www.ijecs.in

International Journal Of Engineering And Computer Science
Volume 9 Issue 10 October 2020, Page No. 25240-25245
ISSN: 2319-7242 DOI: 10.18535/ijecs/v9i10.4537

Data Mining Model to Predict the Algorithm Development Skills of Computer Science Students: In case of Wolkite University

Melaku Kore

Lecturer, Department of Computer Science, Wolkite University, Ethiopia

Abstract

Performance prediction of students in university courses is a crucial concern to all higher education institutions managements where several factors may affect their students' performance. In Computer Science related programs, programming courses and having the skills of algorithm analysis are made compulsory subjects in most institutions of learning. Having mathematical skills play a vital role to be a good programmer and algorithm analysis. In this study an attempt was made to build a model which predict the performance of students in the Algorithm Analysis based on their grades of other courses, specifically by evaluating preliminary programming courses and different mathematics courses of the curricula. The data was collected from Wolkite University, Computing and Informatics College of Computer Science under graduate students of 2015-2018 batch students. To build the model classification algorithms, namely, Decision tree, Naïve Bayes and Support Vector Machine (SMO) were applied using Weka machine learning tool. The experiment from the study shows that a priori knowledge of the Data structure and Algorithm and Fundamentals of Programming I are essential to excel in the course Algorithm Analysis for computer science students. This work will be of considerable importance in identifying students who has low performance on this courses and the students will be advised to take action on their skill gap in timely manner. Finally, the classification algorithms were tested and performances comparisons were made. The decision tree (J48) algorithm provides a promising result of 78.313%.

Keywords: Data Mining, Classification, Decision tree, Data Structure and Algorithm, Performance, K-Fold Cross Validation,

1. Introduction

The production of qualified students is one of the major objectives of higher education institutions. To be a good programmer, computer science students are expected to master their algorithm development and program coding skills. Having good skills of mathematics play a crucial role to analyze and perform computer programming problems. In computer science, algorithms analysis mean the process of finding the amount of time, storage, or other resources needed to execute a given problem Algorithm analysis helps to understand the nature of the problem in deeper and suggest possible solutions.

No doubt that, Algorithm analysis is difficult and challenging course for Computer Science students. The big question is that knowing the responsible science subjects directly affect the proficiency of Algorithms analysis. These students' gap can be addressed by analyzing their academic records using different data mining techniques. Organizations and institutions uses data mining for their current reporting capabilities, to uncover and understand hidden patterns in vast databases [1].

Data mining is one of the most cardinal areas in recent technologies for retrieving valid information from huge amount of unstructured and distributed data using parallel processing of data [3].Applying the technology of data mining has been used in educational filed to discover the hidden information from educational data sets. To predict students' performance there are different known classification algorithms such as Decision tree algorithms, K-neatest neighbor, Support vector machine (SMO), Artificial neural network (ANN), Rule induction algorithms and soon. In this study Decision tree), Naïve Bayes and

Support Vector Machine (SMO) were applied using Weka machine learning tool.

The main objectives of the study was predicting the algorithm development skills of computer science students of Wolkite University for the year 2015 to 2018. Here, based on the prediction of the model courses which are highly affect algorithm analysis proficiency of students were recognized.

The rest part of the study is organized as follows: Section 2 presents related conducted research works, section 3 describes the proposed data mining classification algorithms to predict the students' proficiency on the course Algorithm analysis, on section 4 experiment part of the study is discussed and finally the conclusion of the study is presented on section 5.

2. Related Works

Before this study different research works related to this study were conducted. While the researchers do their work different data mining classification, clustering or association rules. In this section some of related works are mentioned.

In [4], different data mining classification techniques were applied to predict the performance of students in the case of private University in Northern part of Nigeria. The researchers used WEKA as experimental tool and decision tree algorithms. To evaluate the classification technique 10 folded classification approach was applied and finally the researchers made a comparison of the selected decision tree algorithms, ID3, C4.5 and CART. They found that a promising result in all algorithms but they put that C4.5 gave relatively better accuracy than other algorithms.

As discussed on [7], students' final grade can be predicted by analyzing their previous grades. The researchers use grades for major courses of each semester and predict the final grades of students in the case of King Saud University female computer science students of 2013 batch. Decision tree classification was the model they have used and they put that the experiment will be extended on ANN as their future work.

A research work conducted by [9], students' academic performance prediction was applied based on their academic records. The researchers used two different data set to examine the prediction. The first data set was student's placement data from MOE and the second data set was Dilla University, Horticulture students; data of the year 2009 –2014. Data integration was applied on the different data sets and they used Rapid Data miner tool for the preprocessing, processing and experiment part of the study. The result show that around 27 rules were generated from the model they built.

Prediction and analysis of student performance is an important milestone in educational environment [11]. The researchers on this study try to identify factors associated with students whose academic status is not good and to improve the quality of education by detecting slow learners. While those students whose academic performance is low has been identified the teachers can assist them and the quality of education can be improved. On the study, the accuracy of five data mining classification techniques, Multilayer Perception, Naïve Bayes, SMO, J48 and REPTree were investigated and multilayer perception (75%) gives better result.

The researchers in [14], conducted a classification techniques of Engineering Students of Purvanchal University, Jaunpur (Uttar Pradesh) 2010 session students. The objective of the study was classifying students' performance on the final exam and predict the outcome either pass, promoted to the next year or fail on that semester. To do, so different decision tree algorithms were used and the C4.5 algorithms reveals better performance (67.7778%) than the others.

3. Research Methodology

To build the proposed classification model five steps are followed, namely (1) problem formulation, (2) data preparation which includes collection of raw data, data cleaning, data selection and transformation, (3) apply classification algorithms, (4) result obtained, (5) performance evaluation and conclusion.

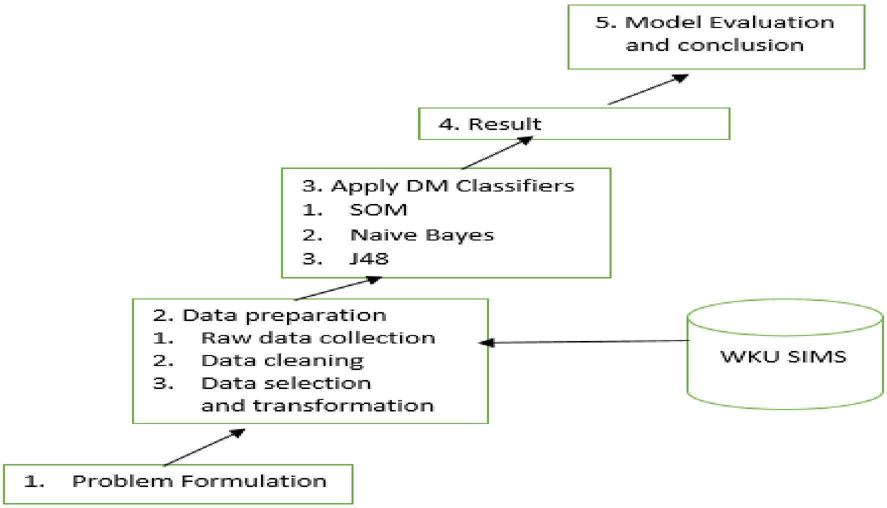
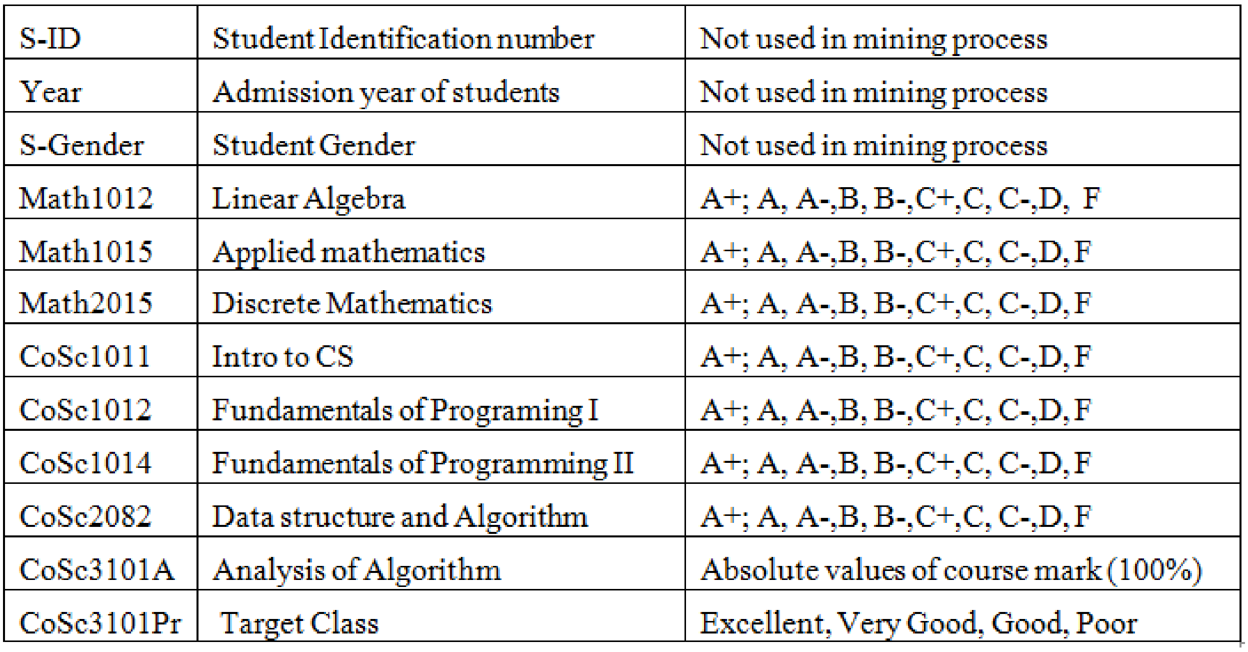


Figure 1: The steps of methodology

3.1 Data Preparation

In data mining classification techniques it's clear that the performance of the classifier depends on the data you prepared. Therefore, the data set should be organized in a well manner. In this study a five year computer science students (2013-2018) of Wolkite University was collected. The data consists of 275 males and 143 females, a total of 418 students. While the data was exported from Student Information Management System (SIMS) the format was in excel format and it consists of 17 attributes. Intact, all attributes are not relevant to predict the performance of students' algorithm development proficiency. For this study 8 determinant variables and 1 class (target) variable were selected. Here, basic programming courses and different mathematics courses were selected as major determinant attributes by applying attribute selection method. The data set was converted CSV format and students who were dropout, dismissed before taking all those selected courses were removed manually from the file.

Table 1: Lists of Student related variable



S-ID	Student Identification number	Not used in mining process
Year	Admission year of students	Not used in mining process
S-Gender	Student Gender	Not used in mining process
Math1012	Linear Algebra	A+; A, A-,B, B-,C+,C, C-,D, F
Math1015	Applied mathematics	A+; A, A-,B, B-,C+,C, C-,D, F
Math2015	Discrete Mathematics	A+; A, A-,B, B-,C+,C, C-,D, F
CoSc1011	Intro to CS	A+; A, A-,B, B-,C+,C, C-,D, F
CoSc1012	Fundamentals of Programing I	A+; A, A-,B, B-,C+,C, C-,D, F
CoSc1014	Fundamentals of Programming II	A+; A, A-,B, B-,C+,C, C-,D, F
CoSc2082	Data structure and Algorithm	A+; A, A-,B, B-,C+,C, C-,D, F
CoSc3101A	Analysis of Algorithm	Absolute values of course mark (100%)
CoSc3101Pr	Target Class	Excellent, Very Good, Good, Poor

3.2 Building The Classification Model

In this section the proposed classification models were discussed. A classifier makes use of a learning algorithm to find a model that best defines the relationship between the attributes and the class label of the training dataset [10]. In machine learning process the performance of the model is evaluated by counting the number of correctly and incorrectly interpreted instances. There are many classifiers and one cannot be the best in all cases, time and result. The variables that are used to determine the predicted variable are called independent variables while the independent variable is called class variable. Here, in our study we used the ranker method of attribute selection and we had took the top 7 variables from 17 attributes. The target variable was CoSc310Pr. The predicted classes were Excellent, Very Good, Good and Poor. Students who scored 80%-100%, 65%-79%, 50%-64% and below 50% on the course Analysis of Algorithm are labeled as Excellent, Very Good, Good and Poor class respectively.

For this study, the data set was tested with three different classification algorithms: Support Vector Machine, Decision tree (J48) and Naive Bayes. The attributes CoSc1012, CoSc2082 and Math2015 were the top attributes that affect the target class. To process the classification 10 Folded Cross Validation was applied on the data set. In 10-fold cross validation, the complete dataset is randomly split into 10 mutually exclusive subsets of approximately equal size. The classification model is trained and tested 10 times. Each time it is trained on nine folds and tested on the remaining single fold. According to [1], 10-fold cross validation does not require more data compared to the traditional single split (2/3 training, 1/3 testing) experimentation.

As the name implies Decision tree are commonly used for gaining information for the purpose of decision making. The prediction of data objects using decision tree classification techniques is simple and one can understand the generated rules easily. There are many decision tree algorithms like, C4.5, ID3, CART, J48 etc. For this study, the J48 algorithm was used. After completing the J48 process the decision tree was visualized and different rules were generated. The attribute CoSc2082 (Data structure and Algorithm) was the root node. On the data set 78.313% was correctly classified using the J48 method.

Support Vector Machine (SOM), is a supervised learning algorithm which is used for classification and regression purpose. In this context, the algorithm is used for classification purpose. It's observed that 73.567% of the data sets are correctly classified in this study.

NaiveBayes is a probabilistic learning algorithm used for classification problems. Bayes Theorem is applicable for this algorithm. On the study the result shows that 68.671% of the data set was classified correctly.

4. Result and Discussion

This section describes the experiment output of the study such as the distribution of students on the predicted variable (target class), performance comparison of the algorithms and the correlation value of the determinant attributes. The result shows that from the data taken for this case study around 50% of the students have Good performance in the course Analysis of Algorithm.

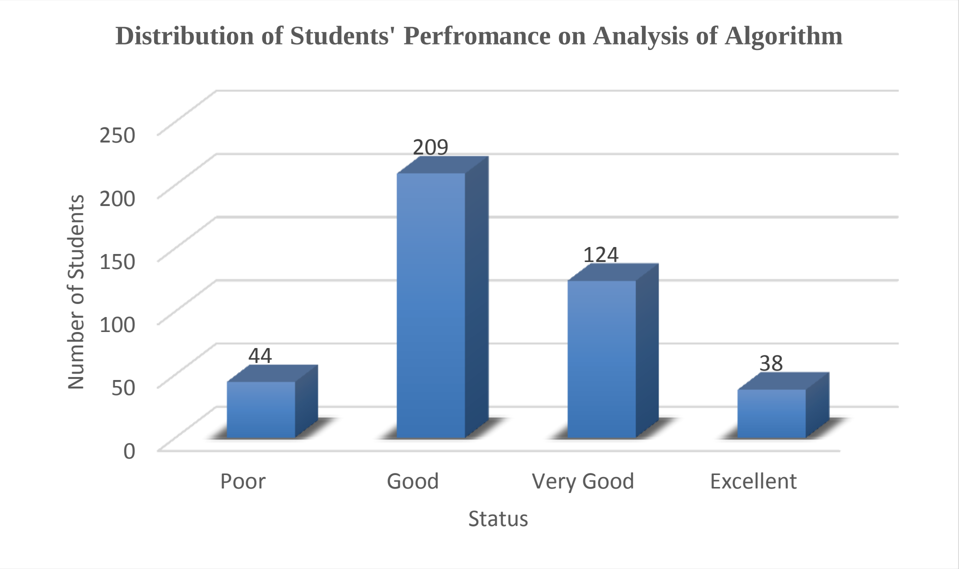


Fig 2: Students' Distribution on Analysis of Algorithm

Correlation Attribute Eval evaluator with ranker method was applied to select the highest attributes which affect the prediction class. From the result observed, we found that the variables CoSc2082 (Data structure and Algorithm Analysis), CoSc102 (Fundamentals of Programing I and Math2015 (Discrete Mathematics) were the top three determinant variables.

Table 2: Correlation Value

No	Variable Name	%distribution
1	CoSc2082	26.67
2	CoSc1012	18.73
3	Math2015	14.33

Three experiments were conducted to identify the best datamining classification algorithm. During the experiment the performance of each selected algorithms were observed. When the result is examined one can be best in execution time and the other on the accuracy result. Finally, the result shows that the decision tree (J48) was the best algorithm. The following table presents the accuracy of the algorithms with execution time.

Table 3: Classifiers Accuracy

	Algorithm	Accuracy (%)	Execution Time
1	J48	78.313	0.4
2	SVM	73.567	1.03
3	NaiveBayes	68.671	0.09

Conclusion

On this study a data mining classification methods were applied to predict the performance of Algorithm development skills of computer science students in the case of Wolkite University. This study employed the use of Decision tree, Support Vector Machine and NaiveBayes data mining classification techniques to predict the performance of students in algorithm development. The study reveals that background knowledge of Data Structure and Algorithm, Fundamentals of Programming I followed by Discrete mathematics and Combinatorics are essential to becoming a good algorithm developer.

The result show that from the collected data set the distribution of students having poor performance is almost similar to excellent performance and almost half of students data record shows good in algorithm development. The Decision tree (J48) algorithm reveals a promising result to predict the performance of the students. In the future work by using large data set, we can increase the performance of the algorithms and identify additional basic courses which affects students' algorithm development proficiency of students'.

Reference

[1.] D. David L. Olson and Dursun, "Advanced Data Mining Techniques, Springer-Verlag: Berlin Heidelberg", 2008.

[2.] Jing Luan (2006). Data Mining Applications in Higher Education, SPSS,www.spss.com/ Downloaded in July 2020.

[3.] Pratiyush Guleria, Manu Sood, "Big Data Anlalytics: Predicting Academic Course Preference Using Hadoop Inspired MapReduce", IEEE, 2017.

[4.] Y. K. Saheed, T. O. Oladele, A. O. Akanni and W. M. Ibrahim. "STUDENT PERFORMANCE PREDICTION BASED ON DATA MINING CLASSIFICATION TECHNIQUES" 2018.

[5.] Anoopkumar M, A. M. J. Md. Zubair Rahman," Model of Tuned J48 Classification and Analysis of Performance Prediction in Educational Data Mining", 2018.

[6.] Lakshmipriya. K, Dr. Arunesh P.K, "PREDICTING STUDENT PERFORMANCE USING DATA MINING CLASSIFICATION TECHNIQUES", 2017.

[7.] Mashael A. Al-Barrak and Muna Al-Razgan,"Predicting Students Final GPA Using Decision Trees: A Case Study", 2016.

[8.] Ghada Badr, Afnan Algobail, Hanadi Almutairi, Manal Almutery." Predicting Students' Performance in University Courses: A Case Study and Tool in KSU Mathematics Department", 2016.

[9.] Fiseha Berhanu, Addisalem Abera, "Students' Performance Prediction based on their Academic Record", 2015.

[10.] Raheela Asif, "Predicting Student Academic Performance at Degree Level: A Case Study", 2015.

[11.] Parneet Kaur, Manpreet Singh, Gurpreet Singh Josan," Classification and prediction based data mining algorithms to predict slow learners in education sector", 2015.

[12.] V.Ramesh, P.Parkavi. K.Ramar, "Predicting Student Performance: A Statistical and Data Mining Approach", 2013.

[13.] Surjeet Kumar Yadav, Saurabh Pal,"Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification", 2012.

[14.] O.S. Akinola, B.O. Akinkunmi, T.S. Alo, "A Data Mining Model for Predicting Computer Programming Proficiency of Computer Science Undergraduate Students", 2012.

[15.] N. M. Norwawi, S. F. Abdusalam, C. F. Hibadullah, B. M. Shuaibu "Classification of Students' Performance in Computer Programming Course According to Learning Style", 2009.

[16.] Qasem A. Al-Radaideh, Emad M. Al-Shawakfa, and Mustafa I. Al-Najjar, "Mining Student Data Using Decision Trees", 2006.