A Study of Factors Affecting Learning Efficiency on Higher Education Student Performance Evaluation Dataset Using Feature Selection Techniques

Kairung Hengpraprohm*, Supoj Hengpraprohm*, and Wannee Sudjitjoon **

Received: September 7, 2022
Revised: October 11, 2022
Accepted: November 15, 2022

* Corresponding Author: Supoj Hengpraprohm, E-mail: supojn@yahoo.com

Abstract

This research aimed to discover the features affecting the learning efficiency on the higher education student performance evaluation dataset. The data were gathered from the student in the Faculty of Engineering and the Faculty of Education, Academic year 2019, to forecast the final learning performance of the students. Data consisted of 33 attributes and 145 records from UCI Machine Learning Dataset. Four feature selection techniques, which were Information Gain, Gain Ratio, Correlation Coefficient, and Chi-Square, were applied, along with four data classification methods: K-Nearest Neighbor, Random Forest, Artificial Neural Network, and Linear Regression. Findings demonstrated that the best feature selection techniques were Information Gain and Gain Ratio. When analyzing the relationship of feature data using Pearson's correlation, the feature that had a positive relationship with the data class could be adapted. Further, when considering five features: CUML_GPA, EXP_GPA, READ_FREQ, COURSE ID, and KIDS: meaning when the student had a high cumulative grade point average of the last semester, high academic achievement expectation score, frequency of reading non-scientific books, and divorced or dead parents, they had satisfying learning achievement. Additionally, the attributes, which were STUDY_HRS, AGE, SALARY, IMPACT, had a negative relationship with the data class. It meant the low weekly study hour, young age, low income, and positive impact of the project or activity on the success led to satisfying learning achievement. Thus, it could be concluded that the factors affecting learning efficiency were the accumulated grade point average, achievement expectation score, frequency of reading non-scientific books, and low weekly study hours. All features could be the guideline for designing the learning management for the learner's highest learning efficiency.

Keywords: Feature selection, Data classification, Learning Efficiency, Higher Education.

1. Introduction

Learning efficiency is the ability to connect knowledge to current information. The increasing competency for effective operation by changing the behavior permanently is a result of experience, knowledge, understanding, skills, attitude, and the appropriate application to the situations to achieve the goal. The learning efficiency of the learner is the ultimate goal. Therefore, the recognition of factors affecting learner efficiency is crucial. For this reason, the concerned people, i.e. teachers, executives, or related people in the educational context can utilize the data to improve the education quality, upgrade the student's learning, and design the learning management to be consistent with the factors affecting the learning efficiency. Thus, the study on the relevant aspects is necessary by extracting the data to identify the related factors. Educational authorities collect a great amount of student data. However, only a few could be extracted to gain knowledge.

At present, the application of mathematical and statistical knowledge to data analysis to extract knowledge or knowledge model from the data is popular in many fields, such as

* *Data Science, Science and Technology Faculty, Nakhon Pathom Rajabhat University.*
** *Educational Research and Measurement, Faculty of Education, Nakhon Pathom Rajabhat University.*

medicine [1], banking [2], commerce [3], computer network, and education. Besides, data mining plays more key roles constantly [4], especially in the study to upgrade the student learning efficiency [5], [6]. Forecasting of learning efficiency assists the stakeholders in education take a proactive decision on intervention to enhance the educational quality. Attribute or feature selection for forecasting the student efficiency has a crucial role in increasing the forecasting accuracy and helps to create the strategic plan for learning performance improvement. The algorithm for feature selection is different to forecast student efficiency. However, the study showed different strengths and weaknesses. Thus, the selection of the best attribute is necessary and vital to enhance forecasting efficiency.

Quality feature set of the data is the factor affecting the quality improvement of the data classification model directly. Therefore, the study on the factor affecting student learning efficiency must be conducted via the feature selection technique that eliminates unimportant or irrelevant features to obtain the actual factor affecting the data classification. Each feature selection technique has different strengths and weaknesses. Data classification using the data mining technique [7] has varied methods. The popular methods are K-Nearest Neighbor, Random Forest, Artificial Neural Network, and Linear Regression, which have different good points and weaknesses as well.

The objective of this research was to study the factors affecting student learning using the four feature selection techniques: Information Gain, Gain Ratio, Correlation Coefficient, and Chi-square [8], [9], [10] which would identify which method and feature has the highest impact on the student learning efficiency, so the stakeholder could use the results to develop the education quality. Besides, the research results would give the precise analysis approach as the guideline for analysis and design of a smart decision support system for evaluating the learning competency in artificial intelligence (AI) for the primary school students to create the AI learning innovation for primary school students.

2. Theoretical background and related researches

2.1 Feature selection method

Feature selection is to select a subset of features from a set of current features. The new subset of features aims to increase the efficiency of the performance and minimize the working time because non-important current features will be eliminated. Feature selection is classified into two approaches: (1) Filter Approach, this method selects the important features by calculating the weight or relation value of the feature and keeps only important features, such as Correlation Based Feature Selection, Information Gain, Gain Ratio, and Chi-Square technique. And (2) Wrapper Approach, this method selects the key features according to the accuracy in data classification by using certain classification algorithms.

This research studied the Filter Approach that does not depend on the data classification algorithm. The feature selection methods used in this work are as follows:

1) Correlation Coefficient Base Feature Selection (CFS) is the consideration based on the relationship of the features obtained from the ability evaluation of forecasting the selected features for data classification. It also manages irrelevant features. CFS sorts the sub-group of data dimension and selects the sub-group with a high relationship with the class and does not relate to other classes. The irrelevant data or data with a low relationship with the class will be deleted. The redundant data dimension will be eliminated from the data group with a high relationship. The equation of CFS sub-group evaluation is shown in Equation (1) [10]

$$M_s = \frac{k\bar{r}_{cf}}{\sqrt{k + k(k+1)\bar{r}_{ff}}} \qquad (1)$$

where M_s is the value of data dimension of sub-group S consisting data dimension K

$\bar{r}_{cf}$ is the average of relationship between the variable and class

$\bar{r}_{ff}$ is the average of relationship between the data dimensions

2) Information Gain [10] is the technique that indicates the least desired feature selection by calculating the gain of each data dimension. The data dimension with the highest gain will be selected to use for identification. Equation 2 displays the calculation of information gain or entropy of all data sets. Equation 3 shows the calculation of the entropy of the data dimension of each feature. Equation 4 is to find the Information Gain to consider feature A. The calculation of information gain is to measure entropy before dividing the data by the data dimension and see the efficiency if it is better or not. If the efficiency improves, the information gain is high.

$$E(D) = -\sum_{i=1}^{n} p_i \log_2(p_i) \qquad (2)$$

$$E_A(D) = \sum_{i=1}^{m} \frac{|D_j|}{D} E(D_j) \qquad (3)$$

$$Gain(A) = E(D) - E_A(D) \qquad (4)$$

where P_i is the probability that a record will have a data category

3) Gain Ratio [9], [10] is the indicator of data clustering into subset, which is developed from the Information Gain. Using the Information Gain technique to divide the data set might have partiality; when many considered feature has a high gain, the selected feature is incorrect. For instance, a consideration of a feature that is the specific indicator, such as data code, the code separation will result in many sub-data, and each sub-data set has only one record. When finding the Information Gain, there are numbers of high value.

With partiality, the new indicator of information splitting is developed, which is Gain Ration: it applies the normalization of Information Gain using "Split Information" that can be calculated as shown in Equation 5.

$$SplitInfo_A(D) = -\sum_{i=1}^{m} \frac{|D_j|}{|D|} x \log_2 \frac{|D_j|}{|D|} \qquad (5)$$

where *SplitInfo(A)* means the amount of data that is considered by splitting information in data set D to sub-data set m based on feature A. After calculating the $SplitInfo_A(D)$, the Gain Ration can be calculated as shown in Equation 6.

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)} \qquad (6)$$

4) Chi-squared (X^2) [9], [10] is the feature evaluation by calculating Chi-square to examine whether the frequency distribution of the feature variable follow the model or not, as shown in Equation 7.

$$(x^2) = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \qquad (7)$$

where $O_1, O_2, \ldots, O_N$ is the frequency of the variable obtained from the study

$E_1, E_2, \ldots, E_N$ is the expected frequency

2.2 Data Classification

The data classification technique is a process to create the model to be in the designated group from the sample data, which is called the training data that each data row consists of several fields or attributes. The attribute might be the continuous or categorical value. The classifying attribute is the indicator of the data class. The model obtained from data classification helps to consider the data class that has not yet been classified in the future. This technique is applied in many fields, such as marketing customer classification, abnormality examination, and medical analysis. The data classification algorithms used in this work are as follows:

1) K-nearest neighbor algorithm (KNN) [11] has a simple and understandable method. The principle is to compare the similarity of the interesting data with the training data whether it is similar or near to which data the most at k rows. Then, find the conclusion whether the answer to the interested data should be the same answer as the K-nearest roes.

2) Random Forrest [12] is to create the models using the Decision Tree approach by randomizing the variable and combining the results of each model to count the most repeated result to extract the final result. The decision Tree method is the technique that the result is in a tree structure form. The tree consists of nodes, which each one will be the feature as a test, a branch that exhibits the possible value of the test ed feature, and a leaf that is at the bottom of the tress representing

the data class, which is the forecasting result. The benefit of this method is the accurate forecasting result and fewer overfitting problems.

3) Artificial Neural Network (ANN) [13] is one of the techniques for data mining that is the mathematical model for processing the information with the connectionist calculation to imitate the function of the neural network in the human brain aiming to invent a tool. It can learn pattern recognition and knowledge extraction as same as the human brain. The original concept of this technique is from the study of the bioelectric network in the brain which comprises of nerve cells or "neurons" and "synapses". Each neuron contains the "dendrite" which is the input and the "axon" which is the output of the cell. All cells function with the chemical electrochemistry reaction; when it is activated with the external drive or other cells, the nerve impulse flows through the dendrite to the nucleus which decides whether it should activate other cells or not. If the nerve impulse is strong enough, the nucleus will activate other cells via its axon.

4) Multiple Linear Regression [14] is the regression analysis that most independent variables are the quantitative variables, and the dependent variables are the quantitative variables only. The relationship between the independent and dependent variables can be presented with the linear model. The multiple linear regression comprises of one dependent variable and two or more independent variables. The analysis will be determined the size of the relationship, and the mathematic equation that forecasts the value of the independent variable is created using the studied dependent variables.

2.3 Performance Measures Criteria

The experimental results in this research will be shown in terms of the accuracy [15] which is calculated as in the equation:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (8)$$

where

TP (True Positive) is the number forecasted by the model that it is True and the data is True.

FP (False Positive) is the number forecasted by the model that it is True, but the data is False.

TN (True Negative) is the number forecasted by the model that it is False rue and the data is False.

FN (False Negative) is the number forecasted by the model that it is False, but the data is True.

Test results from this research will be the average of accuracy from the student performance dataset classification model.

3. Literature review

W. Punlumjeak and N. Rachburee.[16] studied the improvement of data classification efficiency using four feature selection methods: genetic algorithms, support vector machine, information gain, and minimum redundancy and maximum relevance, together with the four data classification methods: naive bays, decision tree, k-nearest neighbor, and neural network. The experimental results showed that the minimum redundancy and maximum relevant feature selection method with 10 features selected gave the best result of 91.12% accuracy with a k-nearest neighbor classifier. The results of the present study show that the advantage of future selection to find a minimum and significance of feature is more effective to classify the student performance.

L. Rahman, N. A. Setiawan and A. E. Permanasari [17] proposed feature selection techniques in improving Student's Academic Performance classification accuracy. The algorithm used was Naïve Bayes, Decision Tree, and Artificial Neural Network, which were applied to the features selection: wrapper, and information gain. The application of feature selection was intended to obtain a higher accuracy value. When compared to the embedded method in the studies, the feature selection in this experiment had a lower accuracy rate.

In this study, M. R. Ahmed etc., [18] proposed the calculation of the academic performance of undergraduate students with a predictive data mining model using feature selection techniques with classification algorithms. For this purpose, 800 student's data of the final year studying at the undergraduate level of the department of Computer Science and Engineering

from North Western University, Khulna were used to evaluate the performance of four feature selection methods: genetic algorithms, gain ratio, relief, and information gain and five classification algorithms: K-Nearest Neighbor, Naïve Bayes, Bagging, Random Forest, and J48 Decision Tree. The experimental results depicted that the Genetic algorithms method provided the best accuracy with the K-NN classifier.

M. Wafi, U. Faruq, and A. A. Supianto [19] examined the most popular classification approach, K- Nearest Neighbor (K-NN) which had many adjustments to increase the efficiency of accuracy in data classification and the best feature selection. The article proposed the genetic algorithm, which contained the positive heuristic quality for automatic feature selection, and modified K-NN was used as the classification method that selected the K-NN and modified K-NN appropriately.

R. Suguna etc., [20] examined the regression analysis to analyze the relationship between the dependent and independent variables in the dataset and suggested the influence level of the independent variable on the result prediction using the Multiple Linear Regression method. The model with more than one predictor was created by identifying the statistical relationship with the model. The research evaluated and analyzed the efficiency of several Multiple Linear Regression models and advised the efficiency improvement of the best model with the feature selection. The data set used in this research was the student's academic performance dataset from Kaggle, and it was tested with the classifier: Logistic, KNN, Kernel SVM, and Naïve Bayes.

4. Research Methodology

This research aimed to determine the features affecting the learning efficiency on the higher education student performance evaluation dataset. Data from the students in the Faculty of Engineering and Faculty of Education, Academic Year 2019 were collected to forecast the student's semester performance. Data comprised of 33 attributes and 145 records from UCI Machine Learning Dataset [21].

The experimental settings were presented in Figure 1 which were studied from the four different feature selection techniques: Information Gain, Information Gain Ratio, Correlation Coefficient, and Chi-square, using the four data classification methods: K-Nearest Neighbor, Random Forest, Artificial Neural Network, and Linear Regression, to identify the method with the best efficiency; it means the method has the highest impact on the data classification of student learning efficiency.

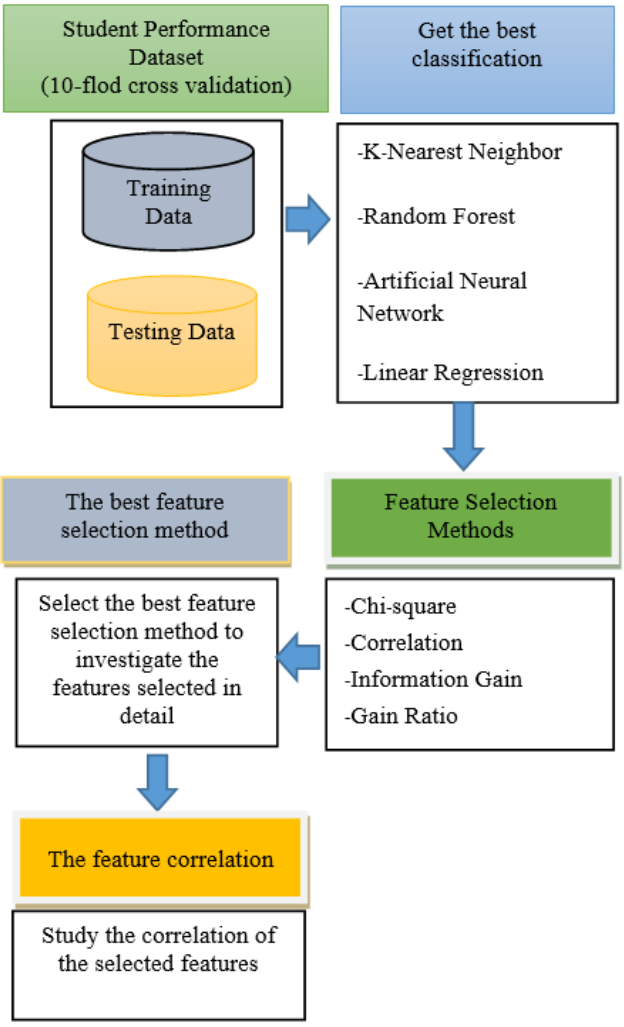

Figure 1. *Research Methodology.*

Then, the selected features from the best feature selection technique were examined by determining the relationship between the selected feature and the learning achievement to see the direction and factors affecting the learning efficiency.

5. Result and Discussion

The experiment looked for the features affecting the efficiency of learning efficiency classification with the different feature selection techniques: Information Gain, Gain Ratio, Correlation Coefficient, and Chi-square [2], [3] and studied the impact on the current four popular data classification methods: K-Nearest Neighbor, Random Forest, Artificial Neural Network, and Linear Regression by testing with the Weka program. The efficiency of four data classification methods was investigated, as shown in Table 1 by testing with the Higher Education Students Performance Evaluation dataset, which was collected from the students in the Faculty of Engineering and Faculty of Education, Academic Year 2019, to forecast the final semester performance. Data consisted of 33 attributes, and 146 records from eight classes: 0: Fail, 1: DD, 2: DC, 3: CC, 4: CB, 5: BB, 6: BA, and 7: AA, from UCI Machine Learning Dataset.

Then, the impact on the data classification using the feature selection was examined again, as shown in Table 2-4.

Table 1. *The efficiency of classification methods on student performance dataset with original data.*

Method	Accuracy
K-Nearest Neighbor	38.05
Random Forest	**70.44**
Artificial Neural Network	39.47
Linear Regression	49.56

The results from Table 1 demonstrated that the technique that gave the best learning efficiency for the student was the Random Forest method; the accuracy was 70.44%, followed by the Linear Regression, Artificial Neural Network and K-Nearest Neighbor, which the accuracy was 49.56, 39.47 and 38.05%, respectively. Consequently, it was used to study the impact of feature selection, as the results shown in Table 2.

Table 2. *The efficiency feature selections with Random Forest classification method.*

Feature selection method	Accuracy
All Features	70.44
Chi-square	74.22
Correlation Coefficient	74.89
Information Gain	**75.82**
Gain Ratio	**76.07**

The results in Table 2 illustrated that the feature selection technique that gave the best learning efficiency for the Random Forest method was the Gain Ratio technique, which precision was 76.07%, followed by the Information Gain, Correlation Coefficient, and Chi-square, at 75.82, 74.89, and 74.22% respectively. As a result, Gain Ratio and Information Gain techniques were utilized to test with all data classifiers, and the efficiency was compared to the feature selection. Results were shown in Table 3.

Table 3. *The efficiency feature selections with classification methods.*

Classification Method	All Feature	Gain-Ratio	Information Gain
K-Nearest Neighbor	38.05	**48.53**	41.61
Random Forest	70.44	**76.07**	75.82
Artificial Neural Network	39.47	34.71	**40.00**
Linear Regression	49.56	43.88	**54.37**

Table 3 revealed that the Gain-Ratio gave the best efficiency for the Random Forest and K-Nearest Neighbor methods, and Artificial Neural Network and Linear Regression gave less efficiency than using all features. Meanwhile, the Information Gain gave better efficiency to all methods if compared to the use of all features. However, if compared to the Gain Ratio, the Information Gain gave the best efficiency to

the Artificial Neural Network and Linear Regression method but a lower efficiency to the Random Forest and K-Nearest Neighbor method.

Therefore, the researcher picked the selected features using the Information Gain and Gain Ratio, and discovered that the key features that were selected included the mother's education, father's education, curriculum ID, last semester's accumulated grade point average, gender, expected accumulated grade point average, the impact of the project/activity toward the success, total income, age, weekly study hour, type of residence, frequency in reading non-scientific books, Frequency in reading scientific books, note taking, preparation for mid-term exam, and parental status.

Thus, the data classification was tested again, and the results confirmed that the selected features gave the better classification efficiency as shown in Table 4.

Table 4 illustrated the study results on the impacts of the feature selection technique using the Information Gain and Information Gain Ratio. It was found that when all features were classified with K-Nearest Neighbor, Random Forest, Artificial Neural Network, and Linear Regression method, the precision increased from 38.05 to 57.44%, 70.44 to 75.89%, 39.47 to 44.84%, and 49.56 to 51.04%, respectively. In short, the selected features gave a better efficiency in data classification.

Table 4. *The influence on selected feature with classification methods.*

Method	Accuracy			
	old	IG	IG-Ration	Selected feature
K-Nearest Neighbor	38.05	**41.64**	**48.53**	**57.44**
Random Forest	70.44	**75.82**	**76.07**	**75.89**
Artificial Neural Network	39.47	**40.00**	34.71	**44.84**
Linear Regression	49.56	**54.37**	43.88	**51.04**

Then, the relationship of the features was analyzed using Pearson's correlation technique, as shown in Table 5. It was found that a correlation of CUML_GPA, EXP_GPA, READ_FREQ, COURSE ID, KIDS, which were the features that had a positive relationship with the data class was 0.34, 0.32, 0.25, 0.14, and 0.07 respectively. Meanwhile, a correlation of the STUDY_HRS, AGE, SALARY, and IMPACT, which were the feature that had a negative relationship with the data was -0.03, -0.10, -0.17, and -0.20, respectively.

Analysis results of the data relationship of the features using Pearson's correlation technique shown in Table 5 indicated that the features with a positive relationship with the data class when considering the five features, which were CUML_GPA, EXP_GPA, READ_FREQ, COURSE ID, and KIDS, it meant that in each curriculum when the student had the high accumulated grade point average of the last semester, high expectation of accumulated grade point average, frequent reading of non-scientific books, and the divorced or died parents, the learning achievement was good.

If the features had a negative relationship with the data class, which were STUDY_HRS, AGE, SALARY, and IMPACT, which meant having low weekly study hours, young age, low total income, and positive impact of project/activity toward success, the learning achievement was good.

6. Conclusions

This research aimed to discover the features affecting the learning efficiency on the higher education student performance evaluation dataset. The data were collected from the student in the Faculty of Engineering and Faculty of Education, Academic year 2019 to forecast the final learning performance of the students. Data consisted of 33 attributes and 145 records from UCI Machine Learning Dataset.

Four feature selection techniques: Information Gain, Gain Ratio, Correlation Coefficient, and Chi-Square, were applied, along with four data classification methods, K-Nearest Neighbor, Random Forest, Artificial Neural Network, and Linear Regression, to determine the technique that had

Table 5. *The attributes correlation with classes.*

No.	Attributes	Meaning	Correlation
1	GENDER	Gender	0.34
2	CUML_GPA	Last semester accumulated grade point average	0.32
3	EXP_GPA	Expected accumulated grade point average	0.25
4	READ_FREQ	Frequency in reading non-scientific books	0.20
5	COURSE ID	Curriculum ID	0.14
6	KIDS	Parental status	0.07
7	MOTHER_EDU	Mother's education	0.07
8	FATHER_EDU	Father's education	0.06
9	NOTES	Note taking	0.05
10	LIVING	Type of residence	0.02
11	PREP_STUDY	Preparation for mid-term exam	0.01
12	READ_FREQ_SCI	Frequency in reading scientific books	0.003
13	STUDY_HRS	Weekly study hour	-0.03
14	AGE	Age	-0.10
15	SALARY	Total income	-0.17
16	IMPACT	Impact of project/activity on the success	-0.20

the highest impact on the data classification of student learning efficiency and the technique that had the highest impact on the student learning efficiency.

The results illustrated that the best feature selection technique was the Information Gain and Gain Ratio. Both techniques selected the same 16 features out of all selected features from each technique. They comprise of gender, last semester accumulated grade point average, expected accumulated grade point average, frequency in reading non-scientific books, curriculum ID, parental status, mother's education, father's education, note taking, type of residence, preparation for mid-term exam, frequency in reading scientific books, weekly study hour, age, total income, and impact of project/activity towards the success.

When analyzing the relationship of the features using Pearson's correlation technique shown in Table 5, the five features that had a positive relationship with the data class, when considering the five features: CUML_GPA, EXP_GPA, READ_FREQ, COURSE ID, and KIDS meaning in each curriculum when the student had the high accumulated grade point average of the last semester, high expectation of accumulated grade point average, frequent reading of non-scientific books, and the divorced or died parents, the learning achievement was good. At the same time, the features that had a negative relationship with the data class: STUDY_HRS, AGE, SALARY, and IMPACT, meaning having low weekly study hours, young age, low total income, and positive impact of project/activity towards success, the learning achievement was good.

All in all, the factors affecting learning efficiency could be determined, which included the accumulated grade point average, the expectation of graduation, frequency of reading non-scientific books, and low weekly study hours. These factors would be the guidelines for designing the learning management for the highest learning efficiency.

7. Acknowledgement

We are thankful for the research fund for the fiscal year 2021 from Thailand Research Fund, Thailand Science Research and Innovation (TSRI), and the research data from the research on The Development of Learning Innovation on Artificial Intelligence for Elementary Students.

8. References

[1] T. Shusaku, and H. Shoji, "Risk Mining in Medicine: Application of Data Mining to Medical Risk Management". *Fundamenta Informaticae,* Vol. 97, No. 1, pp.107 – 121, January, 2010.

[2] H. Nan-Chen, "An integrated data mining and behavioral scoring model for analyzing bank customers". *Expert Systems with Applications,* Vol. 27, No. 4, pp. 623-633, November, 2004.

[3] X. Zheng, G. Zhu, N. Metawa, and Q. Zhou, "Machine learning based customer meta-combination brand equity analysis for marketing behavior evaluation," *Information Processing and Management,* Vol. 59, No. 1, article 102800, 2022.

[4] S. Wang, "Smart Data Mining Algorithm for Intelligent Education." *Journal of Intelligent & Fuzzy Systems,* Vol. 37, No. 1, pp. 9-16, July, 2019.

[5] S. N. Alachiotis et al., "Supervised Machine Learning Models for Student Performance Prediction." *Intelligent Decision Technologies,* Vol. 16, No. 1, pp. 93-106, January, 2022.

[6] O.O. Oladipupo, and O.O. Olugbara, "Evaluation of Data Analytics Based Clustering Algorithms for Knowledge Mining in a Student Engagement Data." *Intelligent Data Analysis,* Vol. 23, No. 5, pp. 1055-1071, January, 2019.

[7] A. Jha, M. Dave, and S. Madan, "A review on the study and analysis of big data using data mining techniques," *International Journal of Latest Trends in Engineering and Technology,* Vol. 6, No. 3, pp. 94–102, 2016.

[8] I. Guyon, and E. André, "An introduction to variable and feature selection." *Journal of machine learning research,* pp. 1157-1182, March, 2003.

[9] N. Rachburee, and W. Punlumjeak. "A comparison of feature selection approach between greedy, IG-ratio, Chi-square, and mRMR in educational mining." *In: 2015 7th international conference on information technology and electrical engineering (ICITEE),* pp. 420-424, 2015.

[10] L. Jia, "A hybrid feature selection method for software defect prediction." *IOP Conference Series: Materials Science and Engineering,* Vol. 394. No. 3, IOP Publishing, 2018.

[11] S. Li, X. Zhang, M. Zong, X. Zhu, and D. Cheng, "Learning k for knn classification." *ACM Transactions on Intelligent Systems and Technology (TIST)*, Vol. 8, No. 3, pp. 1-19, 2017.

[12] C. R. Sekhar, and E. Madhu, "Mode choice analysis using random forest decision trees." *Transportation Research Procedia,* Vol. 17, pp. 644-652, 2016.

[13] Y. Singh Y, and A.S. Chauhan, "Neural Networks in Data Mining." *Journal of Theoretical and Applied Information Technology,* Vol. 5, No. 1, pp. 37-42, 2009.

[14] M. Khashei, A. Z. Hamadani, and M. Bijari, "A novel hybrid classification model of artificial neural networks and multiple linear regression models." *Expert Systems with Applications,* Vol. 39, No. 3, pp. 2606-2620, 2012.

[15] S. Makridakis, "Accuracy measures: theoretical and practical concerns." *International journal of forecasting,* Vol. 9, No. 4, pp. 527-529, 1993.

[16] W. Punlumjeak, and N. Rachburee, "A comparative study of feature selection techniques for classify student performance." *2015 7th International Conference on Information Technology and Electrical Engineering (ICITEE)*, pp. 425-429, 2015.

[17] L. Rahman, N. A. Setiawan, and A. E. Permanasari, "Feature selection methods in improving accuracy of classifying students' academic performance." *2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, pp. 267-271, 2017.

[18] M. R. Ahmed, S. T. I. Tahid, N. A. Mitu, P. Kundu, and S. Yeasmin, "A Comprehensive Analysis on Undergraduate Student Academic Performance using Feature Selection Techniques on Classification Algorithms." *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1-6, 2020.

[19] M. Wafi, U. Faruq, and A. A. Supianto, "Automatic Feature Selection for Modified K- Nearest Neighbor to Predict Student's Academic Performance."

2019 International Conference on Sustainable Information Engineering and Technology (SIET), pp. 44-48, 2019.

[20] R. Suguna, M. D. Shyamala, A. B. Rupali, and S. J. Aparna, "Assessment of feature selection for student academic performance through machine learning classification." *Journal of Statistics and Management Systems,* Vol. 22, No. 4, pp. 729-739, 2019.

[21] University of California, Irvine, School of Information and Computer Sciences. *UCI Machine Learning Repository.* Available Online at https://archive.ics.uci.edu/ml. accessed on 10 September 2022.