PREDICTION OF STUDENTS' PERFORMANCE IN E-LEARNING ENVIRONMENT USING RANDOM FOREST

ABUBAKAR YUSUF

A dissertation submitted in partial fulfilment of the
requirements for the award of the degree of
Master of Computer Science

Faculty of Computing
Universiti Teknologi Malaysia

JUNE 2017

ACKNOWLEDGEMENT

I wish to begin by thanking Almighty Allah for giving me the opportunity, knowledge, strength, courage patience and determination to undertake this study.

I would like to express my profound appreciation to my supervisor Dr. Nor Bahiah Hj Ahmad for her advice and for giving me the opportunity and motivation to undertake this study. She was all the time available for fruitful research discussions and her guidance and encouragement during this period were the strongest contribution to this study. Finally, I appreciate her total confidence and the responsibility granted to me throughout this study period.

To my beloved family and friends, many thanks for all the prayers, support, love and care shown to me throughout the period of my programme. Their endless love has given me the strength and confidence to complete the study. Lastly but not least, I am so deeply indebted to Nuhu Bamalli Polytechnic Management for the approval giving to me to undertake this programme. My heartily appreciation goes to all those that help me directly or indirectly during the course of my study.

Finally, I would like to express my gratitude to my beloved wife Ummakhulsum, my brothers and sisters and my little kids Aisha and Umar for their persistent patience and support which motivated me immensely along the course of this research. To all my friends, i must thank you for your valuable supports and prayers for the successful completion of my programme.

ABSTRACT

The need of advancement in e-learning technology causes educational data to become very huge and increase very rapidly. The data is generated daily as a result of students interaction with e-learning environment, especially learning management systems. The data contain hidden information about the participation of students in various activities of e-learning which when revealed can be used to associate with the students performance. Predicting the performance of students based on the use of e-learning system in educational institutions is a major concern and has become very important for education managements to better understand why so many students perform poorly or even fail in their studies. However, it is difficult to do the prediction due to the diverse factors or characteristics that influence their performance. This dissertation is aimed at predicting students performance by considering the students interaction in e-learning environment, their assessment marks and their prerequisite knowledge as prediction features. Random Forest algorithm, which is an ensemble of decision trees, has been used for prediction and the comparative analysis shows that the algorithm outperforms the popular decision tree and K-Nearest Neighbor algorithms. However, Naive Bayes outperformed Random Forest. In addition to the performance prediction, Random Forest was also used to identify the significant attributes that influence students performance, which was validated by a statistical test using Pearson correlation. The research therefore, revealed that lab task, assignments, midterm and prerequisite knowledge are significant indicators of students performance predictions.

ABSTRAK

Keperluan kemajuan dalam teknologi e-pembelajaran menyebabkan data pendidikan menjadi sangat besar dan berkembang dengan pantas. Data ini dihasilkan setiap hari daripada hasil interaksi pelajar-pelajar dalam persekitaran e-pembelajaran, terutamanya dari sistem pengurusan pembelajaran. Data pembelajaran mengandungi maklumat yang tersembunyi mengenai penyertaan pelajar-pelajar dalam pelbagai ativiti e-pembelajaran yang mana apabila didedahkan boleh diguna untuk dikaitkan dengan prestasi pelajar. Meramalkan prestasi pelajar-pelajar berdasarkan tahap penggunaan pelajar terhadap sistem e-pembelajaran telah menjadi tumpuan utama dan menjadi sangat penting dalam pengurusan pendidikan untuk memahami dengan lebih baik kenapa ramai pelajar-pelajar mempunyai prestasi yang rendah atau gagal dalam pembelajaran mereka. Akan tetapi, ramalan ini agak rumit untuk dilaksanakaan kerana terdapat pelbagai faktor dan ciri-ciri yang mempengaruhi prestasi mereka. Disertasi ini bertujuan untuk meramal prestasi pelajar-pelajar dengan mengkaji interaksi mereka dalam persekitaran e-pembelajaran, markah penilaian pelajar dan prasyarat pengetahuan sebagai ciri-ciri peramalan. Algorithma Random Forest, yang merupakan gabungan pokok keputusan telah digunakan sebagai teknik ramalan, dan hasil analisa kajian perbandingan menunjukkan prestasi Random Forest telah mengatasi pokok keputusan lain dan juga algorithma K-Nearest Neighbour. Sungguhpun begitu, Naive Bayes dapat mengatasi prestasi Random Forest dalam meramal prestasi pelajar. Sebagai tambahan kepada ramalan prestasi pelajar ini, Random Forest juga telah digunakan untuk mengenali ciri-ciri signifikan yang mempengaruhi prestasi pelajar-pelajar, yang telah disahkan oleh ujian statistik menggunakan korelasi Pearson. Kajian ini juga mendedahkan yang tugasan makmal, tugasan, ujian pertengahan penggal dan prasyarat pengetahuan adalah petunjuk ketara kepada ramalan prestasi pelajar.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CART	-	Classification and Regression Trees
CSV	-	Comma Separated Values
EDM	-	Educational Data Mining
E-Learning	-	Electronic Learning
FN	-	False Negative
FP	-	False Positive
ID3	-	Iterative Dichotomiser 3
IP	-	Internet Protocol
KNN	-	K-Nearest Neighbor
LMS	-	Learning Management System
MATLAB	-	Matrix Laboratory
ML	-	Machine Learning
MOOC	-	Massive Open Online Course
MOODLE	-	Modular Object Oriented Dynamic Learning Environment
NN	-	Neural Network
PT	-	Programming Techniques
RF	-	Random Forest
SPSS	-	Statistical Package for the Social Sciences
TN	-	True Negative
TP	-	True Positive
UTM	-	Universiti Teknologi Malaysia
WEKA	-	Waikato Environment for Knowledge Analysis

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Overview

Student performance in educational institutions such as Universities and Colleges is not only a pointer to the effectiveness of the institutions but also major determinant of the future of students in particular and the nation in general. Learning outcomes have become a phenomenon of interest to all and this account for the reason why scholars have been working hard to find out factors that militate against good academic performance (Aremu and Sokan, 2003). Academic achievement of learners has attracted attention of scholars, parents, policymakers and planners. Adeyemo (2001) noted that the major goal of the school is to work towards attainment of academic excellence by students. This student academic performance can be regarded as the observable and measurable behavior of a student in a particular situation. In the context of this research, performance is defined as the final mark acquired by a student at the end of the term/semester in a subject enrolled through learning management system. The final mark is a cumulative sum of the internal assessment (quizzes, assignments, midterm test) and the final examination score.

Performance of students may be influenced by several factors such as gender, age, parents socioeconomic situation, area of resident, nature of school being attended, school medium of teaching, number of study hours spent daily, and nature of accommodation which may be school own hostel or otherwise (Romero *et al.*, 2013). A number of researches about factors affecting students performance at different study levels have been conducted by many authors. Graetz (1995) suggests that a student educational success contingent heavily on social status of students parents/ guardians in the society. Considine and Zappalà (2002) noticed the same that parents income or social status positively affects the student test score in examination. Bratti and

Staffolani (2002) observed that the measurement of students previous educational outcomes are the most important indicators of students future achievement, This concludes that the better the performance of students in previous studies, the better their performance in future attempts.

Students' performance prediction is one of the earliest and most valuable applications of Educational Data Mining (EDM) and its objective is to measure the hidden value of students performance, understanding or grade from the other information, attitude or behavior of those students (Romero *et al.*, 2013). This is a difficult issue to address because of the diverse number of factors or attributes that influences the performance of students such as cultural, family background, psychological history, previous academic performance, parents economic situation, previous schooling, interaction between student and faculty, etc. (Araque *et al.*, 2009).

Several EDM techniques have been used in the prediction of students' performance such as classification, regression and density estimation for predicting variable with categorical value, continuous valued variable and probability density function respectively. It is essential to note that most recent researches on EDM for students performance prediction were primarily applied to cases of University and high school students (Kotsiantis *et al.*, 2010) and specifically, in most cases to e-learning or related mode of instruction (Romero *et al.*, 2013) This is fundamentally as a result of increase in the use of learning management systems (LMSs) such as Moodle, Blackboard, Edmodo, Cornerstone, Schoology, ConnectEdu, Kalboard 360 etc. These learning management systems have the ability to provide unlimited access to learning materials, easily tracks learner performance, and enables easy and convenient expansion of e-learning courses. The LMS collects huge amount of information related to user visits and interactions such as viewing of resources, submission of assignments, participation in discussion forums etc. which are very essential in predicting students performance, analysing students behaviour and assisting instructors, detecting problems and providing improvements (Romero *et al.*, 2008)

Random Forest algorithm being an ensemble machine learning algorithm and these ensemble techniques have become impressive in the area of prediction. These techniques incorporate different machine learning algorithms with the aim of improving the overall prediction accuracy (Dietterich, 2000). The technique is based on the thought that collective decision is more accurate when compared to individual decisions. Ensemble method combines a set of classifiers to create blended model which results to an improved accuracy. Research shows that prediction from ensemble

models gives better outcome when compared to individual classifiers. Several researches have been conducted by different researchers such as (Domingos, 1996), (Opitz and Maclin, 1999) and (Bauer and Kohavi, 1999), and they all proved that combining multiple classifiers provides an improved prediction accuracy. Ensemble techniques have also been successfully applied to diverse real-world tasks. Indeed, they have been found to be useful and usually perform better in almost all places where learning techniques are exploited. These techniques have been specifically applied in areas such as computer vision which has been used by (Viola and Jones, 2004). Other areas where ensemble techniques performed better include Intrusion detection (Giacinto *et al.*, 2003) and medical diagnosis (Zhou and Jiang, 2003)

This study therefore focuses on developing prediction model of students academic performance based on their interaction with learning management system in order to explore the performance of ensemble techniques in the prediction of student performance with the aim of achieving high prediction accuracy.

1.2 Problem Background

In e-learning, predicting the performance of students is a great concern to the education managements. For example, it could give an appropriate warning to students those who are at risk by forecasting the grade of students, and help them to avoid problems such as dropout, probation etc. and overcome all difficulties in their study. However, measuring of academic performance of students that uses e-learning is challenging since students academic performance hinges on diverse factors or characteristics such as assignment submission, forum post, quiz and so on (Guo *et al.*, 2015)

Among those factors, some are more significant than the others in determining students performance when interacting with learning management system such as Moodle, the most significant factors among others can only be identified through experiment on a target dataset. Identifying such significant factors/characteristics can be of great importance as instructors may focus and ensure students engage themselves on such activities during their course work which will help in improving students performance. The most significant attributes can be identified through feature selection process which include a variety of techniques (such as filter method, wrapper method and embedded method) for selecting features that are most useful or most relevant in

the target dataset.

Another challenging issue in EDM is the choice of appropriate machine learning technique in predicting students performance due to the fact that there is no one single algorithm that obtains the best classification accuracy in all cases as highlighted by (Romero and Ventura, 2010). The performance of machine learning method is heavily dependent of the choice of data representation. A number of machine learning related approaches for predicting students performance on various learning management systems have been proposed but the problems still remains open to research.

Romero and Ventura (2010) applied data mining techniques to predict the marks that university students will obtain in the final exam of a course. The author compared the performance of KNN, MLP, C4.5 and CART algorithms in classifying data of 438 students who use the Moodle LMS. The attributes used for this study include course identification number, total number of assignments done, number of quizzes passed, number of quizzes failed, number of messages sent to forum, number of messages read on forum, total time used on assignments, total time used on quizzes, total time used on forum and final mark the student obtained in the course. The study revealed that in general there is not one single algorithm that obtains the best classification accuracy in all cases (with all datasets). Also, pre-processing task like filtering, discretization or rebalancing can be very important to obtain better or worse results.

Manne *et al.* (2014) used Decision Tree, K-Nearest Neighbour, Naive Bayes and Support Vector Machine algorithms to predict the final grade obtained by each student in a particular course. Moodle was the data source for the research. The attributes considered for the analysis include course ID number, total number of assignments submitted, number of quizzes attended, number of quizzes passed, number of quizzes failed, number of messages sent to teacher, number of messages sent to forum, number of messages sent to chat, number of messages read, total time used in forum, total time used in quizzes and final mark obtained by the student in the at the end of the course period. The study revealed that Decision Tree outperformed other techniques with the predictive accuracy of 85% as such the author recommended the suitability of the technique for developing student performance predictive models.

Sisovic *et al.* (2015) used classification methods to detect connections between prior knowledge and Moodle course activity in relation to passing or failing a course

based on Final grade. The approach was evaluated with data from 153 students obtained on Moodle LMS using J48, JRIP and PART algorithms. The attributes used for this study are number of course views during semester, number of assignment views during semester, number of assignments uploads during semester, number of resource views during semester, attendance of preparatory seminar, maths level, maths result, informatics result and finally the dependent variable pass indicating if the student passes or failed. The study confirmed that Moodle activities, i.e. assignment uploads and course views are better predictors of final success than the Matura examination results. The study also shows that the preparatory seminar contributes to the course success. The feature related to the preparatory seminar is specific because although it belongs to the prior knowledge set of features, it can be considered a university activity. The foregoing confirms the significance of continuous work, despite the high level of prior knowledge.

Akçapınar (2016) develops classification model that predicts learners approach such as deep learners and surface leaners using data obtained from Moodle learning management system of Computer Education and Instructional Technology Department in Turkey. The interaction data for the students were recorded for a semester. The study considered six features related to various activities, the features include total number of assignment, total number of views in the discussion forum, total number of different days of login, total number of updates (assignments, forum posts etc.), total number of logins and total number of activities in the Moodle environment. K-Nearest Neighbour algorithm was used for the prediction.

1.3 Problem Statement

In e-learning, identifying students who may be at risk of failure is very important in that early identification of such students can lead to providing them with necessary assistance which may prevent them from failure (Dewan *et al.*, 2015). The earlier the at-risk students can be identified, the quicker their problems can be addressed by the education managements by taking appropriate actions to prevent them from poor performance. The prediction of students performance is expected to help develop some special arrangements that educational institutions organize in order to help students perform better in their courses. As a result of technological development, e-learning platforms now provide a means of monitoring students learning behaviour and their interactions with the system through the systems log files, the data can then be analyzed using data mining techniques. However, careful selection of those activities

and machine learning techniques is critical. Inappropriate choice of activities and machine learning techniques may result in an unreliable prediction result.

One of the popular machine learning techniques used today in other to improve performance of classification models are the use of ensemble classifier. These ensemble classifiers are a set of machine learning algorithms whose individual results are combined using weighted or unweighted voting technique to predict unseen data. One of the most effective research area in predictive analytics has been to study techniques for developing good ensemble of classifiers. It has therefore been discovered that ensemble techniques are much more accurate that the individual base classifiers that make them up (Dietterich, 2000)

A number of classification models prevails recently which has been implemented by many researchers on student performance prediction. Most of them are successfully implemented by using different classification algorithms on different datasets, only few of the researchers used ensemble classifiers (such as Bagging, Boosting and RandomForest) on e-learning data especially the Moodle. Another trending issue is that there are so many factors or features that contribute to the students performance in e-learning, defining those factors has been an open research interest. This research is concerned with predicting student academic performance of undergraduate students in Computer Science Department taking Data Structure course at Faculty of Computing Universiti Teknologi Malaysia and also to identify variables that are most significant in predicting students performance. The research will also investigate the performance of Random Forest in predicting the category a student belongs to with respect to their performance in the particular subject. The research is expected to answer the following questions.

1. What are the most significant attributes/features in predicting student performance in particular subject?
2. What is the performance of Random Forest in analyzing and predicting performance of students that use Moodle Learning Management System?
3. What is the impact of Moodle LMS activities and prior knowledge on students academic performance in data structure course?

1.4 Research Aim

The aim of this dissertation is to predict student academic performance based on their interaction with MOODLE learning management, assessment and prerequisite knowledge system using Random Forest algorithm

1.5 Research Objectives

The objectives of this dissertation are as follows:

1. To investigate the features/attributes that are significant indicators of students academic performance based on their interaction with e-learning, assessment and prerequisite knowledge
2. To investigate the performance of Random Forest in predicting students academic performance using interaction data from MOODLE learning management system
3. To correlate the effects of students interaction, assessment and prerequisite knowledge to their academic performance in order to improve the e-learning environment

1.6 Scope of the Study

The scopes of this dissertation are described below:

1. The dissertation mainly focuses on students performance prediction considering their interaction with learning management system (MOODLE) as the input data
2. The data are obtained from MOODLE platform of Department of Computer Science, Universiti Teknologi Malaysia, Computer Science undergraduate students taking Data Structure course

1.7 Significance of the Research

In Educational Data Mining, predicting the performance of a student is a very important to the education stakeholders. For instance, it could give an appropriate warning to students who may be at risk of failing by forecasting their grades, and help them to avoid problems and overcome all difficulties in their study. However, this research is aimed at predicting students performance in e-learning environment. It demonstrates how data mining methods are implemented to analyze the activities of students as they interact with learning management system. The output of the research is a model that classify students based on the way they interact with the e-learning environment. The research result is useful to lecturers by providing feedback about the learning behavior of students and how it affects students performance. It can also be used to trace students who are at risk of failing in order to provide early assistance. The research can also help the academic managements in improving the structure of a course to avoid student failure. Therefor the overall achievement of this research work is to improve the comprehensibility of e-learning environment by understanding the students behavior.

REFERENCES

Abyaneh, H. Z., Nia, A. M., Varkeshi, M. B., Marofi, S. and Kisi, O. (2010). Performance evaluation of ANN and ANFIS models for estimating garlic crop evapotranspiration. *Journal of Irrigation and Drainage Engineering*. 137(5), 280–286.

Adeyemo, D. (2001). Teacher job satisfaction, job involvement, career and organizational commitments as correlates of students academic performance. *Nigerian Journal of Applied Psychology*. 6(2), 126–135.

Ahmad, A. and Dey, L. (2007). A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set. *Pattern Recognition Letters*. 28(1), 110–118.

Akçapınar, G. (2016). PREDICTING STUDENTS'APPROACHES TO LEARNING BASED ON MOODLE LOGS.

Akçapỳnar, G., Altun, A. and Cosgun, E. (2014). Investigating Students' Interaction Profile in an Online Learning Environment with Clustering. In *Advanced Learning Technologies (ICALT), 2014 IEEE 14th International Conference on*. IEEE, 109–111.

Amatriain, X., Jaimes, A., Oliver, N. and Pujol, J. M. (2011). Data mining methods for recommender systems. In *Recommender Systems Handbook*. (pp. 39–71). Springer.

Amrieh, E. A., Hamtini, T. and Aljarah, I. (2016). Mining Educational Data to Predict Students academic Performance using Ensemble Methods. *International Journal of Database Theory and Application*. 9(8), 119–136.

Anuradha, C. and Velmurugan, T. (2015). A comparative analysis on the evaluation of classification algorithms in the prediction of students performance. *Indian Journal of Science and Technology*. 8(15).

Araque, F., Roldán, C. and Salguero, A. (2009). Factors influencing university drop out rates. *Computers & Education*. 53(3), 563–574.

Aremu, A. and Sokan, B. (2003). A multi causal evaluation of academic performance

of Nigerian learners: Issues and implications for national development. *Department of Guidance and counseling, University of Ibadan*.

Baker, R. S. and Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *JEDM-Journal of Educational Data Mining*. 1(1), 3–17.

Bauer, E. and Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine learning*. 36(1-2), 105–139.

Bhardwaj, B. K. and Pal, S. (2012). Data Mining: A prediction for performance improvement using classification. *arXiv preprint arXiv:1201.3418*.

Blackorby, J., Edgar, E. and Kortering, L. J. (1991). A third of our youth? A look at the problem of high school dropout among students with mild handicaps. *The Journal of Special Education*. 25(1), 102–113.

Bratti, M. and Staffolani, S. (2002). Student time allocation and educational production functions.

Chen, M.-S., Han, J. and Yu, P. S. (1996). Data mining: an overview from a database perspective. *IEEE Transactions on Knowledge and data Engineering*. 8(6), 866–883.

Considine, G. and Zappalà, G. (2002). The influence of social and economic disadvantage in the academic performance of school students in Australia. *Journal of Sociology*. 38(2), 129–148.

Delavari, N., Beikzadeh, M. R. and Phon-Amnuaisuk, S. (2005). Application of enhanced analysis model for data mining processes in higher educational system. In *Information Technology Based Higher Education and Training, 2005. ITHET 2005. 6th International Conference on*. IEEE, F4B–1.

Dewan, M. A. A., Lin, F., Wen, D. *et al.* (2015). Predicting Dropout-Prone Students in E-Learning Education System. In *Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom), 2015 IEEE 12th Intl Conf on*. IEEE, 1735–1740.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International workshop on multiple classifier systems*. Springer, 1–15.

Domingos, P. (1996). Unifying instance-based and rule-based induction. *Machine Learning*. 24(2), 141–168.

Edgar, E. and Polloway, E. A. (1994). Education for adolescents with disabilities: Curriculum and placement issues. *The Journal of Special Education*. 27(4), 438–452.

Efrati, V., Limongelli, C. and Sciarrone, F. (2014). A Data Mining Approach to the Analysis of Students Learning Styles in an e-Learning Community: A Case Study. In *International Conference on Universal Access in Human-Computer Interaction*. Springer, 289–300.

Efron, B. (2013). Bayes' theorem in the 21st century. *Science*. 340(6137), 1177–1178.

El Gamal, A. (2013). An educational data mining model for predicting student performance in programming course. *International Journal of Computer Applications*. 70(17).

Filippidi, A., Tselios, N. and Komis, V. (2010). Impact of Moodle usage practices on students performance in the context of a blended learning environment. *Proceedings of Social Applications for Life Long Learning*, 2–7.

Ghassemzadeh, S., Shafflie, M., Sarrafi, A. and Ranjbar, M. (2013). The importance of normalization in predicting dew point pressure by ANFIS. *Petroleum Science and Technology*. 31(10), 1040–1047.

Giacinto, G., Roli, F. and Didaci, L. (2003). Fusion of multiple classifiers for intrusion detection in computer networks. *Pattern recognition letters*. 24(12), 1795–1803.

Gneiting, T. and Raftery, A. E. (2005). Weather forecasting with ensemble methods. *Science*. 310(5746), 248–249.

Graetz, B. (1995). Socioeconomic status in education research and policy. *Socioeconomic status and school education*, 23–51.

Guo, B., Zhang, R., Xu, G., Shi, C. and Yang, L. (2015). Predicting Students Performance in Educational Data Mining. In *Educational Technology (ISET), 2015 International Symposium on*. IEEE, 125–128.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*. 3(Mar), 1157–1182.

Hansen, L. K. and Salamon, P. (1990). Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*. 12(10), 993–1001.

Hinton, P. R., McMurray, I. and Brownlow, C. (2014). *SPSS explained*. Routledge.

Hu, Y.-H., Lo, C.-L. and Shih, S.-P. (2014). Developing early warning systems to predict students online learning performance. *Computers in Human Behavior*. 36, 469–478.

Kotsiantis, S., Patriarcheas, K. and Xenos, M. (2010). A combinational incremental ensemble of classifiers as a technique for predicting students performance in distance education. *Knowledge-Based Systems*. 23(6), 529–535.

Liang, J., Yang, J., Wu, Y., Li, C. and Zheng, L. (2016). Big Data Application in Education: Dropout Prediction in Edx MOOCs. In *Multimedia Big Data (BigMM), 2016 IEEE Second International Conference on*. IEEE, 440–443.

Luan, J. (2002). Data Mining and Knowledge Management in Higher Education-Potential Applications.

Manne, S., Yelisetti, S., Kakarla, M. and Fatima, S. (2014). Mining VRSEC student learning behaviour in moodle system using datamining techniques. In *Computer and Communications Technologies (ICCCT), 2014 International Conference on*. IEEE, 1–7.

Namdeo, J. and Jayakumar, N. (2014). Predicting Students Performance Using Data Mining Technique with Rough Set Theory Concepts. *International Journal*. 2(2).

Opitz, D. and Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*. 11, 169–198.

Pandey, U. K. and Pal, S. (2011). Data Mining: A prediction of performer or underperformer using classification. *arXiv preprint arXiv:1104.4163*.

Panigrahi, S., Kundu, A., Sural, S. and Majumdar, A. K. (2009). Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning. *Information Fusion*. 10(4), 354–363.

Patidar, P., Dangra, J. and Rawar, M. (2015). Decision Tree C4. 5 algorithm and its enhanced approach for Educational Data Mining. *Int. J. Futuristic Trends Eng. Technol.* 2(2), 14–24.

Pyle, D. (2005). Data Preparation and Preprocessing. In *Do Smart Adaptive Systems Exist?* (pp. 27–53). Springer.

Quinlan, J. R. (1979). *Induction over Large Data Bases.* Technical report. DTIC Document.

Romero, C., López, M.-I., Luna, J.-M. and Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers & Education*. 68, 458–472.

Romero, C. and Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*. 33(1), 135–146.

Romero, C. and Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*

(Applications and Reviews). 40(6), 601–618.

Romero, C., Ventura, S. and García, E. (2008). Data mining in course management systems: Moodle case study and tutorial. *Computers & Education*. 51(1), 368–384.

Shyamala, K. and Rajagopalan, S. (2006). Data mining model for a better higher educational system. *Information Technology Journal*. 5(3), 560–564.

Sisovic, S., Matetic, M. and Bakaric, M. B. (2015). Mining student data to assess the impact of moodle activities and prior knowledge on programming course success. In *Proceedings of the 16th International Conference on Computer Systems and Technologies*. ACM, 366–373.

Tair, M. M. A. and El-Halees, A. M. (2012). Mining educational data to improve students' performance: a case study. *International Journal of Information*. 2(2).

Veenstra, C. P., Dey, E. L. and Herrin, G. D. (2008). Is Modeling of Freshman Engineering Success Different from Modeling of Non-Engineering Success? *Journal of Engineering Education*. 97(4), 467–479.

Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International journal of computer vision*. 57(2), 137–154.

Walters-Williams, J. and Li, Y. (2010). Comparative study of distance functions for nearest neighbors. In *Advanced Techniques in Computing Sciences and Software Engineering*. (pp. 79–84). Springer.

West, D., Dellana, S. and Qian, J. (2005). Neural network ensemble strategies for financial decision applications. *Computers & operations research*. 32(10), 2543–2559.

Yadav, S. K. and Pal, S. (2012). Data mining: A prediction for performance improvement of engineering students using classification. *arXiv preprint arXiv:1203.3832*.

Zafra, A., Romero, C., Ventura, S. and Herrera-Viedma, E. (2009). Multi-instance genetic programming for web index recommendation. *Expert Systems with Applications*. 36(9), 11470–11479.

Zhou, Z.-H. (2012). *Ensemble methods: foundations and algorithms*. CRC press.

Zhou, Z.-H. and Jiang, Y. (2003). Medical diagnosis with C4. 5 rule preceded by artificial neural network ensemble. *IEEE Transactions on information Technology in Biomedicine*. 7(1), 37–42.

Zhou, Z.-H., Wu, J. and Tang, W. (2002). Ensembling neural networks: many could be better than all. *Artificial intelligence*. 137(1-2), 239–263.