processing has increased with advent of big data analysis and the increased adoption of online learning environments. The collection and analysis of data from learner and their environment is appearing useful in supporting and obtaining insight into students learning activities [5] [6] [7].

One of the key areas of the application of EDM is the development of student performance prediction models that would predict the student's performance in educational institutions based on some underlying factors that are given as input. This student prediction model is vital for educational institutions to build strategy for improving the quality of teaching process. It is helpful in identifying the students at risk of low academic achievements [1, 2]. This prediction is beneficial issuing early warnings to students with low academic performance and the in-time interventions can help students improve their performance.

Introductory programming courses are taught at the early stages of degree courses. Since the novice programmers have hardly any idea that learning programming courses are much more different that the subjects they usually study during their school, therefore, they tend to make numerous mistakes and face hardships which often results in their failure to achieve higher grades [3] [4] [5] [6]. The institutions are much concerned about the failure rates and worried about an increase in retention of such subjects. In this paper, we use Machine Learning Algorithms to identify the students who are likely to obtain low grades in introductory course.

The structure of the paper is: in next section, we provide an overview of Machine Learning and popular algorithms, section-3 highlights the related work and then we give our contributions briefly in section-4. Section-5 provides our methodology and in subsequent sections we provide details of our data source, pre-processing, experiments, and findings. Section-9 concludes the paper and section-10 provides our future plans for extending our work.

2. MACHINE LEARNING METHODS

Machine Learning Algorithms is a set of useful tools used for developing student performance prediction models. Assessing students' performance prediction is very complex issue and various algorithms are used for this purpose. Machine learning (ML) is a branch of Artificial Intelligence [7] that may refer to learning from past experience (previous data) to improve future performance automatically without any external assistance from human [8]. More formally, Machine Learning is defined as: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" [9]. The prediction models would not only inform the students during their study, but also provide real time advice to rectify their problems.

The main difference between computer and human is the ability to think and learn from experience. Computers do not learn from experience, they execute the human-made algorithms to learn from it, get training and make predictions. ML algorithms have been found that are effective for some types of learning tasks. They are particularly useful in situations where the human may not have sufficient knowledge and it became harder to develop

978-1-5386-8046-9/19/$31.00 ©2019 IEEE

effective knowledge-engineering algorithms. ML algorithm receives input data which it explores to get training and produce hypotheses. This training make possible for them to make predictions about future. In this section, we briefly explain some of the widely-used ML algorithms.

A. Naive Bayes Algorithm:

Naïve Bayes (NB) [10], is a classification technique, based on Bayesian theorem [17]. It is a simple and efficient classification algorithm [3] [18]. It proves efficient in situations which rely on (a) class condition independence [19]; which assumes that the effect of an attribute value on a given class is independent of the values of the other attributes and (b) assuming no hidden attributes could affect the prediction process.

B. Multilayer Perceptron (MLP):

Multilayer Perceptron (MLP) is one of the commonly used neural networks algorithm [18]. It is particularly appropriate in situations where the relationship between input and output attributes is vague; which is called as approximation of classification function [3].

C. Support Vector Machine (SVM):

Support Vector Machine (SVM) [20], derive features from variables, and manage them in linear combination to achieve a prediction decision (either classification or regression) [21] [22].

D. Decision Tree:

Decision Tree uses a recursive technique to construct a tree and achieve maximum prediction accuracy. Decision Tree uses different mathematical algorithms, such as Gini index, information gain, and Chi-square statistics. The decision tree starts with a root node, and branching through intermediate nodes (called leaf nodes) it ends at the last node (called as end node) [23].

Some of the popular decision tree algorithms areID3, C4.5, C5 [24] [25] and CART [26]

3. RELATED WORK

This section provides an overview of the work of various authors performed for prediction of student's academic performance.

The work presented in [11] combine data recorded from students' programming process (using Test My Code) with ML algorithms to classify student as high and low performing after the early stages of an introductory programming course. In the paper [12] classifications techniques are used to predict student's overall performance and categorize them as slow, average and fast learners. The study in [13] presents a comparative study on the effectiveness of educational data mining techniques for early identification of students at risk of failing programming course. The authors also addressed the impact of data pre-processing and algorithms fine-tuning on ML techniques. Through survey cum experimental methodology, [14] discover the factors that are considered to have influence on the students' performance and then use ML algorithms to identifying slow learners so that teachers can intervene in good time and help them individually to improve their performance. [15] compares ML algorithms to find out the suitable algorithm which can accurately predict student's performance in distance learning and thus helps tutor recognize students with a high probability of poor performance. The work in [16] investigates the reasons behind the lack of success in cores subjects of Portuguese secondary education. The results show that student performance is highly affected by grades from previous first/second school periods. In their work [17] used five classifiers, with different configurations settings, and observed an accuracy measure up to 92% with Decision Tree and Neural network appearing with better performance based on predictive accuracy as well as misclassification cost measure. [18] applied decision tree algorithms on students' past performance data to generate the model which can be applied to identify the dropouts and students who require special attention due to their low academic performance. [19] use CRISP framework to discover the key attributes that may affect the student performance and then make use of classification techniques to build a prediction model which allows students to predict the final grade in a course addressed in their study. [20] describe the results of a case study aimed at predicting students drop out after the first semester of their studies or even before they enter the addressed program. Three supervised data mining algorithms were applied in [21] to build a model which can predict student success in course. [22] propose WAVE architecture to identify the students at dropout risk, using only time-varying student data. The proposed architecture allows monitoring the academic progress and identifying the students with low academic performance. The work in [23] apply White-Box classifiers algorithms (for rules and for decision tree) for the detection of students' failure on real time data in order to improve their academic performance and to prevent them dropping out. [24] use white-box classification methods (induction rules and decision trees) to builds prediction model for predicting students at risk of failing or dropout at middle-school at Zacatecas, México. The experiments were performed first with using all the available attributes; and next, by selecting the best attributes; and finally, involving rebalancing data and using cost sensitive classification. The paper [25] presents the initial results from a data mining research project implemented at a Bulgarian university. The paper is focused on the implementation of data mining techniques to discover the hidden patterns in the available data which can be useful in predicting students' performance based on their personal and pre-university characteristics. Some other works include [26] [27] [28] and [29]

4. OUR CONTRIBUTIONS

Introductory programming courses are amongst the starter packs for the students aiming to make their future in Computer Science. However, the novice programmer finds it difficult to digest the suddenly appearing difficulties at the initial stages of their studies. Therefore, the institutions have keen interest to know the reasons behind the failure rate in programming courses. We want to develop a model to predict student's final grades in introductory programming course at an early stage of the semester. The model will be helpful to notify them about their probable outcome at the early stages of the semester, so that they can modify their academic activities to increase their grades. In order to achieve this task, we need to analyse the student's data using Machine Learning algorithms to build a model which can predict students' final grades at the early stages of the semester. We conduct series of experiments in WEKA to see the performance of various classifiers over the dataset with carefully pre-processed data. We will find answers to the following questions:

Q1: How can we predict the introductory programming course student's performance is the early stages of semester?

5. METHODOLOGY

We used the process of Knowledge Discovery and Data Mining methodology [24] in predicting student's academic performance (figure-1). This method consists of 4 main

phases a) Data Gathering b) Data Pre-Processing c) Data Mining d) Interpretation

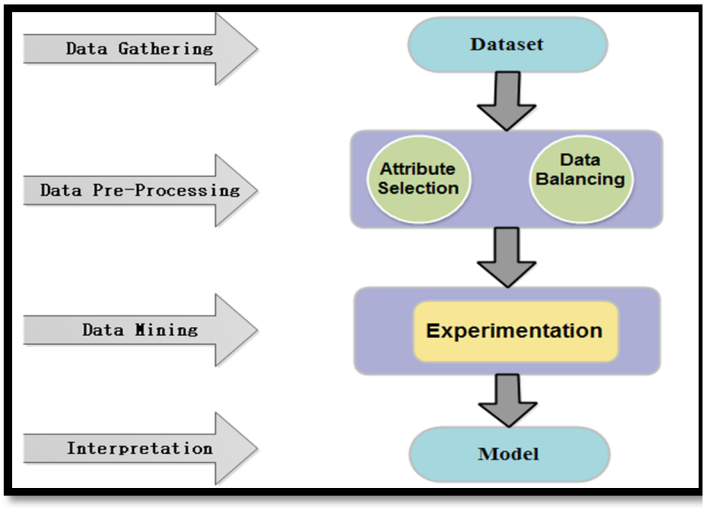

Fig. 1 The methodology used in our approach

a) Data Gathering: All the available information about students are collected and arranged into a dataset. We collected student's information from college Registration department which includes their demographic information, and their academic records. We formed a dataset where we defined necessary student's attributes and defined whether each attribute have real or nominal value. For example, "Gender" is student attribute which have nominal value (Male, Female). The final dataset contains complete record of 50 students is stored in the format required by WEKA.

b) Data Pre-Processing: A number of issues with the dataset may affect the accuracy of the ML algorithms; therefore, the dataset is further scrutinized to get better model. We performed pre-processing of data using different algorithms provided in WEKA (explained later in this paper). By the end of this stage our dataset is ready for Machine Learning implementation.

c) Data Mining: The final dataset is provided as input so ML algorithms which can apply and predict student's academic performance. We selected 13 algorithms from 5 categories of ML i.e. Bayes (Naive Bayes), Function (Support Vector Machine, Multilayer Perceptron (MLP), Lazy (IBK), Rules (Decision Table, JRip, OneR, PART and ZeroR) and Tree (J48, Random Forest, Random Tree, and SimpleCART). All the algorithms have been have been executed, evaluated and compared in order to determine which one obtains the best results.

d) Interpretation: Finally, the obtained results are analysed to produce a model for prediction of student academic performance. We designed our Machine Learning model with the algorithm providing the highest accuracy. We can then test our model with another dataset (as Supplied test set in WEKA) to evaluate its performance, accuracy and correctness.

We used WEKA [30] to perform the pre-processing of our data and performing classification evaluation experiments. WEKA is open source software that offers a collection of data mining algorithms for data pre-processing, classification, regression, clustering, and association rules.

6. Data Gathering (Dataset)

The data source used in this project is from Buraimi University College (BUC), an educational institution located in Oman. The data source, comprising of 50 instances, comes from 2 sections. Here a semester lasts for 16-17 weeks and students are assessed throughout the semester with exams and continuous assessment tools (quizzes, assignments, presentations).

We want to predict the student's outcome after 15% of total grades are awarded. To build a classification model for the early warning system, the prediction variable (Grades) consists of 3 classes, i.e. Low (marks less than 60), Medium (60-84) and High (More than 85). The pass marks are 50, but the grade value for marks in range 50-60 is 1 which means it will not affect the overall CGPA.

7. Data Pre-Processing

Initially, we manually removed the features such as Student_ID, Student_Name, Course_Section and other attributes which do not seem to have an impact over the student performance. Therefore, we started with the following set of features.

TABLE 1.THE MAJOR ATTRIBUTES

Feature	Domain
Gender	Nominal (Male, Female)
Course_Code	Nominal (Course IDs of five courses)
Course_Level	Nominal (Two, Three)
Session	Nominal (Morning, Evening)
Class_Duration	Nominal (1, 1.5)
CGPA	Real
Major	Nominal (CS, IS, SE)
Degree	Nominal (Bachelor, Diploma)
Year	Nominal (Year-1, Year-2, Year-3, Year-4)
Attendance	Real
Test1_Marks	Real
Grade	Nominal (High, Medium, Low)

In order to deal with the important problems that may exist in educational data mining, we performed the pre-processing of data source. High Dimensionality (dataset having large number of attributes) and Unbalanced Data (dataset where the number of instances from one class is much larger than the number of instances from other class) are the two major problems likely to affect the performance of classification [24] [31].

A) Removing High Dimensionality

Our data source suffers from high dimensionality; therefore, we evaluated attributes selection algorithms provided by WEKA on our data source. We performed feature selection to decrease the number of overlapping features, possible over fitting, and to probably improve predictive accuracy of the feature set [11]. We used correlation-based feature subset selection [32] (CorrelationAttributeEval) and Information-gain feature subset [33] (InfoGainAttributeEval) both with Ranker search methods.

Table-2 highlights the results of both algorithms. We keep 0.20 as cut-off significance value and ignore attributes below this value. The table-2 illustrates that Test1_Marks and CGPA appears vital in both the algorithms.

TABLE 2. THE ATTRIBUTE SELECTION

Attribute	InfoGainAtribute	CorrelationAttri	CFSSUBsetEval
Test1_Marks	0.66463	0.3405	✓
CGPA	0.63985	0.3231	✓
Attendance	0.20942	0.1148	
Major	0.11002	0.1693	
Gender	0.07081	0.1823	✓
Year	0.08614	0.1782	

Further we used CFSSUBsetEval with GreedyStepwise search method provides us CGPA, Test1_Marks, and Gender as essential attributes.

We can conclude from the above comparison that CGPA, Test1_Marks, and Attendance are the strong predictors. Further, CFSSUBsetEval is emphasising on Gender attributes as well. Finally, we select CGPA, Test1_Marks, Attendance, and Gender as the set of attributes we will use in our prediction model.

B) Balancing Data

There are 16% of students in "Low" class in our dataset which result in a serious class imbalance problem. It is evident from [34] that the adjustment of the ratio of class samples can improve the machine's learning performance. In order to balance the instances of each class, we applied SMOTE [35] to modify the distribution of instances.

8. Experimentation and Interpretation

We used WEKA to construct the classification model. The evaluation is performed using k-fold cross validation in WEKA (using 10-fold cross validation), which is a common practice in predictive data mining applications [36].

A) Performance Measurement

We compared the effectiveness of applied classifiers through their F-Measure value [37]. F-Measure (Eq.(1)) is the harmonic mean between precision (Eq.(2)) and Recall (Eq.(3)), as described below [13]:

$$\text{Fmeasure} = 2x\ ((\text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall})) \qquad \text{(Eq. 1)}$$

$$\text{Precision} = TP/(FP + TP) \qquad \text{(Eq. 2)}$$

$$\text{Recall} = TP/(FN + TP) \qquad \text{(Eq. 3)}$$

True Positive (TP): Positive instances, and classified as positive

False Positive (FP): Negative instances, but classified positive

False Negative (FN): Positive instances, but classified negative.

B) Results and Findings

We compared the performance of each of the classifier. The graphs in figure-2 compare the performance (correctly classified instances %) of all classifiers. It is observed that Decision Tree (J48) have achieved an accuracy of 88% followed by Naïve Bayes (nB) and Decision Table, 84% and 83% respectively. Overall, the Decision Tree family algorithms have achieved higher accuracy comparing to other categories of algorithms.

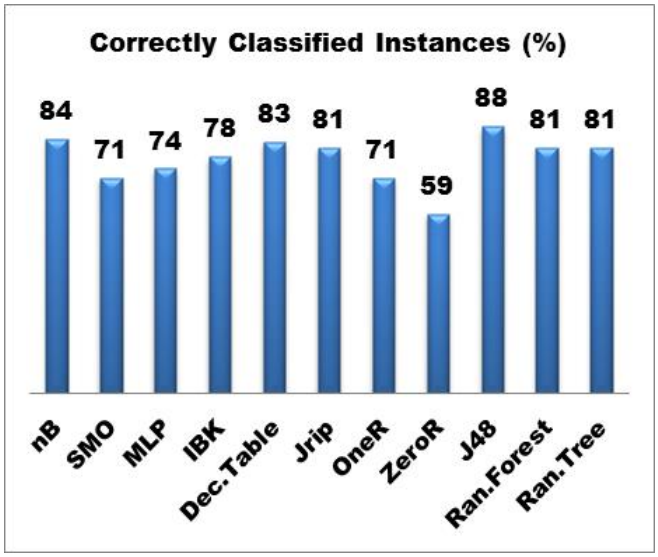

Fig. 2 Comparative analysis of classifiers

In order to further performed in-depth analysis, we selected the classifiers with high accuracy and performed experiment in WEKA experimenter. WEKA experimenter compares the classifiers. We compared the classifiers through t-test (with F-Measure) to compare the degree of their effectiveness. The results snapped from WEKA (Figure-3) experimenter shows that J48 (0.88 F-Measure) outplays other algorithms.

```
Dataset                    (1) bayes.N | (2) rule (3) tree
-----------------------------------------------------------
'3-Class Linear Dataset_T(100)   0.83 |   0.79     0.88
-----------------------------------------------------------
                              (v/ /*) |  (0/1/0)  (0/1/0)
```

Fig. 3 WEKA Experimenter. t-test (with F-Measure)

Similarly, the t-test (with True Positive (TP)) also confirms that Decision Tree (J48) is performing better than other classifiers. This is depicted in Figure 4.

```
Dataset                    (1) bayes.N | (2) rule (3) tree
-----------------------------------------------------------
'3-Class Linear Dataset_T(100)   0.83 |   0.84     0.90
-----------------------------------------------------------
                              (v/ /*) |  (0/1/0)  (0/1/0)
```

Fig. 4 WEKA Experimenter. t-test (with True Positive (TP))

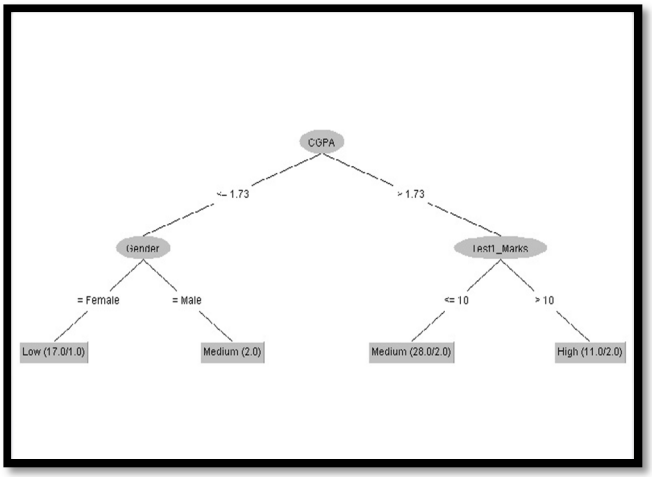

Fig. 5 Decision Tree for J48

The analysis confirms that Decision Tree (J48) is an appropriate classifier for modelling our early prediction. T-test for F-Measure shows that J48 Decision Tree is able to reach an effectiveness of 88% when the students have taken their first exams (15% of overall grade).

9. CONCLUSION

Our aim is to develop a prediction model which can notify a student about his/her probable outcomes at the early stages of the semester. In order to achieve this we applied 11 classification algorithms over a data source using WEKA. We concluded that Decision Tree family algorithms achieve high accuracy with J48 to be the most appropriate with 88%.

This study will help to the students and the teachers to improve the performance of the students. This study will also work to identify those students which needed special attention and will also work to reduce fail ratio and taking appropriate action for the next semester examination.

10. FUTURE WORK

In future we are aiming to develop a Prediction Model integrated in Java Language which will track student performance through the semester and will notified them in case they are in danger of failing the course.

We also agree that the dataset selected for our experiments is very small and will plan for a much bigger dataset in the future.

REFERENCCES

1. Musso, M. and E. Cascallar, NEW APPROACHES FOR IMPROVED QUALITY IN EDUCATIONAL ASSESSMENTS: USING AUTOMATED PREDICTIVE SYSTEMS IN READING AND MATHEMATICS. Problems of Education in the 21st Century, 2009. 17.
2. Ramaswami, M. and R. Bhaskaran, A CHAID based performance prediction model in educational data mining. arXiv preprint arXiv:1002.1144, 2010.
3. Watson, C. and F.W. Li. Failure rates in introductory programming revisited. in Proceedings of the 2014 conference on Innovation & technology in computer science education. 2014. ACM.
4. Hanks, B., et al. Program quality with pair programming in CS1. in ACM SIGCSE Bulletin. 2004. ACM.
5. Iepsen, E.F., M. Bercht, and E. Reategui. Detection and assistance to students who show frustration in learning of algorithms. in Frontiers in Education Conference, 2013 IEEE. 2013. IEEE.
6. Tan, P.-H., C.-Y. Ting, and S.-W. Ling. Learning difficulties in programming courses: undergraduates' perspective and perception. in Computer Technology and Development, 2009. ICCTD'09. International Conference on. 2009. IEEE.
7. Marsland, S., Machine learning: an algorithmic perspective. 2015: CRC press.
8. Das, K. and R.N. Behera, A Survey on Machine Learning: Concept, Algorithms and Applications. International Journal of Innovative Research in Computer and Communication Engineering, 2017. 5(2): p. 1301-1309.
9. Michalski, R.S., J.G. Carbonell, and T.M. Mitchell, Machine learning: An artificial intelligence approach. 2013: Springer Science & Business Media.
10. Domingos, P. and M. Pazzani, On the optimality of the simple Bayesian classifier under zero-one loss. Machine learning, 1997. 29(2-3): p. 103-130.
11. Ahadi, A., et al. Exploring machine learning methods to automatically identify students in need of assistance. in Proceedings of the eleventh annual International Conference on International Computing Education Research. 2015. ACM.
12. Mhetre, V. and M. Nagar. Classification based data mining algorithms to predict slow, average and fast learners in educational system using WEKA. in Computing Methodologies and Communication (ICCMC), 2017 International Conference on. 2017. IEEE.
13. Costa, E.B., et al., Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. Computers in Human Behavior, 2017. 73: p. 247-256.
14. Kaur, P., M. Singh, and G.S. Josan, Classification and prediction based data mining algorithms to predict slow learners in education sector. Procedia Computer Science, 2015. 57: p. 500-508.
15. Kotsiantis, S., C. Pierrakeas, and P. Pintelas, PREDICTING STUDENTS'PERFORMANCE IN DISTANCE LEARNING USING MACHINE LEARNING TECHNIQUES. Applied Artificial Intelligence, 2004. 18(5): p. 411-426.
16. Cortez, P. and A.M.G. Silva, Using data mining to predict secondary school student performance. 2008.
17. Ramaswami, M., Validating Predictive Performance of Classifier Models for Multiclass Problem in Educational Data Mining. International Journal of Computer Science Issues (IJCSI), 2014. 11(5): p. 86.
18. Yadav, S.K., B. Bharadwaj, and S. Pal, Data mining applications: A comparative study for predicting student's performance. arXiv preprint arXiv:1202.4815, 2012.
19. Al-Radaideh, Q.A., E.M. Al-Shawakfa, and M.I. Al-Najjar. Mining student data using decision trees. in International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan. 2006.

20. Dekker, G.W., M. Pechenizkiy, and J.M. Vleeshouwers, Predicting Students Drop Out: A Case Study. International Working Group on Educational Data Mining, 2009.
21. Osmanbegović, E. and M. Suljić, Data mining approach for predicting student performance. Economic Review, 2012. 10(1): p. 3-12.
22. Manhães, L.M.B., S.M.S. da Cruz, and G. Zimbrão. WAVE: an architecture for predicting dropout in undergraduate courses using EDM. in Proceedings of the 29th Annual ACM Symposium on Applied Computing. 2014. ACM.
23. Khobragade, L.P. and P. Mahadik, Students' academic failure prediction using data mining. International Journal of Advanced Research in Computer and Communication Engineering, 2015. 4(11): p. 290-298.
24. Márquez-Vera, C., C.R. Morales, and S.V. Soto, Predicting school failure and dropout by using data mining techniques. IEEE Revista Iberoamericana de Tecnologias del Aprendizaje, 2013. 8(1): p. 7-14.
25. Kabakchieva, D., Predicting student performance by using data mining methods for classification. Cybernetics and information technologies, 2013. 13(1): p. 61-72.
26. Athani, S.S., et al. Student performance predictor using multiclass support vector classification algorithm. in Signal Processing and Communication (ICSPC), 2017 International Conference on. 2017. IEEE.
27. Kyndt, E., et al., Predicting academic performance in higher education: Role of cognitive, learning and motivation. 2011.
28. Paliwal, M. and U.A. Kumar, A study of academic performance of business school graduates using neural network and statistical techniques. Expert Systems with Applications, 2009. 36(4): p. 7865-7872.
29. Şen, B., E. Uçar, and D. Delen, Predicting and analyzing secondary education placement-test scores: A data mining approach. Expert Systems with Applications, 2012. 39(10): p. 9468-9476.
30. Hall, M., et al., The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 2009. 11(1): p. 10-18.
31. Gu, Q., et al. Data mining on imbalanced data sets. in Advanced Computer Theory and Engineering, 2008. ICACTE'08. International Conference on. 2008. IEEE.
32. Hall, M.A., Correlation-based feature selection for machine learning. 1999.
33. Kullback, S. and R.A. Leibler, On information and sufficiency. The annals of mathematical statistics, 1951. 22(1): p. 79-86.
34. Tan, P.-N., Introduction to data mining. 2007: Pearson Education India.
35. Chawla, N.V., et al., SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 2002. 16: p. 321-357.
36. Olson, D.L. and D. Delen, Advanced data mining techniques. 2008: Springer Science & Business Media.
37. Han, J., J. Pei, and M. Kamber, Data mining: concepts and techniques. 2011: Elsevier.