JOIV : Int. J. Inform. Visualization, 7(4) - December 2023 2460-2468

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Case Study: Using Data Mining to Predict Student Performance Based on Demographic Attributes

Nursyuhadah Alghazali binti Muhammad Zahruddin [a], Nur Diyana Kamarudin [a,b,*], Ruzanna Mat Jusoh [a], Nur Aisyah Abdul Fataf [b], Rahmat Hidayat [c]

[a] *Faculty of Defense Science and Technology, National Defense University of Malaysia, Sungai Besi, Kuala Lumpur, Malaysia*
[b] *Cyber Security and Digital Revolution Industry Centre, National Defense University of Malaysia, Sungai Besi, Kuala Lumpur, Malaysia*
[c] *Department of Information Technology, Politeknik Negeri Padang, West Sumatera, Indonesia*

*Corresponding author: *nurdiyana@upnm.edu.my*

***Abstract*—This study predicts student performance at Universiti Pertahanan Nasional Malaysia (UPNM) based on their socio-demographic profile; it also determines how a prediction algorithm can be used to classify the student data for the most significant demographic attributes. The analytical pattern in academic results per batch has been identified using demographic attributes and the student's grades to improve short-term and long-term learning and teaching plans. Understanding the likely outcome of the education process based on predictions can help UPNM lecturers enhance the achievements of the subsequent batch of students by modifying the factors contributing to the prior success. This study identifies and predicts student performance using data mining and classification techniques such as decision trees, neural networks, and k-nearest neighbors. This frequently adopted method comprises data selection and preparation, cleansing, incorporating previous knowledge datasets, and interpreting precise solutions. This study presents the simplified output from each data mining method to facilitate a better understanding of the result and determine the best data mining method. The results show that the critical attributes influencing student performance are gender, age, and student status. The Neural Networks method has the lowest Root of the Mean of the Square of Errors (RMSE) for accuracy measurement. In contrast, the decision tree method has the highest RMSE, which indicates that the decision tree method has a lower performance accuracy. Moreover, the correlation coefficient for the k-nearest neighbor has been recorded as less than one.**

***Keywords*— Demographic profiling; student performance prediction; UPNM; WEKA; data mining; knowledge discovery database.**

Manuscript received 15 Aug. 2023; revised 9 Oct. 2023; accepted 21 Nov. 2023. Date of publication 31 Dec. 2023.

I. INTRODUCTION

Higher education institutions constantly strive to improve students' academic achievement to ensure their graduates are qualified professionals when they enter the job market. Students must have fundamental rational and numerical skills, such as originality, flexibility, curiosity, observation skills, analytical capability, the willingness to seek new knowledge, and be proactive in finding solutions to problems in their chosen professions [1]. Demographic profiling can reveal the crucial determiners of student achievement and accomplishment [2]. Student performance can be predicted using several data mining methods and algorithms, either classification or clustering methods.

One of the Knowledge Discovery Database (KDD) processes is Data Mining [1], which identifies the unseen data and the pattern in a large volume of data collected from multiple sources, such as data warehouses and data marts. Data mining secretly pulls out data and provides vital information to make appropriate judgments. Data pre-processing involves information cleansing to reduce noise, conducting relevant research to reduce insignificant features, and improving the accuracy of forecasts, flexibility, and controllability [3]. Data mining is a method for extracting knowledge from large volumes of data. It is primarily used to collect the required data to achieve better results. This process includes quantitative and numerical models. Data mining involves exploring collected data, including text, online, two-image processing, and graphics [2]. The most crucial stage is obtaining information through knowledge discovery. Extraction of relevant data involves several steps [2], [3]. It combines techniques from various fields, including image processing, e-commerce, retail, and mining patterns. Data mining is beneficial in education data mining (EDM), a field for discovering information from extensive data on education [4]. The objective of EDM is to

identify educational data patterns and enhance the quality of education. EDM is the education research study of many techniques that assess economic, cultural, social, personal, geographical, environmental, and other elements [4]. In this research, education data mining helps educators predict student performance.

Population research based on age, race, sex, and other variables is a demographic analysis. Demographic data relates to statistically stated socioeconomic information, including jobs, education, income, marriage, birth, and mortality rates [30]. The demographic attributes in this research are marital status, gender, UPNM student status (civilian or cadet officer), and whether the student is from a rural or urban area. Other demographic attributes are race (Malay, Chinese, Indian, and others) and whether the student is from Peninsular Malaysia or Sabah and Sarawak.

The indicator for student achievement in the higher education system is the Cumulative Grade Point Average (CGPA), which measures the overall cumulative student performance throughout the enrolled academic session. This study uses a 3.0 CGPA and student demographic profile as student performance indicators each semester. The students' study status also indicates their educational performance. This study considers students with a 2.75 CGPA to pass the semester, while those with a CGPA less than 2.75 fail. Because university educators play a critical role in ensuring good student performance, this research analyzes the fundamental factors making learning easier [5], [6]. Besides being transparent regarding administration, performance prediction must also be implemented to provide evidence gathered during the learning process (for example, the metadata for education material, contextual data, data on student engagement, and demographics) [7]. University educators need help to observe student performance easily. Therefore, this study analyses the student demographic attributes and five achievement data to identify the positive patterns in the student study routine and formulate measures to improve their grades [8]. Moreover, it is impossible to manually observe the main factors that could contribute to decreasing or rising grades from students.

This study predicted student performance by analyzing students' demographic attributes and their grades and CGPA each semester employing the Neural Network and Decision Tree from the supervised learning in Machine Learning methods and the k-Nearest Neighbor (k-NN) from unsupervised method [9], [10], [11]. Student performance is the primary indicator of achievement in higher education institutions [12], and poor student performance could tarnish the reputation of higher-learning institutions. Therefore, it is essential to implement measures that ensure good student performance. This study aimed to provide the UPNM management, lecturers, and students with fundamental insights for improving student performance by using analytical measurement and data mining methods to determine the optimum student-centered learning method and provide a conducive learning environment in UPNM.

A. Literature review

Yaacob *et al.* [13] used supervised techniques in data mining, namely the Naïve Bayes, Logistic Regression Model, Decision Tree, and k-nearest Neighbor methods, to predict student performance and identify the fundamental data. The precision, precision gauge, ROC curve, and findings showed that Naïve Bayes produced better results than the other classification techniques. Tomasevic *et al*. [9] considered previous student performance, demographic data, and student engagement as the parameters for predicting student performance using Support Vector Machines (SVM), Decision Trees,), Naïve Bayes, Artificial Neural Networks (ANN), Logistic Regression and k-Nearest Neighbors (k-NN). The accuracy was reduced by Artificial Neural Networks (ANN) for both the categorization and regression assignments through the input of student engagement data and previous performance data. In comparison, demographics were used to demonstrate no significant impact on the performance prediction accuracy. Yang and Li [7] employed the classification in Back Propagation Neural Network (BP-NN) to propose an undergraduate attribute matrix. The student performance predictor can estimate undergraduate students' attribute points, such as subject scores and learning skills grades. Malini and Kalpana [14] used boosting, bagging, and artificial neural networks (ANN) to predict student performance based on the student's grades, social and demographic attributes, and school-related factors. This study has shown that economic status influences student achievement. The MLP had a 72% accuracy, 88% on Bagging classifiers, and 86% accuracy with economic background features from MultiBoost classifiers. The higher accuracy showed that economic background influenced the students' learning behavior.

Khan *et al*. [15] utilized Random Wheel Classifiers, Artificial Neural Networks, and Naïve Bayes in their research. To predict student performance, they considered estimated teaching quality, scoring ease, past backlogs, student quality, and domain knowledge. The suggested classifiers successfully estimated more than 80% of failures and successes. Baradwaj and Pal [16] predicted student performance using the ID3, Naïve Bayes, SVM, and C4.5 methods and found that SVM is the most accurate classification method.

Widyahastuti and Tjhin [17] examined student demographic attributes using the Rule Based, Naïve Bayes, and Decision Tree to improve student standards and ability through active engagement. The performance predictions helped weaker students to recognize the difficulty of courses. Amrieh *et al*. [2] employed Artificial Neural Networks, Boosting Bagging, Naïve Bayesian, Random Forest, and Decision Tree to determine academic achievement and students' attitudes. The study showed that the suggested model accuracy using attitude features gained up to 22.1% improvement. The accuracy was 80% when assessing newcomer students. Moreover, [18] use Fuzzy Soft Set Classification (FSSC) reached up to accuracy results to be able to detect students at risk in the early stages of education. So, the study found that higher education can minimize students not graduating on time or dropout by providing appropriate treatment and designing strategic programs.

II. MATERIALS AND METHOD

A. Data Preparation

This research conducted data mining using the database knowledge discovery (KDD) methodology. KDD is the

process of discovering practical data-gathering knowledge and comprises data selection, preparation, purification, incorporating previous dataset knowledge, and interpretation of precise solutions. The datasets consist of 97 instances of the students in the Department of Science Computer, Faculty of Defense Science and Technology, UPNM. The dataset has nine demographic attributes and performance measures: age, gender, race, marital status, student status, residential area, study status, geographic location, and CGPA.

The first step in data processing involves cleaning, which aims to identify and rectify or remove inaccurate or inconsistent data within a dataset, index, or database. This research employed the Excel platform to clean the data manually, following steps such as eliminating duplicates and replacing missing values with the average points derived from the corresponding column. The subsequent step is merging data from multiple sources to ensure reliable data access and transfer across various topics and configurations. Afterward, the process continues with data selection, which involves determining and retrieving the pertinent data for analysis. This study guided data selection by Neural Networks, Decision Trees, and k-nearest Neighbor (k-NN) techniques, providing recommendations for selecting relevant data. Data transformation converts or compacts the data by executing summary and aggregate operations in the mining formats appropriate for the samples. The datasets were transformed into numerical attributes to make them readable in WEKA. The demographic characteristics are gender, residential area, race, state location, student status, marital status, CGPA, and study status. Table 1 presents the transformation data.

Data mining is forecasting the results by distinguishing the abnormalities, forms, and associations in large datasets. The data mining techniques used in this research are the K-Nearest Neighbor, Neural Networks, and Decision Tree; the software for analyzing the data to predict student performances is the Waikato Environment for Knowledge Analysis (WEKA). Pattern evaluation identifies the patterns expressing knowledge using the provided measurements. Knowledge presentation uses visualization tools to describe strategies and present the outcomes of data mining. It involves generating statements, plotting graphs, establishing discriminant regulations, and categorizing the guidelines and regulations.

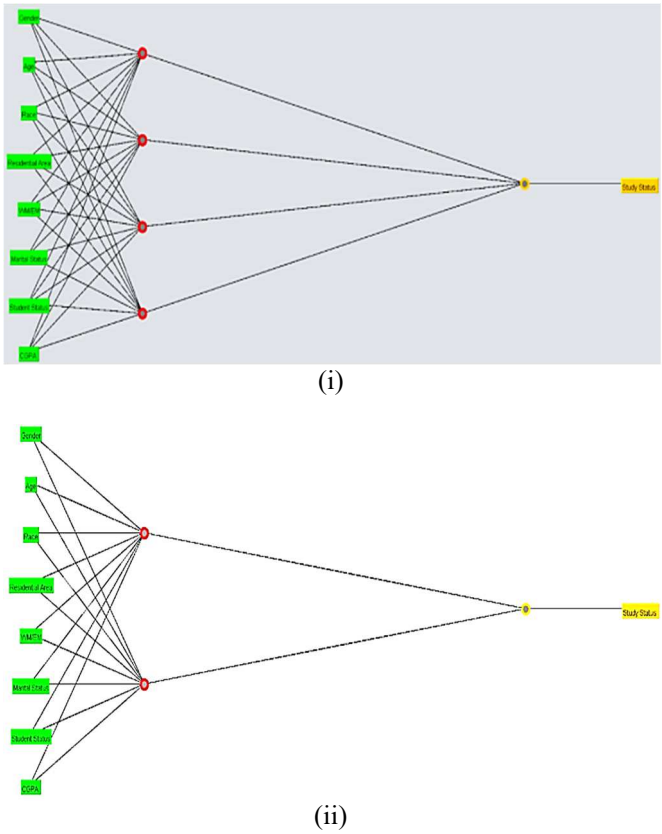
TABLE I
DATA TRANSFORMATION

Data	Integer Datatype
Gender {male}	100
Gender {female}	200
Residential area {city}	201
Residential area {rural}	202
Race {Malay}	101
Race {Chinese}	102
Race {Indian}	103
Race {Others}	104
State Location {Peninsular Malaysia}	1
State Location {Sabah and Sarawak}	2
Student status {cadet officer}	33
Student status {civilian}	43
CGPA {< 2.75}	10
CGPA {> 2.75}	20
Marital status {single}	10
Study status {pass}	1001
Study status {fail}	1002

III. RESULTS AND DISCUSSIONS

A. Neural Networks

Artificial Neural Networks (ANNs) consist of node layers, each having an input and output layer and one or more unknown layers. Each node connects to another node and has a weight and threshold value. A node is triggered if the result exceeds a threshold value and transmits the data to the next level. In these classifiers, Multi-Layered Perceptron (MLP) formed in linked levels. The input level brings in the response patterns. The output layer consists of classification or output indicators represented by input samples. The network discovers by analyzing individual data, predicts each data, and modifies the weights if the prediction is incorrect. This process is repeated several times, and the network continues improving its forecasts until one or fewer uncertain conditions are reached.



(i)

(ii)

Fig. 1 (i) Architecture Layer of Neural Network Output 1; (ii) Architecture Layer of Neural Network Output 2

Figure 1(i) shows the results employed in the next layer of the neural networks. The nodes with heavier weights indicate a more significant contribution by the attributes. This technique has four hidden layers and four assigned nodes. The threshold of Node 0 is 0.4916, and the initial weight of Node 1 is -1.769, Node 2 is 1.178, Node 3 is -0.036, and Node 4 is -0.486. The primary focus of the test mode is the Study Status, Pass or Fail. The threshold value for Node 1 is -0.0157; the demographic characteristics exceeding this threshold value are gender, age, residential area, student status, and CGPA, indicating that they are the most influential attributes in Node 1. The threshold value for Node 2 is -0.0036; the demographic attributes exceeding this threshold value are gender, age, race,

marital status, and student status. Node 3 has a threshold value of -0.279. The analysis showed that the values for gender, age, race, marital status, student status, residential area, state location, and CGPA are above the threshold value. Similarly, the demographic and performance attributes have values higher than the - 0.166 threshold value for Node 4.

TABLE II
SUMMARY OF THE NEURAL NETWORK OUTPUT 1

Summary	
Correlation Coefficient	1
Mean absolute error	0.0004
Root mean squared error	0.0006
Relative absolute error	0.1009%
Root relative squared error	0.1363%
Total number of instances	97

TABLE III
SUMMARY OF NEURAL NETWORK OUTPUT 2

Summary	
Correlation coefficient	1
Mean absolute error	0.0001
Root mean squared error	0.0002
Relative absolute error	0.0285%
Root relative squared error	0.0431%
Total number of instances	97

TABLE IV
SUMMARY OF NEURAL NETWORK OUTPUT 1

Classifier model (complete training set)	
Linear Node 0	
Inputs	**Weights**
Threshold	0.19463625347114
Node 1	-1.678892747987823
Node 2	1.177531256577951
Node 3	-0.036088061391542255
Node 4	-0.486481985148401
Sigmoid Node 1	
Inputs	**Weights**
Threshold	-0.01568404904205123
Gender	-0.00312883786019948
Age	0.0432760097560809
Race	-0.04927478698544483
Residential Area	-0.01338502461394065
WM/EM	-0.0907835812693506
Marital Status	-0.04708924872783642
Student Status	0.04057432079317868
CGPA	1.8650853127915232
Sigmoid Node 2	
Inputs	**Weights**
Threshold	-0.0036759384361545583
Gender	0.0034376799558019536
Age	0.1276323100144554
Race	0.0013795697693746109
Residential Area	-0.0083594463353499
WM/EM	-0.05935137855675985
Marital Status	-0.0016329809894899822
Student Status	0.03397599586950739
CGPA	-1.1027923273449178
Sigmoid Node 3	
Inputs	**Weights**
Threshold	-0.2793520743480427
Gender	0.22425259305861237
Age	0.22058504861237
Race	0.24255716507127298
Residential Area	-0.11870716316688698
WM/EM	0.1272793423797521
Marital Status	0.04719501209537452
Student Status	0.12204991873302977
CGPA	0.417091712081249
Sigmoid Node 4	
Inputs	**Weights**
Threshold	-0.16574860884809206
Gender	-0.004132928929352015
Age	0.17073634385903924
Race	0.17073634385903924
Residential Area	0.013874870521445619
WM/EM	0.03839982138436964
Marital Status	0.048911715075140555
Student Status	-0.011320528541976684
CGPA	0.787829744343726

Compared to the neural network techniques tested with the hidden layer, Table 5 used a neural network output with two hidden layers with similar ten-fold cross-validation. The RMSE of two hidden layers is less than 0.0002, while the values for four hidden layers are 0.0006, indicating a higher accuracy of the method. The neural network with two hidden layers in the first node has a threshold value of -0.138, which displays that all the attributes pass the threshold value. Node 2 has a threshold value of -0.214. The weight of all characteristics exceeded the attributes of the threshold values and thus contributed to the prediction. The hidden layers with higher weights indicate the more significant contribution of the characteristics. In other words, the hidden layers show the more accurate and precise neural networks method, where the characteristics' contribution increases as the weights of the hidden layer increase.

TABLE V
SUMMARY OF NEURAL NETWORK OUTPUT 2

Classifier model (complete training set)	
Linear Node 0	
Inputs	**Weights**
Threshold	1.3127788461512266
Node 1	-2.006415105349033
Node 2	-0.7221785464954951
Sigmoid Node 1	
Inputs	**Weights**
Threshold	-0.138330427801449
Gender	0.0024422870954161832
Age	-0.016550117708939317
Race	-0.0136560117708939317
Residential Area	-0.009744188878761717
WM/EM	-0.05412689389521921
Marital Status	0.022429503357886876
Student Status	0.00974005150793991
CGPA	2.2681718613926494
Sigmoid Node 2	
Inputs	**Weights**
Threshold	-0.2135816291708891
Gender	-0.002754242948901339
Age	0.019980977027931147
Race	0.016842127646253962
Residential Area	.011433596818833876
WM/EM	0.0715397088841275
Marital Status	0.01773541612498744
Student Status	-0.011797327012342535
CGPA	1.1583086724850309

Node 2 has a threshold value of -0.214. The weight of all characteristics exceeded the attributes of the threshold values and thus contributed to the prediction. The hidden layers with higher weights indicate the greater contribution of the

characteristics. In other words, the hidden layers show the more accurate and precise neural networks method, where the characteristics' contribution increases as the weights of the hidden layer increase.

In summary, the weights of all nodes indicate that gender, age, and student status contributed equally. These demographic characteristics are the most influential factors in predicting student performance [9], [19]–[24]. The nodes in a neural network can process numerical, algebraic, or signal data flow. Neural networks are excellent approximators of generic functions and often outperform other prediction methods [25]–[28], sometimes significantly. One only needs mathematical and statistical understanding to train or use neural networks. WEKA has several attributes that prevent some of the potential drawbacks of neural networks, including instantaneously discovering an adequate connectivity topology and sensitivity analysis (as shown in the variable importance chart) to aid in network interpretation, pruning, and validation to preclude overfitting.

B. Decision Tree

Decision Tree is a supervised learning method for solving regression and classification problems, although it is most frequently used to explain classification jobs [29]–[31]. The internal nodes in the tree-structured classifier contain dataset attributes, and the branches correspond to the decision rules. Each leaf node represents the result. The key challenge when using a Decision Tree is identifying the characteristics of the root node at each level. The test mode of the Decision Tree model was split by 20% of the test. The size of the decision tree is 11. Figure 2(i) shows that the characteristics of the Decision Tree are race, gender, age, and residential area. CGPA ≥ 2.75 indicates 20, which means the students passed the exam, while ≤ 2.75 shows 10 in the output as the students are Fail.

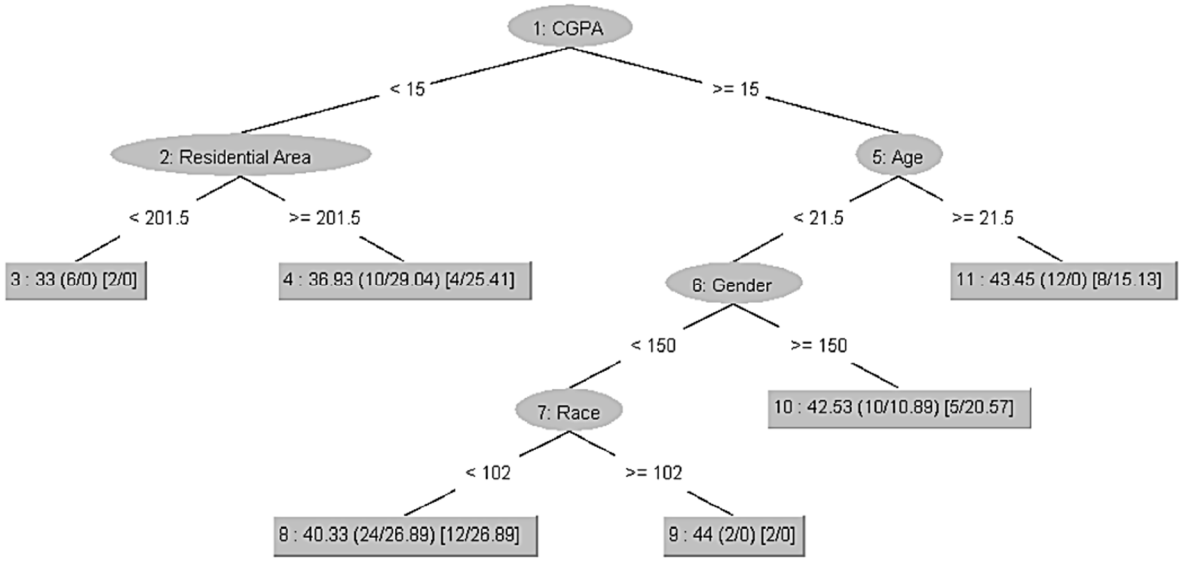


Fig. 2 (i) Decision Tree Output

TABLE VI
SUMMARY OF THE DECISION TREE OUTPUT

Summary	
Correlation Coefficient	0
Mean absolute error	4.9804
Root mean squared error	5.1604
Relative absolute error	100%
Root relative squared error	100%
Total number of instances	78

Four students who failed the semester are from rural areas, and three are from urban areas. The seven students who passed the semester are ≥ 21 years old. Most students who pass the semester are male students less than 21 years old. Moreover, most students who passed the semester are Malay, given that there are many of the students in UPNM. Most Indian and Chinese students and those from other races passed the semester. These results are crucial in predicting student performances based on demographic characteristics. The RMSE for the Decision Tree technique is high because it did not detect all attributes, which means it has lower accuracy, as shown in Figure 3(ii).

```
=== Classifier model (full training set) ===


REPTree
============

CGPA  < 15
|   Residential Area < 201.5 : 33 (6/0) [2/0]
|   Residential Area >= 201.5 : 36.93 (10/29.04) [4/25.41]
CGPA  >= 15
|   Age < 21.5
|   |   Gender < 150
|   |   |   Race < 102 : 40.33 (24/26.89) [12/26.89]
|   |   |   Race >= 102 : 44 (2/0) [2/0]
|   |   Gender >= 150 : 42.53 (10/10.89) [5/20.57]
|   Age >= 21.5 : 43.45 (12/0) [8/15.13]

Size of the tree : 11
```

Fig. 3 (ii) Decision Tree Output

C. k-Nearest Neighbor

There are four ways to calculate the distance between data points and their closest neighbor [32], [33]: the Euclidean distance, Hamming distance, Manhattan distance, and Minkowski distance. The Euclidean distance (y) is the most frequently employed distance function or metric and is computed using the sum of the squared difference of the square root between the new point (x) and an old point (y). The k-value in this method is 3. The lower value was chosen as the optimal k value because it could be affected by the outliers and noise, and thus, there is a high probability of overfitting. The larger k-values often give a smoother outcome boundary, but the value should be manageable since the classes with smaller datasets will be outnumbered by the other datasets. Furthermore, large k values are computationally costly. In the output of the k-NN from Figure 4, Figure 5, Figure 6, Figure 7, Figure 8, Figure 9, Figure 10, to Figure 11. The summary of the output can be seen in Table 7.

TABLE VII
SUMMARY OF THE K-NEAREST NEIGHBOR OUTPUT

Summary	
Correlation coefficient	0.9892
Mean absolute error	0.0089
Root mean squared error	0.0622
Relative absolute error	2.4812%
Root relative squared error	14.5201%
Total number of instances	98

The Integer Datatype 1000 on the y-axis represents pass, while 1002 represents fail. Figure 4 is the plot for gender vs the predicted study status, where variable 100 on the x-axis represents male, and variable 200 is female. The output shows that more male students pass the semester than females. Most male students did not pass the semester, but no female students failed. Figure 5 shows that most students who passed the semester are ≤ 26 years old, and a few are ≤ 30 years old. Most students who failed the semester were ≤ 26 years old. Figure 6 shows that most students who passed the semester were Malay, followed by Indian, Chinese, and other races. However, the figure also shows that most Malay students failed the semester. Figure 7 shows that students who failed the semester were from rural areas. Figure 8 shows that most students who passed the semester are from Peninsular Malaysia.

Figure 9 shows that all students who passed the semester are single because there were no married students for the dedicated semester. Figure 10 shows that fewer civilians failed the semester than cadet officers. The correlation coefficient of the k-nearest Neighbor technique was less than 1, and the MAE and RMSE were low. On the other hand, the method's accuracy is recorded as second place in accuracy.

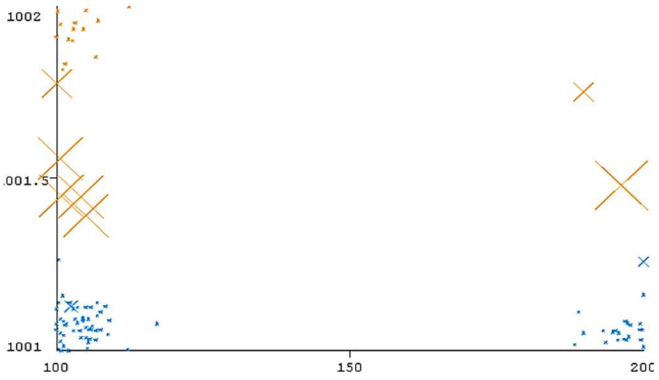


Fig. 4 Gender (x) vs the predicted study status (y) by the k-Nearest Neighbor Output

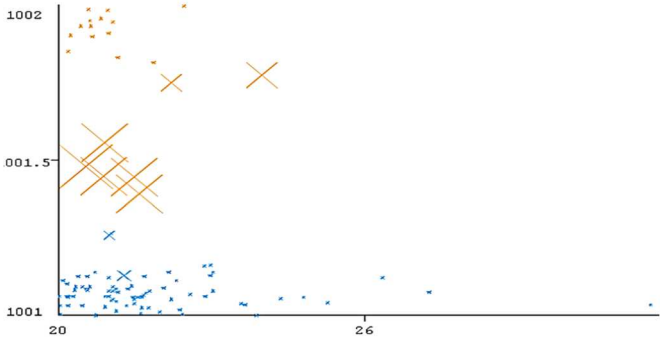


Fig. 5 Age (x) vs the predicted study status (y) by k-Nearest Neighbor Output

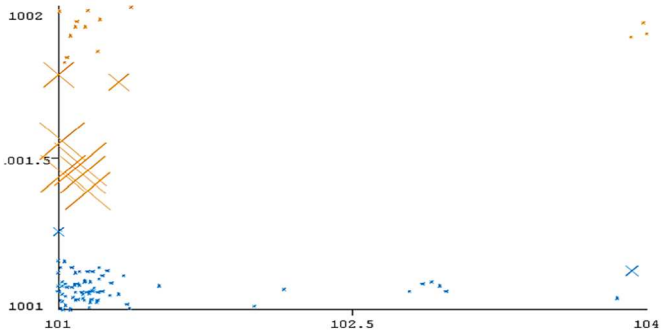


Fig. 6 Race (x) vs the predicted study status (y) by k-Nearest Neighbor Output

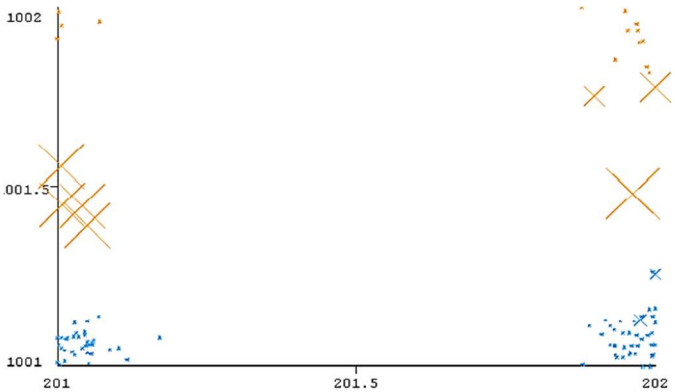


Fig 7 Residential Area (x) vs. the predicted study status (y) by the k-Nearest Neighbor Output

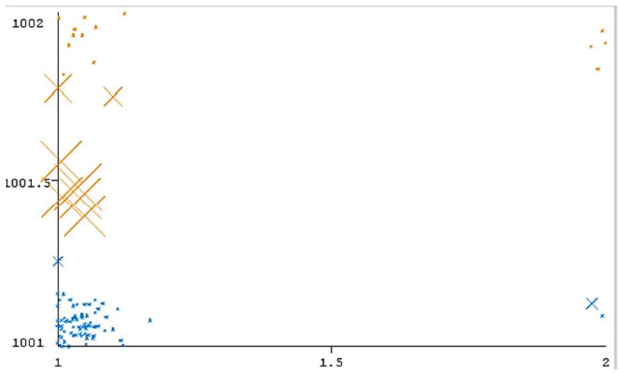


Fig. 8 Peninsular Malaysia/Sabah and Sarawak (x) vs the predicted study status (y) by the k-Nearest Neighbor Output

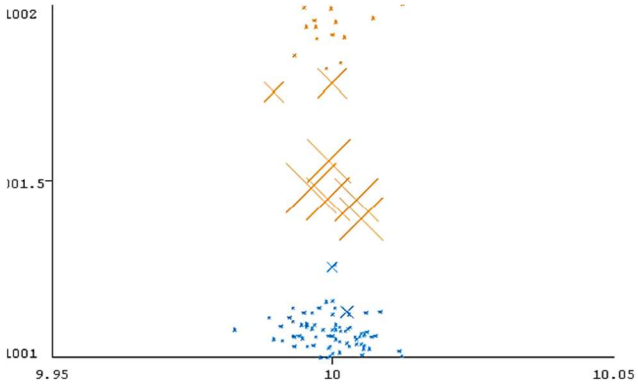


Fig. 9 Marital status vs the predicted study status by k-Nearest Neighbor Output

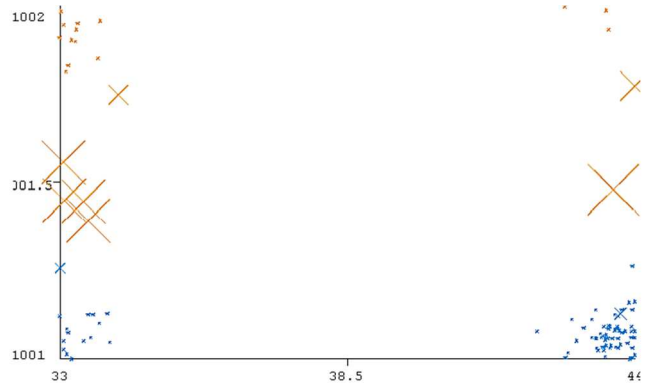


Fig. 10 Student status (x) vs the predicted study status (y) by the k-nearest Neighbor Output

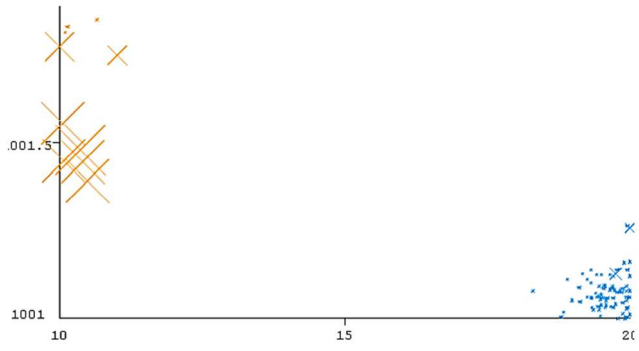


Fig. 11 Cumulative Grade Pointer Average (CGPA) (x) vs the predicted study status (y) by the k-Nearest Neighbor Output

This study measured the accuracy of each data mining method. The regression analysis used the Mean Squared Error, Mean Absolute Error, Root Mean Squared Error, and R-squared or Coefficient of the purpose metrics to evaluate the model performance. The values predicted by a well-fitting regression model are close to the actual values. The Mean Absolute Error represents the standard of the absolute distinction between the real and predicted values in the dataset. It computes the mean of the residuals in the dataset. Mean Squared Error is the median of the squared difference between the dataset's initial and predicted values and shows the variance of the residuals. The square root of the Mean Squared Error is the Mean Squared Error, which gives the residuals' standard deviation. Therefore, this study determined the accuracy of the data mining approaches using the Root Mean Squared Error (RMSE). Low MAE, MSE, and RMSE values indicate a more precise regression model, while smaller RMSE values indicate higher model accuracy. The Root Mean Square Error (RMSE) shows how accurately the model forecasts the reaction and is the most crucial fit criterion if the primary objective is prediction. The accuracy of the Neural Networks, k-nearest Neighbor, and Decision Tree was determined using the RMSE as in Table 8.

TABLE VIII
RMSE OF THE DATA MINING METHODS

Method	RMSE
Neural Networks	0.0006
k-Nearest Neighbor (k-NN)	0.0622
Decision Tree	5.1604

The accuracy of the data mining methods is crucial to ensure that the regressions are well-fitting models. The output of the models showed that each data mining method has a specific way of identifying, classifying, and analyzing the data. The Neural Networks method is a computer system with interconnected nodes that operate like the nerve cells in the human brain. It uses algorithms to recognize the unseen patterns and relationships in raw data, group and classify the datasets, and continuously discover and enhance the model over time. The Decision Tree method uses internal nodes in the tree-structured classifier containing the dataset attributes, where the branches represent decision rules, and each leaf node represents the result. The k-nearest Neighbor (k-NN) algorithm is a simple supervised machine learning algorithm that explains classification and regression problems. The RMSE of the data mining methods showed that Neural Networks predicted student performance with the highest accuracy. It classified most demographic attributes that influenced student performance. Gender, age, and student status attributes exceeded the threshold values for each node. K-Nearest Neighbor (k-NN) has the second-highest accuracy. It breaks down the data using the predicted study status as the independent variable. The remaining nine attributes were the dependent variables. The RMSE of the k-Nearest Neighbor is lower than the Decision Tree method because it used a grouping method to categorize the students. The Decision Tree method was the least accurate because it only predicted a few demographic attributes and classified selected datatypes in the datasets at the leaf nodes.

IV. Conclusion

This study used data mining methods from supervised machine learning, namely Decision Tree and Neural Network, and unsupervised machine learning, k-Nearest Neighbor (k-NN), to predict student performance. The data mining methods have different ways of predicting student performance. The Neural Networks method predicted most of the demographic variables influencing student performances. The k-Nearest Neighbor gave a more accurate student performance using 44 k-values and data types. The Decision Tree method failed to predict student performance comprehensively because of its limited ability, although it predicted the most influential variables in the datasets. Each method has a unique way of classifying and predicting the student's performances. The RMSE of the data mining methods indicates their performance accuracy. Neural Networks has the highest, followed by k-Nearest Neighbor (k-NN) and Decision Tree.

The data mining methods employed in this study have their advantages and disadvantages. The Neural Networks and Decision Tree methods are unsupervised learning methods which allow the datasets to act without supervision. It is a type of machine learning that trains prototypes using unlabeled datasets. The k-Nearest Neighbor method is a supervised machine learning method. This study ran all data mining methods using the WEKA analytical software.

The Neural Networks method had the lowest RMSE value and was the most accurate, followed by the k-Nearest Neighbor method, which predicted student performance for each characteristic by clustering the data values using the k values. The Decision Tree method has the poorest accuracy because it only predicted a few demographic characteristics and classified only selected data types in the datasets at the leaf nodes. The accuracy of these models showed their abilities to classify, cluster, and analyze when making predictions.

Acknowledgment

The authors thank the Faculty of Defense Science and Technology (FSTP) and UPNM for the financial support and research facility. The authors declare no conflict of interest regarding the publication of this paper.

References

[1] Q. A. Al-Radaideh, E. M. Al-Shawakfa, and M. I. Al-Najjar, "Mining student data using decision trees," in *International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan*, 2006.

[2] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Mining Educational Data to Predict Student's academic Performance using Ensemble Methods," *Int. J. Database Theory Appl.*, vol. 9, no. 8, pp. 119–136, Aug. 2016, doi: 10.14257/ijdta.2016.9.8.13.

[3] S. Ayesha, T. Mustafa, A. R. Sattar, and M. I. Khan, "Data mining model for higher education system," *Eur. J. Sci. Res.*, vol. 43, no. 1, pp. 24–29, 2010.

[4] B. K. Bhardwaj and S. Pal, "Data Mining: A prediction for performance improvement using classification," *arXiv Prepr. arXiv1201.3418*, 2012.

[5] Z. N. Khan, "Scholastic Achievement of Higher Secondary Students in Science Stream.," *Online Submiss.*, vol. 1, no. 2, pp. 84–87, 2005.

[6] V. Ramesh, P. Parkavi, and K. Ramar, "Predicting Student Performance: A Statistical and Data Mining Approach," *Int. J. Comput. Appl.*, vol. 63, no. 8, pp. 35–39, Feb. 2013, doi: 10.5120/10489-5242.

[7] F. Yang and F. W. B. Li, "Study on student performance estimation, student progress analysis, and student potential prediction based on data mining," *Comput. Educ.*, vol. 123, pp. 97–108, Aug. 2018, doi:10.1016/j.compedu.2018.04.006.

[8] W. Xing, R. Guo, E. Petakovic, and S. Goggins, "Participation-based student final performance prediction model through interpretable Genetic Programming: Integrating learning analytics, educational data mining and theory," *Comput. Human Behav.*, vol. 47, pp. 168–181, Jun. 2015, doi: 10.1016/j.chb.2014.09.034.

[9] N. Tomasevic, N. Gvozdenovic, and S. Vranes, "An overview and comparison of supervised data mining techniques for student exam performance prediction," *Comput. Educ.*, vol. 143, p. 103676, Jan. 2020, doi: 10.1016/j.compedu.2019.103676.

[10] M. Pandey and S. Taruna, "Towards the integration of multiple classifier pertaining to the Student's performance prediction," *Perspect. Sci.*, vol. 8, pp. 364–366, Sep. 2016, doi:10.1016/j.pisc.2016.04.076.

[11] K. David Kolo, S. A. Adepoju, and J. Kolo Alhassan, "A Decision Tree Approach for Predicting Students Academic Performance," *Int. J. Educ. Manag. Eng.*, vol. 5, no. 5, pp. 12–19, Oct. 2015, doi:10.5815/ijeme.2015.05.02.

[12] A. Gonzalez-Nucamendi, J. Noguez, L. Neri, V. Robledo-Rella, R. M. G. García-Castelán, and D. Escobar-Castillejos, "The prediction of academic performance using engineering student's profiles," *Comput. Electr. Eng.*, vol. 93, p. 107288, Jul. 2021, doi:10.1016/j.compeleceng.2021.107288.

[13] W. F. W. Yaacob, S. A. M. Nasir, W. F. W. Yaacob, and N. M. Sobri, "Supervised data mining approach for predicting student performance," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 16, no. 3, p. 1584, Dec. 2019, doi: 10.11591/ijeecs.v16.i3.pp1584-1592.

[14] J. Malini and Y. Kalpana, "Investigation of factors affecting student performance evaluation using education materials data mining technique," *Mater. Today Proc.*, vol. 47, pp. 6105–6110, 2021, doi:10.1016/j.matpr.2021.05.026.

[15] A. Khan, S. K. Ghosh, D. Ghosh, and S. Chattopadhyay, "Random wheel: An algorithm for early classification of student performance with confidence," *Eng. Appl. Artif. Intell.*, vol. 102, p. 104270, Jun. 2021, doi: 10.1016/j.engappai.2021.104270.

[16] B. K. Baradwaj and S. Pal, "Mining educational data to analyze students' performance," *arXiv Prepr. arXiv1201.3417*, 2012.

[17] F. Widyahastuti and V. U. Tjhin, "Predicting students performance in final examination using linear regression and multilayer perceptron," in *2017 10th International Conference on Human System Interactions (HSI)*, IEEE, Jul. 2017, pp. 188–192. doi: 10.1109/HSI.2017.8005026.

[18] I. T. Riyadi Yanto, E. Sutoyo, A. Rahman, R. Hidayat, A. A. Ramli, and M. F. M. Fudzee, "Classification of Student Academic Performance using Fuzzy Soft Set," in *2020 International Conference on Smart Technology and Applications (ICoSTA)*, IEEE, Feb. 2020, pp. 1–6. doi: 10.1109/ICoSTA48221.2020.1570606632.

[19] K. M. Hamdan, A. M. Al-Bashaireh, Z. Zahran, A. Al-Daghestani, S. AL-Habashneh, and A. M. Shaheen, "University students' interaction, Internet self-efficacy, self-regulation and satisfaction with online education during pandemic crises of COVID-19 (SARS-CoV-2)," *Int. J. Educ. Manag.*, vol. 35, no. 3, pp. 713–725, Apr. 2021, doi:10.1108/IJEM-11-2020-0513.

[20] N. B. Pokhrel, R. Khadayat, and P. Tulachan, "Depression, anxiety, and burnout among medical students and residents of a medical school in Nepal: a cross-sectional study," *BMC Psychiatry*, vol. 20, no. 1, p. 298, Dec. 2020, doi: 10.1186/s12888-020-02645-6.

[21] F. Giannakas, C. Troussas, I. Voyiatzis, and C. Sgouropoulou, "A deep learning classification framework for early prediction of team-based academic performance," *Appl. Soft Comput.*, vol. 106, p. 107355, Jul. 2021, doi: 10.1016/j.asoc.2021.107355.

[22] A. Alhadabi and A. C. Karpinski, "Grit, self-efficacy, achievement orientation goals, and academic performance in University students," *Int. J. Adolesc. Youth*, vol. 25, no. 1, pp. 519–535, Dec. 2020, doi:10.1080/02673843.2019.1679202.

[23] H. Wu, S. Li, J. Zheng, and J. Guo, "Medical students' motivation and academic performance: the mediating roles of self-efficacy and learning engagement," *Med. Educ. Online*, vol. 25, no. 1, Jan. 2020, doi: 10.1080/10872981.2020.1742964.

[24] H. A. Mengash, "Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems," *IEEE Access*, vol. 8, pp. 55462–55470, 2020, doi:10.1109/ACCESS.2020.2981905.

[25] H.-B. Ly, T.-A. Nguyen, H.-V. Thi Mai, and V. Q. Tran, "Development of deep neural network model to predict the compressive strength of rubber concrete," *Constr. Build. Mater.*, vol. 301, p. 124081, Sep. 2021, doi: 10.1016/j.conbuildmat.2021.124081.

[26] F. Granata and F. Di Nunno, "Neuroforecasting of daily streamflows in the UK for short- and medium-term horizons: A novel insight," *J. Hydrol.*, vol. 624, p. 129888, Sep. 2023, doi:10.1016/j.jhydrol.2023.129888.

[27] Y. Xu, F. Li, and A. Asgari, "Prediction and optimization of heating and cooling loads in a residential building based on multi-layer perceptron neural network and different optimization algorithms," *Energy*, vol. 240, p. 122692, Feb. 2022, doi:10.1016/j.energy.2021.122692.

[28] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Muller, "Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications," *Proc. IEEE*, vol. 109, no. 3, pp. 247–278, Mar. 2021, doi: 10.1109/JPROC.2021.3060483.

[29] S. Garg, S. Sinha, A. K. Kar, and M. Mani, "A review of machine learning applications in human resource management," *Int. J. Product. Perform. Manag.*, vol. 71, no. 5, pp. 1590–1610, May 2022, doi:10.1108/IJPPM-08-2020-0427.

[30] S. Sharma, G. Singh, and M. Sharma, "A comprehensive review and analysis of supervised-learning and soft computing techniques for stress diagnosis in humans," *Comput. Biol. Med.*, vol. 134, p. 104450, Jul. 2021, doi: 10.1016/j.compbiomed.2021.104450.

[31] J. Choi, B. Gu, S. Chin, and J.-S. Lee, "Machine learning predictive model based on national data for fatal accidents of construction workers," *Autom. Constr.*, vol. 110, p. 102974, Feb. 2020, doi:10.1016/j.autcon.2019.102974.

[32] P. Cunningham and S. J. Delany, "k-Nearest Neighbour Classifiers - A Tutorial," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–25, Jul. 2022, doi: 10.1145/3459665.

[33] M. B. Cohen, B. T. Fasy, G. L. Miller, A. Nayyeri, D. R. Sheehy, and A. Velingker, "Approximating Nearest Neighbor Distances," 2015, pp. 200–211. doi: 10.1007/978-3-319-21840-3_17.