Interesting Attributes of Student Performance Using Machine Learning Models Based on Family and Educational Backgrounds in the Faculty of Agriculture and Technology at Rajamangala University of Technology Isan

Sakchan Luangmaneerote* and Anyawee Chaiwachiragompol*

Received : January 23, 2021
Revised : March 3, 2021
Accepted : March 31, 2021

Abstract

This research studied used six machine learning models to predict student performance and studied interesting attributes that cause influence students to achieve their degree based on their family background and background studies. The student's data was obtained from Education service system (ESS) of University using data that was collected the last 5 years, and used five machine learning models to predict the student performance. Results had been clearly showed that the data related to student's family collected in ESS is was not sufficient for accurate prediction, while student grades, course learning and their study location during high school is was a distinguishing feature caused influencing students to achieve their degrees. Additionally, the best accuracy result was Random forest which can predicted an accuracy of 0.62, while the other models did not show satisfactory results. The results of this research may enable the faculty to develop an improved database to predict student performance and promote faculty resource management.

Keywords: Family and Educational Backgrounds, Student Performance, Feature Importance.

1. Introduction

The Faculty of Agriculture and Technology, Rajamangala University of Technology Isan, Surin Campus has been using Education service system (ESS) for over five years in order to facilitate student enrollment [1]. However, no one ever used the data stored in the ESS for in-depth analysis.

In the past five years, 4,107 students have been enrolled through ESS, while the number of students who dropped out of the faculty stands at 1,441. Some students are not able to achieve their degrees due to many factors such as dislike of the field of study, financial support, poor educational system or an absence of motivation or interest. These factors play a crucial role in student performance.

There are a number of reasons why educational organizations in Thailand ignore this data, even though they realize that it is more important. One of the main reasons is the lack of adequate educational methods to predict student performance and a deficient understanding of what factors affect student performance. Even though, currently, many published research worldwide attempt to propose their approaches, the context of each area of study is quite different.

Student performance is an essential part of higher education, although many educational organizations consider remarkable record of academic achievement to be the best indicator. There are numerous criteria based on literature review. U. Bin Mat [2] specified that student performance can be measured through learning assessment and co-curriculum. However, several recent studies [3], [4], [5] have suggested that graduation is the best indicator of student success.

Many published research mentioned student performance depending on various factors such as different economic, social, mental, and environmental factors [6], [7]. These factors had been stressed that affected student performance. Some of the research attempts to explain links of dropout caused by gender, race, grade level, school location, school type, week peers, distance from learning place [8], [9], [10].

* *Department of Computer Technology, Faculty of Agriculture and Technology, Rajamangala University of Technology Isan, Surin campus.*

ln order to address these problems, this research set the following objectives:

1. To study feature importance affecting the student performance.

2. To find the best machine learning model from six machine learning models for predicting the student performance.

The results of this research will be used to update all data collected in ESS and apply machine learning model based on family background and educational background to predict the student performance in the future.

2. Theoretical background and related researches

The systematic review has been conducted in this section with the purpose of discovering the possible attributes which are likely to be used to predict student performance. One of the most frequently used attributes is the Cumulative Average Score [11], [12], [13], [14]. The main reason is that it is a substantial value which can easily measure the student performance and can also be considered an indicator of actual academic potential [2]. As mentioned by Christian and Ayub [15], they exactly stated that GPA is the most influential attribute of regarding the survival of the students in their studies.

In addition, some studies have mentioned student demographic consisting of age, family background and gender [13], [16]. Furthermore, blood group is an interesting attribute because some papers stated that the type of blood group possibly affects student performance at different levels [17], [18]. The main reason as to many researchers selecting gender is because male and female have different approach of learning styles [2].

One of the most significant current discussions in student performance is family background. Most researchers have attempted to explain that the impact of family backgrounds is much stronger than the effect of school resources [19], [20]. According to Fang [21] and Sun [22], they stated that parent's income and educational levels of parents contribute to the achievement of students.

Nowadays, a number of algorithms, such as Decision tree, Neural networks, Naïve Bayes and so on, have been proposed to predict the student performance.

Decision tree is a popular method because it uses if-then rule which is easy to understand [23]. There are approximately ten papers that have used decision trees as a way to assess student performance [24].

Neural networks is another popular technique. The advantage of neural networks is that all possible interactions between variables can be detected. Besides, more than eight articles using this algorithm have been published [20].

Naive Bayes is another interesting model. Among the 30 papers published are four studies using the Naive Bayes model to predict the student performance with the aim of finding the best predictor [18].

The gaussian Naive Bayes technique follows the Bayes theorem. This technique is described as the probability of an event based on the prior knowledge of conditions associated with the event. It provided an accuracy of 74% which is an interesting result for further investigation [25].

Logistic regression is a classification relationship between one or more existing independent variables. It is used to predict binary results based on relationships between one or more remaining independent variables. Some researchers stress that using logistic regression show outperformance [26].

Random forest is a classification algorithm that consists of many decision structures. Many researchers have suggested that this algorithm can assist them to predict accurately and provide satisfactory results [27], [28].

Almost all research has not mentioned feature attributes which play a crucial role in assisting to improve the accuracy of prediction. This research not only requires finding the best model accuracy but also still needs to find the important attribute in order to update the data in the ESS.

3. Research Methodology

This research utilized data stored in ESS between 2015 and 2019, which collected 4,107 records and selected 18 attributes, related to family and educational background, out of 87 attributes stored in the ESS and used scikit-learn in Python for analysis of data and used other packages for visualization. Almost parameters used default value from scikit-learn in Python except for Neural network using solver='lbfgs', alpha=1e-5, hidden_layer_sizes=(10, 10), random_state=1. This research is based on the CRISP-DM process, which consists of six stages.

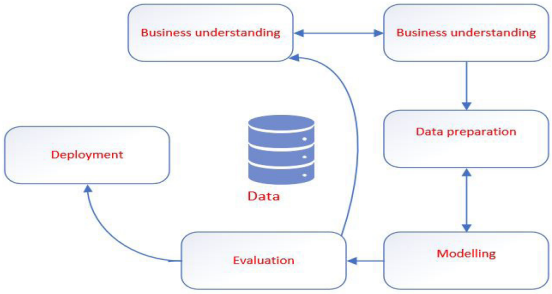

Figure 1. *CRISP-DM process.*

Business understanding: It is the first step in understanding the real problem. The researcher found that the number of students studying in the Faculty of Agriculture and Technology had dropped out at a high rate. If the data stored in ESS could be utilized to predict student performance accurately, it could be proved that the data stored in the ESS is sufficient for prediction.

Data understanding: It is the second step in understanding the data stored in ESS. The hypothesis is that the family and educational background should have a significant effect on the student performance. Therefore, data about the Grade Point Average of students in the high school, the study location during high school and all data related to the family and educational background were taken into accounts.

Data preparation: This step is the selective process of converting the collected data into data that can be analyzed in the next step. At this step, data may be modified such as converting the data to the same range or filling in missing data. This step consists of three sections as follows.

- Selection: This is process of the gathering data and data selection process, identifying the sources of data needed for processing and analyzing. All data associated with family and educational background would be selected in order to use for analysis.
- Data cleaning: It is a process of ensuring that all data is flawless. This is the process of preparing data for analysis by deleting, correcting inaccurate or incomplete data. Missing values are missing data, mostly caused by a process of faulty storage. There are several ways to correct missing values such as using average values or deleting missing values out. Inaccurate value is false data from reality, mostly caused by incorrect data entry such as the number of children to be raised by parents. Duplicate data is a type of data that has a copy in the same system and can be caused by poor data entry such as a student's name and surname. The percentage of missing information is quite low. Therefore, this research selected a deletion method that affected some data less than others.
- Data transformation: Normally, machine learning models require categorized data in the form of numeric data. Therefore, some data will be converted to numeric data after cleaning. For this research, 18 attributes were converted whose details of shown in Table 1 have 2,798 records after cleaning as shown in Table 1. Furthermore, the attribute of status of student is the class label and status of student is studying had been removed before put it into the machine learning model.

Modelling: it is part of verifying that the data is in good condition, and it makes it possible to find useful patterns in the data of interest. This step creates an interesting model and find the required answer. In this study, six models that many studies claim as highly effective were selected.

Table 1. *Selected attributes used for prediction.*

Attributes	Value	Description
High school Grade Point Average (GPA)	GPA between 0.00 and 4.00	Previous grades during high school may have an impact on student achievement.
The location of study of high school	Area is in the range of 0 to 164	The inequality of learning in different areas might affect students' academic achievement.
Gender	Male = 0 Female = 1	A man and woman may have different learning styles.
Curriculum	In the range of 0 to 6	Educational levels, diploma or graduation could affect the student performance.
Program of study	In the range of 0 to 20	Course of study could affect student achievement.
The number of siblings in the family.	In the range of 0 to 8	The number of siblings in the family can affect student achievement.
Father status	Not specified = 0 Dead = 1 Still alive = 2	Status of father; dead or alive
Father's income	100,001-120,000 Baht/year = 0 120,001-130,000 Baht/year = 1 130,001-149,999 Baht/year = 2 150,000 - 300,000 Baht/year = 3 Over 300,000 Baht/year = 4 80,001-100,000 Baht/year = 5 < 80,000 Baht/year = 6 No income = 7 Not specified = 8	Income of father might affect student achievement.
Father's occupation	Self-employed = 0 Government employee = 1 Salaried employee =2 State enterprise = 3 Government official =4 Other =5 Local farmer = 6 No income = 7 Not specified = 8	Father's occupation might affect student achievement.
Mother status	Not specified = 0 Still alive = 2 Dead = 1	Status of mother; dead or alive
Mother's income	100,001-120,000 Baht/year = 0 120,001-130,000 Baht/year = 1 130,001-149,999 Baht/year = 2 150,000 - 300,000 Baht/year = 3 Over 300,000 Baht/year = 4 80,001-100,000 Baht/year = 5 < 80,000 Baht/year = 6 No income = 7 Not specified = 8	Income of mother might affect student achievement.

Attributes	Value	Description
Mother's occupation	Self-employed = 0 Government employee = 1 Salaried employee =2 State enterprise = 3 Government official =4 Other =5 Local farmer = 6 No income = 7 Not specified = 8	Mother's occupation might affect student achievement.
Parent status	Father passed away = 0 Father and mother passed away = 1 Both are in new marriages =2 Mother passed away =3 Mother is in a new marriage =4 Divorced =5 Live together =6 Separated =7 Not specified =8	Status of parent, living together or divorced.
Patron's income	100,001-120,000 Baht/year = 0 120,001-130,000 Baht/year = 1 130,001-149,999 Baht/year = 2 150,000 - 300,000 Baht/year = 3 Over 300,000 Baht/year = 4 80,001-100,000 Baht/year = 5 < 80,000 Baht/year = 6 No income = 7 Not specified = 8	Income of patron might affect student achievement.
Patron's occupation	Self-employed = 0 Government employee = 1 Salaried employee =2 State enterprise = 3 Government official = 4 Other = 5 Local farmer = 6 No income = 7 Not specified = 8	Patron's occupation might affect student achievement.
Blood groups	A = 0 AB = 1 B = 2 O = 3	Blood groups might affect student performance.
High School Program	In the range of 0 to 433	Programs that students study in the course of high school.
Status of student	Studying = 0 Drop out = 1 Completed degree = 2	Unable to complete degree or complete degree.

Evaluation: It is a part of measuring performance of an algorithm. This step is required to find out which models give the best results

Deployment: this step is how to combine machine learning models with existing production environment to make data-driven business decisions.

In this research, six machine learning models, consisting of Random Forest, k-nearest neighbors, Logistic regression, Gaussian Naive Bayes, neural networks, and Decision Tree, were experimented with the purpose of predicting student performance in the Faculty of Agriculture and Technology.

With a cleaned dataset, a machine learning algorithm usually works in two stages. Firstly, the data is split into 20%-80% between testing and training stages. Next, 80 percent of the data is used to train the model while the 20 percent is used to test the accuracy, precision, recall, and F1 score.

4. Results of experiments

In part of results of experiments consists of two main parts. First part demonstrated the classification metrics of six models, prediction of student dropout and completing degree while second part demonstrated the feature importance assists machine learning to predict.

4.1 Results of six models

Four types of outcomes occur when the model performs classification predictions. For this experiment, the results of six models consist of accuracy, recall, precision, and F1. True positive (TP) is when a model predicts an observation belongs to an actual class. True negative (TN) is when a model predicts an observation does not belong to an actual class. False positive (FP) occurs when a model predicts an observation belongs to a class, while in fact it does not. Finally, False negative (FN) happens while a model predicts an observation does not belong to a class while in fact it does.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Accuracy is the ratio of prediction our model got right which is shown in Equation 1.

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

The recall is the measure of a model correctly identifying which is shown in Equation 2.

$$Precesion = \frac{TP}{TP + FP} \quad (3)$$

The precision shows how the model is dependable in classifying samples as Positive as shown in Equation 3.

$$F1 = \frac{2\ x\ Precision\ x\ Recall}{Precesion + Recall} \quad (4)$$

Finally, F1 reflects the weighted average of Precision and Recall which can be calculated as shown in Equation 4.

Figure 2 chart illustrates the accuracy of six machine learning models, consisting of: Random Forest, k-nearest neighbors, Logistic regression, Gaussian Naive Bayes, neural networks, and Decision Tree.

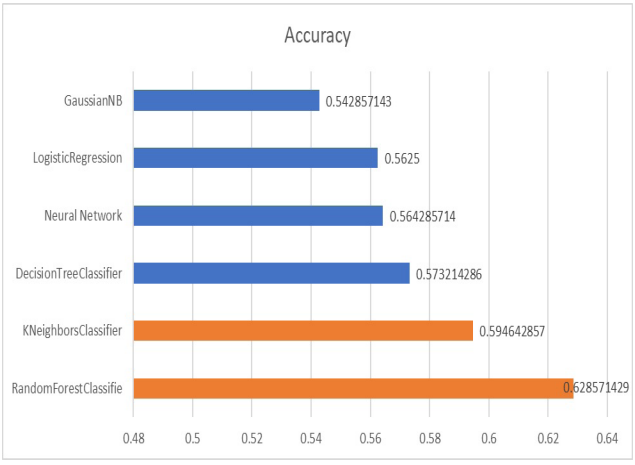

Figure 2. *Accuracy of six models.*

Overall, it can be clearly seen that Random Forest stood at 0.628 which was the most accurate model compared with the other models. The second accurate model was k-nearest neighbors, standing at 0.594, the remaining models showed slightly different results between 0.573 and 0.542.

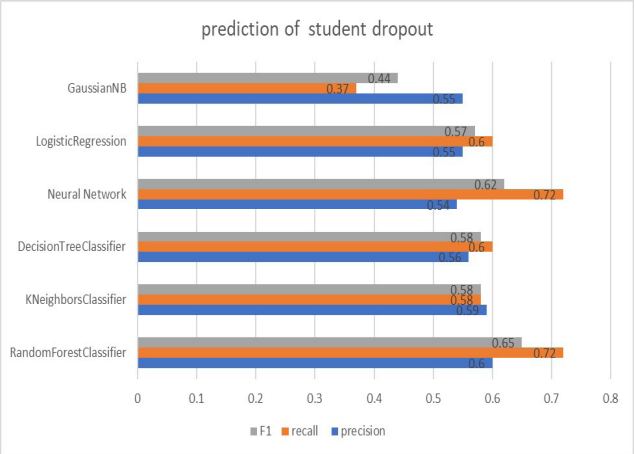

Figure 3. *Prediction of student dropout.*

Figure 3 shows the recall value as the best indicator to show the correct prediction of student dropout from all data. Both Neural network and Random Forest model were regarded as the best model to predict the recall value of 0.72 from all the data. Decision tree and Logistic regression model were the second models to stand at 0.6, while the Gaussian Naive Bayes showed the worst performance of correct prediction of student dropout at 0.37. Regarding the precision value, it was about the precision of the prediction rate of a student dropping out from attempts to predict all. The result of the precision value showed that all models had a negligibly different result with a narrow range from 0.6 to 0.54.

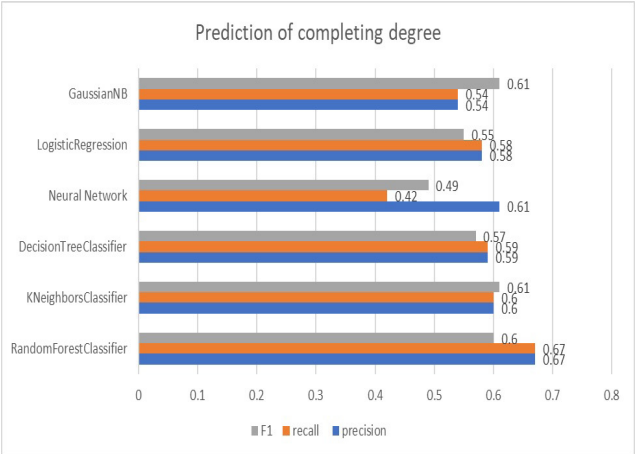

Figure 4. *Prediction of completing degree.*

Figure 4 indicates the model's predictive ability in the case of a student completing a degree. The result of the experiment clearly showed that Random Forest model was still the best model with a value of recall and precision of 0.67. The rest of models suggested insignificant difference of precision and recall value except the model of Neural Network which showed the worst performance with a recall value of 0.42

4.2 Finding Feature Importance

Analyzed input features are capable of assigning a score which is technically known as feature importance. This technique assigns scores to input features of the predictive model which reveals the relative importance of each feature used to predict. The obtained scores can be applied in various situations in a predictive modeling problem, such as improving the way to collect the data and reduce the number of input features.

This research selected Random forests because it a more popular and highly accurate learning algorithm [29]. Normally, the random forest algorithm has accompanied scikit-learn package in Python. Random forests (RF) constructs several decision trees at training. Therefore, predictions from all trees will be used to the final prediction. For each decision tree, Scikit-learn computes nodes

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - W_{right(j)} C_{right(j)} \quad (5)$$

ni_j = the importance of node j
w_j = weight number of samples reaching node j
$C_{left(j)}$ = the impurity value of node j
$left_{(j)}$ = child node form left split on node j
$right_{(j)}$ = child node from right split on node j

importance using Gini Importance as shown in Education 5.

The importance for each feature on decision tree is then

$$fi_i = \frac{\Sigma_{j:node\ j\ splits\ on\ feature\ i} ni_j}{\Sigma_{k \epsilon all\ nodes} ni_k} \quad (6)$$

fi_i = the importance of feature j
ni_j = the importance of node j

computed in Equation 6.

In the practical step, the research loaded data set and split it for training and testing. Then, the final step was to fit the Random Forest Regressor with 250 decision trees and used the feature_importances_ attribute in scikit-learn.

Table 2. *Feature importance.*

Feature	Score
Mother Status	0.005331
Father Status	0.012942
Gender	0.015941
Patron's income	0.0329
Father's income	0.035572
Parent status	0.036798
Mother's income	0.038512
Mother's occupation	0.039503
Fathers occupation	0.041285
Patron's occupation	0.043459
Blood Group	0.048
Curriculum Name	0.049356
The number of siblings in the family	0.054653
Program of study	0.085637
The location of study of high school	0.107076
High School Program	0.119295
High school Grade Point Average (GPA)	0.233741

Table 3 illustrates scores derived from calculating feature Importance using 17 input features. The result reflected which features play a crucial role in assisting machine model prediction.

The feature that gained the most height score was high school grade point average (GPA). This feature showed a value of 0.233741, followed by High School Program and the location of study of high school with a score of 0.119295 and 0.107076, respectively. It can be clearly seen that feature of High School Grade Point Average, High school program, and the location of study of high school play a crucial role in predicting the student performance. The most interesting feature relating to background family was the number of siblings in the family, which showed a score of 0.054653. Other features relating to background family showed insignificant values. It means that these features did not greatly influence student performance.

4.3 Discussion and future work

The results of this research are consistent in many research [22], [23] while contradict other research [19], [20], [18], [20], [21]. The results of each model may depend on different attributes, different data sizes, and different contexts.

This research is aware of some point in research methodology that needs improvement. Interesting attributes should be selected and then put into the machine model for comparison again. In the step of comparison between models, future research should adjust the parameters of each model with the purpose of finding the actual best model. In addition, with the purpose of reducing bias, fold cross-validation should be used instead of splitting data of training sets and test set, with ratios of 80 and 20.

5. Conclusions

This thesis sought to study data stored in the ESS of the University. Almost all of the collected data are family background and educational background. This research aimed to discover useful information from these data and whether it could be utilized to predict student performance. The research established that factors of family background did not have a major influence on the student performance, while educational background was the most important factor among the other factors. Random Forest was the best model to predict student performance compared to the other models. The next step is to collect additional data related to the educational background

s along with integration of the random forest model which may improve the accuracy of the prediction.

6. Acknowledgement

This research would not have been possible without regular support from the Faculty of Agriculture and Technology, Rajamangala Isan, Surin Campus. I would like to thank my Vice President and Dean for always supporting my research.

7. References

[1] "Rmuti ESS." Available Online at http://ess.surin.rmuti.ac.th/Rmuti/Registration/Account/Login.aspx, accessed on 27 December 2020.

[2] U. Bin Mat, N. Buniyamin, P. M. Arsad, and R. Kassim, "An overview of using academic analytics to predict and improve students' achievement: A proposed proactive intelligent intervention," *2013 IEEE 5th Conference on Engineering Education (ICEED),* pp. 126–130, 2013.

[3] A. Venezia, P. M. Callan, J. E. Finney, M. W. Kirst, and M. D. Usdan. *The Governance Divide: A Report on a Four-State Study on Improving College Readiness and Success. National Center Report# 05-3*. Natl. Cent. Public Policy High. Educ, 2005.

[4] A. Slim, G. L. Heileman, J. Kozlick, and C. T. Abdallah, "Predicting student success based on prior performance," *2014 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pp. 410–415, 2014.

[5] T. Ojha, G. L. Heileman, M. Martinez-Ramon, and A. Slim, "Prediction of graduation delay based on student performance," *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 3454–3460, 2017.

[6] J. B. Hansen, *Student Performance and Student Growth as Measures of Success: An Evaluator's Perspective*. 2000.

[7] Y. Beaumont-Walters and K. Soyibo, "An analysis of high school students' performance on five integrated science process skills," *Research in Science & Technological Education*, Vol. 19, No. 2, pp. 133–145, 2001.

[8] S. T. Hijazi and S. M. M. Naqvi, "Factors Affecting Student's Performance.," *Bangladesh e-journal Sociol.*, Vol. 3, No. 1, 2006.

[9] B. Sacerdote, "Peer effects with random assignment: Results for Dartmouth roommates," *Q. J. Econ.*, Vol. 116, No. 2, pp. 681–704, 2001.

[10] G. R. Goethals, "*Peer Effects, Gender, and Intellectual Performance among Students at a Highly Selective College: A Social Comparison of Abilities Analysis. Discussion Paper.*," 2001.

[11] D. M. D. Angeline, "Association rule generation for student performance analysis using apriori algorithm," *SIJ Trans. Comput. Sci. Eng. its Appl.*, Vol. 1, No. 1, pp. 12–16, 2013.

[12] M. M. N. Quadri and N. V Kalyankar, "Drop out feature of student data for academic performance using decision tree techniques," *Glob. J. Comput. Sci. Technol.*, 2010.

[13] E. Osmanbegovic and M. Suljic, "Data mining approach for predicting student performance," *Econ. Rev. J. Econ. Bus.*, Vol. 10, No. 1, pp. 3–12, 2012.

[14] W. Hämäläinen and M. Vinni, "Comparison of machine learning methods for intelligent tutoring systems," *International Conference on Intelligent Tutoring Systems*, pp. 525–534, 2006.

[15] T. M. Christian and M. Ayub, "Exploration of classification using NBTree for predicting students' performance," *2014 International Conference on Data and Software Engineering (ICODSE)*, pp. 1–6, 2014.

[16] V. O. Oladokun, A. T. Adebanjo, and O. E. Charles-Owaba, "*Predicting students academic performance using artificial neural network*: A case study of an engineering

course," 2008.

[17] B. A. Sherke et al., "Do Blood Groups Determine Academic Performance in Medical Students? A Cross Sectional Study," *Int. J. Physiol*., Vol. 6, No. 3, pp. 66–71, 2018.

[18] R. K. Sharma, "Effect of Blood Group on Academic Achievement of Secondary Students in Mathematics," *J. Teach. Educ. Res.*, Vol. 13, No. 02, pp. 91–99, 2018.

[19] A. Gamoran and D. A. Long, "Equality of educational opportunity a 40 year retrospective," *International studies in educational inequality, theory and policy*, Springer, pp. 23–47, 2007.

[20] J. E. Cheadle, "Educational investment, family context, and children's math and reading growth from kindergarten through the third grade," *Sociol. Educ*., Vol. 81, No. 1, pp. 1–31, 2008.

[21] C. Fang and X. Feng, "Family background and academic achievements: a study of stratum differences in compulsory education," *Zhejiang Soc. Sci.*, Vol. 24, No. 8, pp. 47–55, 2008.

[22] Z. Sun, Z. Liu, and B. Sun, "Family, school, and children's academic achievements—based on the study of rural areas in Gansu Province," *J. Beijing Norm. Univ. (Social Sci. Ed.)*, Vol. 37, No. 5, pp. 103–115, 2009.

[23] C. Romero, S. Ventura, P. G. Espejo, and C. Hervás, "*Data mining algorithms to classify students*," 2008.

[24] A. M. Shahiri and W. Husain, "A review on predicting student's performance using data mining techniques," *Procedia Comput. Sci.*, Vol. 72, pp. 414–422, 2015.

[25] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Predicting Student's Performance in Distance Learning using Machine Learning Techniques," *Appl. Artif. Intell*., Vol. 18, No. 5, pp. 411–426, 2004.

[26] N. Thai-Nghe, L. Drumond, A. Krohn-Grimberghe, and L. Schmidt-Thieme, "Recommender system for predicting student performance," *Procedia Comput. Sci*., Vol. 1, No. 2, pp. 2811–2819, 2010.

[27] P. Cortez and A. M. G. Silva, "Using data mining to predict secondary school student performance," 2008.

[28] Y. Abubakar and N. B. H. Ahmad, "Prediction of students' performance in e-learning environment using random forest," *Int. J. Innov. Comput.*, Vol. 7, No. 2, 2017.

[29] M. T. Uddin and M. A. Uddiny, "A guided random forest based feature selection approach for activity recognition," *2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, pp. 1–6, 2015.