Data Mining Approach to Predict Students' Performance

Deresse Demeke, Worku Muluye** and Tewodros Gizaw****

In recent years, the amount of data has been growing tremendously in all areas due to the rapid development of technology. The need for discovering novel and most useful information from these large amounts of data has also grown. With the advent of data mining, different mining techniques have been applied in different application domains, such as education, banking, retail sales, bioinformatics and telecommunications. The paper aims at predicting the Cumulative Grade Point Average (CGPA) of the student at the end of eighth semester based on the grade point in computer programming language and mathematics courses. It is observed from the student results of various semesters that the computer programming and mathematics courses result has correlation with the CGPA they scored. In order to predict the CGPA at the end of eighth semester, the different predictive machine learning algorithms were trained with the given dataset. Based on the result, it is observed that the success of students highly depends on the programming and mathematics course achievement. Consequently, the researchers recommend the curriculum designers to include fundamentals of computer programming in the curriculum of high schools or preparatory schools.

Keywords: Data mining, Educational data mining, Prediction of students performance, Root Mean Square Error (RMSE), Cumulative Grade Point Average (CGPA)

Introduction

In recent years, due to the rapid development of technology, the amount of data has been growing tremendously in all areas. The need for discovering novel and most useful information from these large amounts of data has also grown. With the advent of data mining, different mining techniques have been applied in different application domains, such as education, banking, retail sales, bioinformatics and telecommunications, to extract useful information to fulfill the needs of the industry. With the enormous amount of data stored in files, databases and other repositories, it is increasingly important, though not necessary, to develop a powerful means for

* Lecturer, College of Computing and Informatics, Wolkite University, Wolkite, Ethiopia; and is the corresponding author. E-mail: derese.demeke@wku.edu.et

** Lecturer, College of Computing and Informatics, Wolkite University, Wolkite, Ethiopia. E-mail: worku.muluye@wku.edu.et

*** Lecturer, College of Computing and Informatics, Wolkite University, Wolkite, Ethiopia. E-mail: tewodros.gizaw@wku.edu.et

analysis and interpretation of such data and for the extraction of interesting knowledge that could help in decision making. It is intended to obtain meaningful and valuable information that is not previously known from these data by applying data mining technique (Han, 2006).

Educational data mining is the process of applying data mining tools and techniques to analyze data in educational institutions (Al-Razgan *et al.*, 2013). Hence, the ability to predict students' academic performance is very important in educational environments. Predicting academic performance of students in case of programming domain area is challenging since the students' academic performance specially in programming domain area depends on diverse factors such as personal, socioeconomic, psychological, previous understanding about technology, the curriculum nature and other environmental variables.

Predictive data models are used to predict certain data values based on results identified from the given dataset (Sondwale, 2015). Specifically, the researchers are motivated to conduct this research when they observed the result of the students; there were programming courses' and some mathematics courses' contributions for drop out and academic dismissal of students from the university. The focus of this study is to predict students' performance based on programming and mathematics courses results.

2. Literature Review

Ahmeda (2016) focused on predicting the instructor performance and the factors that affect students' achievements. It indicates the most important factors of students' academic performance such as the previous academic records, family background, economic status and demographic data, and the prediction methods. The dataset was collected from University of California-Irvine (UCI) and from Gazi University in Ankara, Turkey that contains a total of 5,820 evaluation scores by the student. The dataset is divided into two sets where 66% is used for training and 34% is used for testing. The attributes were course-specific questions, the class label, the level of attendance values and level of difficulty of the course. The dataset is tested and analyzed using four data classifier such as J48 decision tree, multilayer perception, Naïve Bayes and sequential minimal optimization. The J48 algorithm achieves the best performance with an accuracy of 84.8%. The feedback given by students for specific course can be considered as an attribute which affects the student's success. Furthermore, to improve educational quality the researchers used CGPA and internal assessment attributes to predict students' performance and removing the worst ranked attributes on dataset increased the algorithms accuracies.

Shahiria (2015) provided an overview of the data mining techniques that are used to predict students' academic performance and the prediction algorithm to identify the most important attributes. The most important factors in predicting students' performance are attributes and prediction methods. CGPA and internal assessment (assignment mark, quizzes, lab work, class test and attendance) were the most

frequently used attributes. Next, the most often used attributes were student's demographic (gender, age, family background and disability) and external assessments (final exam result for a particular subject). The three attributes mostly used in predicting students' performance were extracurricular activities, high school background and social interaction network. Also, psychometric factors such as student interest, study behavior, engage time and family support attributes were used in predicting student's performance.

In educational data mining, predictive modeling is used to predict students' performance. Finally, the authors concluded that CGPA and internal assessment were more frequently used dataset, the classification method was frequently used for prediction techniques, and neural network and decision tree were the two methods highly used by the researchers for predicting students' performance (Shahiria, 2015).

Dina and AlHammadi (2013) reported data mining-based completed research in the higher education sector such as colleges and universities. As data mining becomes a great asset to the educational system, many researchers use data mining to enhance the quality of the education system. The authors studied three different cases, viz., student's level, academic guidance level and decision makers level to show the role of data mining and to enhance the quality of higher education system. Finally, they concluded that data mining is an essential tool in all levels of educational systems. And they presented an effective data mining cycle to clarify data mining process.

Owolabi *et al.* (2014) presented the relationship between mathematics and computer programming courses. The authors identified six variables, viz., mathematics ability, mathematics anxiety, computer anxiety, programming anxiety, age and gender and achievement in basic programming. In their correlational design, mathematics ability, mathematics anxiety, computer anxiety, programming anxiety and age and gender serve as the independent variables and achievement in basic programming as the dependent variable. They determined the relationship between six variables, the composite effect of the six predictor variables and the relative effect of the six predictor variables.

Owolabi *et al*. (2014) conducted a study at the Federal colleges of education in Lagos and Ogun states of Nigeria. The authors used computer anxiety rating scale, programming anxiety rating scale and mathematics anxiety rating scale instruments. The authors also used Algebra course to determine the mathematics ability of students and the basic programming course to determine student's achievement in basic programming. The relationship between mathematics ability and the performance of computer/mathematics students is statistically significant in basic programming. The data were analyzed using multiple regression analyses and Pearson product moment correlation coefficient. The merged effect of the six predictor variables on the performance of computer/mathematics students in basic programming was 20.8%. And mathematics ability variable has a significant relative effect on the performance of computer/mathematics students in basic programming. Finally, in their recommendation college of computer science, students should be motivated to improve their mathematics

ability and develop a more positive attitude towards mathematics courses, because good mathematics ability improves good academic achievement in computer programming (Owolabi *et al.*, 2014).

3. Methodology

The objective of this study to predict students' performance based on the result of programming and mathematics course results. In the following sections, the researchers present the source of dataset, the attributes selected for the study and tools used to for data preprocessing, feature extraction and analysis.

3.1 Data Source

For the experiment purpose, the researchers had collected necessary data related to regular students from three Ethiopian public universities specifically from college of computing and informatics. Within each university, they do have some departments, namely, information technology, computer science, information systems and software engineering, under the college.

3.2 Data Preprocessing

One of the important steps of data mining process is data preprocessing. Data preprocessing is used in identifying the missing values, noisy data and irrelevant and redundant information from dataset. After collecting the data from three universities, the researchers stripped a two-batch student mathematics and programming course results from their PDF grade report. Then identified the common mathematics and programming courses for Information Systems, Computer Science and Information Technology. Next to this, the three-department student mathematics and programming course results were organized and merged in CSV format.

A total of 341 records were taken for the analysis. Attribute selection was used to find out the best attributes in the data. In the data, two attributes, i.e., student's semester cumulative GPA and their name, were removed. Then select attributes were used to rank the attributes. Attribute selection involves searching through all possible combination of attributes in the data to find which subset of attributes works best for prediction. To do this, two objects must be set up: an attribute evaluator and a search method. After eliminating incomplete data, the sample comprised 341 students who graduated in 2012-13. The model of students' success was created, where success as the output variable was measured with the success in the CGPA at the end of last semester before graduation. As input to the model, 13 variables were used.

RapidMiner has a built-in tool which is used to preprocess and attribute selection. Using preprocessing module of RapidMiner, the dataset was prepared by merging the records of 410 students who graduated in 2012-13. After removing the incomplete data, the dataset contained 341 rows having 12 independent variables and one target attribute. Its names and description are shown in Table 1.

Table 1: List of Attributes and Target Variables		
S. No.	Attributes	Description
1.	Name	Full Name of the Student
2.	Sex	Gender of the Student
3.	ID	Identification Number of the Student
4.	INSY1021	Fundamentals of Programming-I
5.	Stat	Introduction to Statistics
6.	INSY1022	Fundamentals of Programming-II
7.	INSY1023	Discrete Mathematics and Combinatory
8.	INSY2031	Object-Oriented Programming
9.	INSY2036	Data Structure and Analysis
10.	INSY2033	Event-Driven Programming
11.	INSY3081	Introduction to Internet Programming
12.	INSY2082	Advanced Internet Programming
13.	CGPA	Cumulative Grade Point Average (Target Variable)

4. Results and Discussion

The student dataset were analyzed using RapidMiner. As cited by Bermudez (2013), RapidMiner is an open-source toolbox for data mining, machine learning and statistical methods. Started in 2001 at the University of Dortmund, this tool has been used in various disciplines such as education, business applications, health informatics and research. RapidMiner provides a wide variety of methods such as statistical measures like correlation or machine learning modeling like classification, regression and clustering algorithms.

RapidMiner studio offers a collection of machine learning and data mining algorithms as well as tools for data preprocessing and visualization. The four common predictive algorithms, namely, Generalized Linear Model, Deep Learning, Decision Tree and Support Vector Machine, are selected to develop the model.

4.1 Regression (General Linear Function)

Regression is a data mining (machine learning) technique used to fit an equation to a dataset. The simplest form of regression, linear regression, uses the formula of a straight line ($y = mx + b$) and determines the appropriate values for m and b to predict the value of y based upon a given value of x. Advanced techniques, such as multiple

regression allows the use of more than one input variable and allows for the fitting of more complex models, such as a quadratic equation (Sahu *et al*, 2013).

4.2 Deep Learning

Deep learning or Artificial Neural Networks (ANN), usually called Neural Networks (NN), is a class of mathematical models that attempts to replicate the structure and processing capabilities of biological neural networks. First discussed in 1943, NN are a group of interconnected nodes or artificial neurons, often organized into layers in which the information is processed using a connectionist approach, i.e., that mental and behavioral perceptions can be described by interconnected networks of simple units. The global behavior of the processing units is determined by the strength (or weights) of these connections. The structure of the network is changed during the learning step modifying the strength of the connections in order to find the desired conditions or data flow.

In an NN, the researchers will have a set of input and output nodes and a group of processing nodes (hidden layer). The connections between these nodes determine the way in which the information will flow, and the strength of these connections (or weights) determines how important the connections are. It is a set of simple processing elements in which global behavior is determined by the connections (weights) and their parameters (Bermúdez, 2013).

According to Soares and Souza (2016), basically, an NN can have different layouts, depending on how the neurons or neuron layers are connected to each other. Every NN architecture is designed for a specific end. NNs can be applied to a number of problems, and depending upon the nature of the problem, the NN should be designed in order to address this problem more efficiently. Basically, there are two modalities of architectures for NNs; based on neuron connections, they can be monolayer networks or multilayer networks, and based on signal flow, it can be classified as feedforward networks or feedback networks.

RapidMiner tool uses multilayer feedforward ANN model to predict the outcomes based on given model.

4.3 Decision Tree

Decision Tree is one of the popular techniques for prediction. Most of the researchers have used this technique because of its simplicity and comprehensibility to uncover small or large data structure and predict the value. The decision tree models are easily understood because of their reasoning process and can be directly converted into set of IF-THEN rules. The students' performance evaluation is based on features extracted from an educational institution. The examples of dataset are student's grades obtained in mathematics course, grades obtained in programming course and final Cumulative Grade Point Average (CGPA). All these datasets were studied and analyzed to find out the main attributes or factors that may affect the students' performance. Then a suitable data mining algorithm will be investigated to predict students' performance (Shahiria, 2015).

4.4 Support Vector Machine

Support Vector Machine is a supervised learning method used for classification. Different researchers have used this method to predict students' performance. This method is considered as their prediction technique because it suited well in small datasets and has a good generalization ability and is faster than other methods. It has acquired the highest prediction accuracy in identifying students at risk of failing (Shahiria, 2015).

4.5 Model Accuracy

A comparison of performance of the models based on tenfold cross-validation is provided in Table 2. Numeric predictive model performance can be measured Root Mean Square Error (RMSE). RMSE is the square root of the mean of the squared errors. RMSE indicates how close the predicted values are to the actual values; hence, a lower RMSE value signifies that the model performance is good. One of the key properties of RMSE is that the unit will be the same as the target variable (Swamynathan, 2017).

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where y_i is the actual value and $\hat{y}$ is the predicted value.

Equation 1: Numeric predictive model performance

Table 2: Predictive Model Performance

Model	Root Mean Squared Error
Generalized Linear Model	0.173+/–0.018
Deep Learning	0.192+/–0.023
Decision Tree	0.226+/–0.015
Support Vector Machine	0.178+/–0.018

Conclusion

In this study, the researchers used deep learning, general linear regression model, support vector machine and decision tree algorithms to predict student's cumulative grade point average. This work may improve student's performance; reduce failing ratio by taking appropriate steps at right time to improve the quality of education.

The researchers concluded that programming and some mathematics course regardless of their sex highly affect the success of the regular students in Ethiopian public university specifically different departments in the college of informatics. Hence, the programming courses determine the success rate of the students. As

many researchers argue, lack of prior experience for computer programming language negatively affects the result of students in university. To the researchers' best of knowledge, the national curriculum of Ethiopia did not include computer programming language course in curriculum of secondary and preparatory schools.

For future work, the researchers hope to refine the technique in order to get more valuable and accurate outputs, which can be used to exactly predict the students' performance by including other factors.

Recommendation: Finally, the researchers suggest curriculum designers to include computer programming course in high school and preparatory school curriculum in order to enhance the quality of education and to increase the success rate of students at university level. On the other hand, the researchers recommend Ethiopian universities to give special attention for programming and mathematics courses.

References

1. Ahmeda Ahmed Mohamed A R A H U (2016), "Using Data Mining to Predict Instructor Performance", 12th International Conference on Application of Fuzzy Systems and Soft Computing, ICAFS, pp. 137-142, August 30.
2. Al-Razgan M, Al-Khalifa A and Al-Khalifa H (2013), "Educational Data Mining: A Systematic Review of the Published Literature 2006-2013", in Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013).
3. Bermúdez Y S (2013), *How to Select the Right Machine Learning Approach?*, Linnaeus University, Sweden.
4. Dina M S A and AlHammadi A Aziz (2013), "Data Mining in Higher Education", *Periodicals of Engineering and Natural Sciences*, Vol. 1, No. 2, pp. 1-4.
5. Han J K M (2006), *Data Mining: Concepts and Techniques,* Morgan Kaufmann Publishers.
6. Owolabi J, Olanipekun P and Iwerima J (2014), "Mathematics Ability and Anxiety, Computer and Programming Anxieties, Age and Gender as Determinants of Achievement in Basic Programming", *GSTF International Journal on Computing*, Vol. 3, No. 4, pp. 109-113, May 23.
7. Sahu Hemlata, Sharma S and Gondhalakar S (2013), "A Brief Overview on Data Mining Survey", *International Journal of Computer Technology and Electronics Engineering (IJCTEE)*, Vol. 1, No. 3, pp. 114-121.
8. Shahiria Amirah Mohamed W H N A R (2015), "A Review on Predicting Student's Performance Using Data Mining Techniques", The Third Information Systems International Conference, pp. 414-422.

9. Soares Fábio M and Souza Alan M F (2016), *Neural Network Programming with Java*, Packt Publishing, Birmingham B3 2PB, UK.

10. Sondwale Pradnya P (2015), "Overview of Predictive and Descriptive Data Mining Techniques", *International Journal of Advanced Research in Computer Science and Software Engineering,* Vol. 5, No. 4, pp. 262-265.

11. Swamynathan M (2017), *A Practical Implementation Guide to Predictive Data Analytics Using Python*, Apress, Bangalore, Karnataka, India.

Reference # 56J-2020-10-02-01