Volume 7, Issue 2, February 2017 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Student Performance Prediction Model using Machine Learning Approach: The Case of Wolkite University

Ermiyas Birihanu Belachew
Lecturer, Software Engineering, Wolkite University, Ethiopia

Feidu Akmel Gobena
Lecture, Computer Sciences, Wolkite University, Ethiopia

DOI: 10.23956/ijarcsse/V7I2/01219

Abstract- A high prediction accuracy of the students' performance is helpful to identify the low performance students at the beginning of the learning process. Machine learning is used to attain this objective. Machine learning techniques are used to discover models or patterns of data, and it is helpful in the decision-making. The ability to predict performance of students is very crucial in our present education system. We applied Machine learning concepts for this study. The dataset used in our study is taken from the Wolkite university registries office for college of computing and informatics from 2004 up to 2007 E.C with respect to each department. In this study, we have been collected student's transcript data that included their final GPA and their grades in all courses. After pre-processing the data, we applied the machine learning methods, neural networks, Naive Bayesian and Support Vector Machine (SMO). Finally, we built the model for each method, evaluate the performance and compare the results of each model. Using machine learning, the aim was to develop a model which can derive the conclusion on students' academic success.

Keywords- Classification, Machine Learning, Higher Education, Prediction, Student Success

I. INTRODUCTION

For higher education institutions whose goal is to contribute to the improvement of quality of higher education. The quality of higher education institutions implies providing the services, which most likely meet the needs of students, academic staff, and other participants in the education system.

Tekeste writes "The golden age of modern education in Ethiopia" is usually dated to the years between 1941 and 1970 (the regime of HIM Hailesellassie). Education was free and it applied more to the poorer section of the population; the rich and the aristocracy were less enticed by the economic returns of education [1][7].

Currently, the Ethiopian Government gives higher education a central position in its strategy for social and economic development. Ethiopia has radically expanded the numbers of its higher education institutions: from two Federal universities to 33; among this 10 of them are opened before 5 years and one of this is Wolkite University.

Nowadays, the data base that store data and information for organization becomes complicated and difficult to analysis [2]; for this case we are going to apply Machine Learning techniques to resolve those problems. Wolkite University has its own student management information system that was developed by Bahir Dar University Course and Curriculum Management System. However, this database contains so much data that it becomes almost impossible to manually analyze them for valuable decision-making information. In order to analysis this complex data base we can able to use machine learning techniques. This Study conducted in 993students from college of computing and informatics within Wolkite University with respective departments. We were using WEKA open source software to test the prediction of the student performance. It provides many different algorithms for data mining and machine learning. WEKA is open source and freely available. It is also platform-independent[3] .We may have various factors for education with in Wolkite University such as environment, family standard of each student, gender, teacher's educational background and education policy[1][4][5], but our research is not going through each factor because it is physiological factor instead of learning once.

This study has the following contributions.

- Provide an overview of existing student's educational standards.
- Provide an overview of the future action by the management of the university and students.
- It becomes an initial framework for future study in the university

The rest of this paper is organized as follows: Section II summaries related work in an application of prediction of students' performance by data mining and machine learning techniques in educational environments; Section III explains the research methodology and machine learning methods used. Section IV discusses experimental results. Finally, we conclude this paper with a summary and describe an outlook for future work.

II. REVIEW OF THE RELATED RESEARCH

The main objective of any higher educational institution is to improve the quality of managerial decisions and to impart quality of education. Good prediction of student's success in higher learning institution is one way to reach the highest level of quality in higher education systems[6] .

Higher Education in Ethiopia includes undergraduate degrees offered for three, four or more years and specialization degrees such as Masters and PHD programs. There is different experiment under took for predicating students' performance in different world universities; here we reviewed some of the paper that are published related with student academic performance[7].

In this study, they have used the data mining techniques first semester Bachelor of Computer Science from University Sultan Zainal Abidin by using three selected classification methods; Naïve Bayes, Rule Based, and Decision Tree; finally, they conclude that from the experiment, the models develop using Rule Based (68.8 %) and Decision Tree algorithm (68.8%) shows the best result compared to the model develop from the Naïve Bayes algorithm (63.3%) from the second experiment[2].

In this experiment, proposed to illustrate how data mining can be used in educational context and predicting the student success or failure in courses. They also model the system architecture what would like, set up the experiment and got the result. The system architecture includes three distinct processes. In the first process, variables are selected and extracted from academic data and gathered in a data set for each course in a format suitable for classifier training. Finally, the researcher conclude that they are using the decision tree classification algorithm (C 5.0) draw conclusion on predication of student performance using data mining techniques from a set of small experiments on academic data base[8].

They proposed to predicate the student academic performance by giving a student to write their comment on the space provided after each lesson then extracting words and speech frequencies and use the LSA technique to reduce the dimensions of a matrix and obtain the most significant vectors. Finally, the researcher conclude that this study expressed the correlation between self-evaluation descriptive sentence written by students and their academic performance by predicting their grade; but what happen if the student does not express the exact filling of themselves. The academic performance of student in education can be affected by inside or outside factor. This can be individual and house hold characteristics, socio-economic situation, school related factor and government policies. According to this study in addition to the above factor there is also poor preparation and commitment, mismatch area of interest and field of placement, poor social integration and lack of appropriately developed instruction and assessment method for deterministic of achieving in cumulative grade point average and cause of student dropout or persistence[6].

The study was done in south Wollo, Ethiopia particularly Dessie; around 13 schools among this 7 of them are government and 6 privates schools and proposed that Gender is among the determinant factors affecting students' academic achievement was the initial idea of the researcher and finally he concluded that female students obtained slightly higher score than the males, but the difference was not statistically significant (P>0.5)[5].

The study conduct to investigate motivation have factor for academic performance in university of Gurjrat and the data are collected from male and female's students from semester 2 and 4 at different disciplines by giving the following questioners with a different category (strongly agree, agree, neutral, disagree) to chosen by those students. Finally, the result shows that males student was more motivated than females[1].

III. THE RESEARCH METHODOLOGY

Educational data mining is an interesting research area which extracts useful, previously unknown patterns from educational database for better understanding, improved educational performance and assessment of the student learning process[9]. In this study, we were used education data mining to extract the needed data from WKU data base and apply the Machine Learning methods especially classification and clustering algorithms such as Neural Nets (MLP), Naive Bayesian and Support Vector Machine.

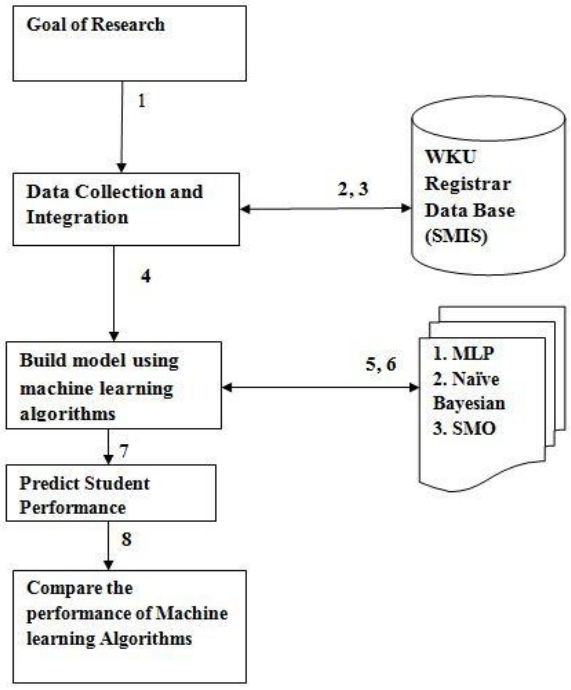

Fig 1**:** Research Process for This Study

A. Goal of the Research;

The main objective of this study were predicate the student performance particularly in college of computing student at Wolkite University based on the course the student took and GPA.

B. Data collection and integration (Academic data base)

WKU registrar data base were included the seven collages with its departments; and our research focus were on college of computing and informatics, because we are near for this collage when comparing with others. Data are gathered from four departments starting from 2004 up to 2007 E.C. At 2004 and 2005 there are not any students who are joined in software engineering department. The data collected from each department at 2004 academic year were 35, 25 and 31 students with respective of Computer Sciences, Information System and Information Technology. The data gathered at 2005 academic year were 34, 36, 32 students with respective computer Sciences, Information System and Information Technology. The data collected at 2006were 70, 93, 117, 79 students with respective computer sciences, Information System, Information technology and Software Engineering. The data collected at 2007 were147, 132, 143, 97 students with respective computer sciences, Information System, information Technology and Software Engineering.

Courses includes both Major and Common courses. Semester CGPA involves the result for the semester with each student CGPA. Courses that are categorizing under course one to course six are operating system, Advanced Programming, data communication and computer network, Microprocessor and Assembly Language programming, Advanced Database System, Discrete Mathematics and Combinatory, Computer Maintenance, introduction to Computer Sciences, Civics and Ethical education, Fundamental of Electronics, Liner Algebra and communication Skill. The data collected would be processed for WEKA as CSV file format as shown in Table 1 attributes description.

Table 1:Attribute Description

Variables	**Descriptions**	**Possible values**
Course 1-Course 6	Course offered in departments within each semester	A+,A-,A,B+,B,B-,C+,C,C-,D,F
Semester GPA	Result of students in current semester	1.00-4.00
Previous GPA	Result of student in the previous semester	1.00-4.00
CGPA	Sum of current with previous semester divided by	1.00-4.00
Status	The student class	First class with great distinction, First class with distinction, First Class, Second Class, Academic warning

C. Machine learning Tool and Algorithm Used

For the purposes of this study WEKA software package was used, that was developed at the University of Waikato in New Zealand. This research work has carried out experiments in order to evaluate the performance and usefulness of different classification algorithms for predicting the student performance in college of computing and informatics at WKU. The data set used for the research work is pre-processed in order to transform them into a suitable format to be used by the prediction tool. Tests were conducted using three algorithms tests for the assessment of input variables: Neural Net (MLP), Naive Bayesian and Support Vector Machine.

- **Neural Net (MLP): -** it is a feed forward network with possibly several hidden layers, one input layers and one output layer, totally interconnected.
- **Naive Bayesian:** -This classifier represents the promising approach to the probabilistic discovery of knowledge, and it provides a very efficient algorithm for data classification.
- **Support Vector Machine:** -SVMs, being computationally powerful tools for supervised learning, are widely used in classification [10, 11].

IV. RESEARCH RESULT

Three experiments were conducted for this study and for all experiments four scenarios were considered (CS, IS, IT and SWE), data base containing 34 attributes and the experiment containing 11 selected attributes for the purpose of this study.

- **Experiment one**

The first experiment was designed to evaluate the performance of a MLP classifier for computer science, Information System and Information Technology student at 2004, 2005, 2006 and 2007 entry in predicting student performance and to investigate the performance of the model and for software engineering students at 2006 and 2007 entry.

After conducting the first experiment the model has performed with 80.6% in 2004 academic year IT students and the algorithm takes 1.85 second. The worst result scored in experiment one was 45.7% in 2004 computer science students with execution time 1.62 second.

- **Experiment Two**

This experiment was designed to evaluate the performance of Naïve Bayesian classifiers for computer Science, information System and Information Technology student at 2004, 2005, 2006 and 2007 entry and for software engineering students at 2006 and 2007 entry.

During experiment two the Naïve Bayesian classifiers model achieved 95.7% with execution time of 0.01 second. The worst result scored in experiment two was 75 % in 2005 IT students with execution time 0 second.

- **Experiment Three**

In this experiment the performance of Support Vector Machine (SMO) will be evaluated for computer Science, information System and Information Technology student at 2004, 2005, 2006 and 2007 entry and for software engineering students at 2006 and 2007 entry.

In experiment 3 an algorithm generates a model having an accuracy of 83.87% with execution time 0.07 second. The lowest result scored in experiment three was 51.4 % in 2004 CS students with execution time 0.11 seconds.

Table 2: Experiment 1 Results

Dept	Academic Year	Neural Net (MLP)	Time Taken to Execute (second)
CS	2004	45.7143%	1.62
	2005	64.7059%	1.39
	2006	75.7143%	5.22
	2007	76.1905%	9.85
IS	2004	72 %	1.75
	2005	72.2222 %	2.71
	2006	64.5161 %	6.76
	2007	71.3178 %	11.77
IT	2004	80.6452 %	1.85
	2005	56.25 %	2.74
	2006	70.9402 %	6.13
	2007	70.6294 %	7.74
SWE	2006	68.0412	6.02
	2007	68.0412	7.1
10-Fold Cross Validation			

Table 3 Experiment 2 Result

Dept	Academic Year	Naive Bayesian	Time Taken to Execute (second)
CS	2004	91.4286 %	0.01
	2005	82.3529 %	0.01
	2006	94.2857%	0
	2007	91.8367 %	0
IS	2004	84 %	0.01
	2005	94.4444 %	0
	2006	89.2473 %	0
	2007	93.7984 %	0.01
IT	2004	93.5484 %	0
	2005	75 %	0
	2006	95.7265 %	0.01
	2007	93.7063 %	0
SWE	2006	93.8144 %	0
	2007	93.8144 %	0
10-Fold Cross Validation			

Table 4 Experiment 3 Results

Dept	Academic Year	SMO	Time Taken to Execute (second)
CS	2004	51.4286%	0.11
	2005	64.7059 %	0.11
	2006	71.4286 %	0.09
	2007	78.2313 %	0.16
IS	2004	68 %	0.05
	2005	72.2222 %	0.05
	2006	64.5161 %	0.1
	2007	72.8682 %	0.12
IT	2004	83.871 %	0.07
	2005	59.375 %	0.1
	2006	75.2137 %	0.1

	2007	74.1259 %	0.12
SWE	2006	62.8866 %	0.1
	2007	62.8866	0.1
10-Fold Cross Validation			

V. CONCLUSION AND FUTURE WORK

A. Conclusion

In this study, the aim was to develop the predictive model for student performance through machine learning techniques. The classification models are experimented with the Naive Bayesian, neural network (MLP) and SVM algorithms. Different models for each of these algorithms are built, and the best overall classifier model from the Naive Bayesian, a neural net (MLP) and SVM algorithm has been selected. Based on the developed model for each department using classifiers naïve Bayesian had high accuracy than the other classifiers, techniques used for model building.

The academic performance of student who had joined the university at 2006 E.C under the department of Information Technology had higher performance using Naïve Bayesian method (95.7%), This implies that Naive Bayesian have higher performance when we compared with the other two selected methods (SMO and MLP).

Even if the performance of 2006 (Information Technology) had high; it does not mean that they have been the Top performance so the college and university management should concentrate to become the performance of student to be 100 %.

B. Future Work

This study is the starting point of education machine learning research in Wolkite University and can be future developed in different ways. To use more data: the research used only one college programs for the model; however, the University Campus has 32 departments of them from 7 colleges. The inclusion of other programs in the model can be able to give deferent ideas and allow the university to gain better understanding for academic performance for students.

Different machine learning techniques will be applied to improve the performance instead of mention in this study.

ACKNOWLEDGMENTS

We would like to thanks Wolkite University collage of computing and Informatics Dean, Ato fedu Akmal and Wolkite University Registrar office dean, Ato Fekadu Mamo for their support giving the data set in order to complete this study.

REFERENCES

[1] A. Asfaw, "Analysis of gender disparity in regional examination:Case of Dessie town; Ethiopia," *Basic Research Journal of Education Research and Review* vol. 4, pp. 29-36, February 2015.

[2] Azwa A and N. Hafieza, "First Semester Computer Science Students' Academic Performances Analysis by Using Data Mining Classification Algorithms," *Proceeding of the International Conference on Artificial Intelligence and Computer Science,* September 2014.

[3] Swasti S and M. J, "A Study on WEKA Tool for Data Preprocessing,Classification and Clustering," *International Journal of Innovative Technology and Exploring Engineering (IJITEE),* vol. 2, pp. 250-253, May 2013.

[4] Tsehaye W and Y. M, "Determinants of Student Attrition at College of Business and Economics, Mekelle University: Econometric Investigation," *Proceedings of the National Symposium on "Establishing, Enhancing & Sustaining Quality Practices in Education",* pp. 110-121.

[5] Z. Afzal, "Impact of Motivation on Academic Performance of University Graduates," *J. Glob. & Sci. Journal of Global and Science* vol. 1, JUNE 2013.

[6] Shaymaa E, Tsunenori M, Kazumasa G, and S. H, "Efficiency of LSA and K-means in Predicting Students' Academic Performance Based on Their Comments Data," *6th International Conference on Computer Supported Education,* pp. 63-74, 2014.

[7] M. o. Education, "Education Statistics Annual Abstract," Addis Ababa, November 2013.

[8] Pedro S, João M, and C. S, "Educational Data Mining: preliminary results at University of Porto," pp. 1-11, JUNE 2014

[9] A.F.ElGamal, "An Educational Data Mining Model for Predicting Student Performance in Programming Course," *International Journal of Computer Applications,* vol. 70, pp. 22-28, May 2013.

[10] J. Nayak, B. Naik, and H. S. Behera, "A Comprehensive Survey on Support Vector Machine in Data Mining Tasks: Applications & Challenges," *International Journal of Database Theory and Application,* vol. 8, pp. 169-186.

[11] S. K. Shevade, S. S. Keerthi, C. Bhattacharyya, and K. R. K. Murthy, "Improvements to the SMO Algorithm for SVM Regression," *IEEE,* vol. 11, pp. 1188-1193, SEPTEMBER 2000.

