Predicting and Analysis of Students' Academic Performance using Data Mining Techniques

Reda M. Ahmed
Department of Mathematics,
Faculty of Science, South Valley
University, Qena 83523, Egypt

Nahla F. Omran
Department of Mathematics,
Faculty of Science, South Valley
University, Qena 83523, Egypt

Abdelmgeid A. Ali
Department of Computer Science,
Faculty of Computers and Information,
Minia University, Al Minia 61519, Egypt

ABSTRACT

The educational database holds on the massive amount of data and it is increasing rapidly. Data mining provides effective techniques for discovering useful knowledge and pattern from students' data. The discovered patterns can be used to understand many problems in the educational field. This paper proposes a framework to predict the achievement of first-year bachelor's students in computer science course. Decision Tree, Naïve Bayes, and Multi-Layer Perceptron classification methods are applied to the students' data using the WEKA Data Mining tool to produce the best prediction model of the students' academic performance. Experiments conducted to detect the best model among the used techniques then the models' accuracy is computed. The extracted knowledge from the prediction model will be utilized to recognize and profile the student to decide the students' level of success in the first semester.

Keywords

Educational Data Mining, Decision Tree, Naïve Bayes, Multi-Layer Perceptron, Prediction, students' academic performance

1. INTRODUCTION

The world holds on huge amounts of data that increase daily and it is very important to analyze this data to discover the useful information and knowledge from it. Data mining (DM) or knowledge discovery from data (KDD) is the process to discover this interesting patterns and knowledge from the stored data [1]. Data mining has various methods used for the analysis process include classification, clustering, and association rules. This knowledge will be used to solve many problems in many fields like education, economic, business, statistics, medicine, and sport. Recent days, the interest in using data mining increased in such field as education. Educational Data Mining (EDM) is a discipline that interested in improving methods besides analyzing educational data, so it can help to demonstrate students' performance to develop teaching and learning domain [2].

In this paper, three different classification techniques are applied to the students' dataset. The three selected classification algorithms are; Decision Tree (DT), Naïve Bayes (NB), and Multilayer Perceptron (MLP). The best technique will propose a predictive model for Students' Academic Performance (SAP). The obtained patterns will be applied to the students' data. This data was obtained from the records of the students at the first semester of the first-year study in Computer Sciences department, Faculty of Sciences, South Valley university, Qena, Egypt.

WEKA (Waikato Environment for Knowledge Analysis) is a DM tool used for managing the experimental analysis for data mining process. It has lots of machine learning algorithms used for data classification, clustering, association rule, and also evaluation. These algorithms are applied directly to datasets. Also, WEKA used widely for predictions due to its ability in exploring and analyzing [3].

The rest of the paper is organized as follows; Section 2 presents related works in educational data mining while,Section 3 presents the proposed methodology. In section 4 experiments and results are reported. Finally, summary remarks and future work are presented in Section 5.

2. RELATED WORKS

El-Halees[4] implemented the educational data mining in order to improve strategies to discover knowledge from the data coming from the educational domain. In his study, educational data mining was used to analyze learning performance. Also, He gathered student's data from database course. The first step was preprocessing the data, then he applied data mining algorithms to find association rules, classification, clustering and outlier detection. The four used methods illustrated students' performance from the extracted knowledge.

Tair and El-Halees [5] conducted the data mining methods for discovering useful information that comes from the educational systems. Educational data mining utilized to improve graduate students' achievement and overcome the issue of low grades of them and to obtain valuable knowledge from the data include fifteen years period from 1993 to 2007 by the college of science and technology. After prepossessing the data, they implemented data mining methods to discover classification, clustering, association and outlier detection rules. In every method, they presented the extracted knowledge and described its importance in the educational domain.

García and Mora [6] introduced a model to predict students' academic achievement by using socio-demographic and academic factors. This model, which depends on utilizing the Naïve Bayes classifier and the Rapid miner tool, achieved the accuracy of 60 % correct classification. Bharadwaj and Pal [7] studied student performance by choosing 300 students from five distinct colleges. By using means of Bayesian classification technique, which investigated

17 attributes, it was found that factors related to students' living area and medium of teaching were highly connected with student academic performance.

Al-Radaideh et al. [8] proposed a predictive model that utilized data mining classification methods to value information that would determine student's performance. They used three different techniques: ID3, Naïve Bayes, and C4.5. They proved that decision tree performance good in forecasting accuracy than the other two models.

Lakshmi et al. [9] described the students' behavior. They utilized the ID3 algorithm for classifying the students' performance and according to which they would allow the area for their master. The ID3 algorithm is the classification algorithm through which one can create decision trees using top down, bottom up, and greedy search methodologies. The metric information gain used for selecting the most useful parameter for the classification of the datasets.

Ali [10] applied a study in educational systems using data mining techniques. He collected the data by the students at the admission time. Data mining methods include classification and clustering applied to students' data based on psychographic, behavioral, and demographic features. It helped him in describing the student's performance whether they are successful or unsuccessful depended on their GPA or percentage achieved during the secondary school.

Sembiring [11] built a model using data mining techniques based on a psychometric investigation of the students. He created a rule model of the students' performance based on their psychometric behavior. The predictor attributes used were: Interest, Family Time, Believe, and Study Behavior. The model developed two main algorithms were kernel k-means clustering, and support vector machine (SVM) classification.

Bhullar and Kaur [12] described a data mining tool that discovered the students that were weak in academics and need assistance. Weka classification tool used to provide stability between speed, precision, and interpretability of results. The J48 algorithm used in classification method.

Sumitha and Vinothkumar [13] developed a data model to predict student's future learning outcomes using senior student's dataset, and they found that J48 was the best algorithm. Saa [14] found out a qualitative model to analyze the student performance based on related social and personal parameters. He examined theoretically multiple parameters of the students' performance in higher education.

3. METHODOLOGY

This section introduces a model of student's performance using classification techniques. The proposed model will be used to predict student performance in the first semester for first year students at the Faculty of Science in Computer Science course, South Valley University. The three main stages involved in this study; Data Collection and Integration, Data Transformation and Patterns Extraction as shown in figure 1 .

3.1 Data Transformation

This stage was performed to improve the quality of input data to produce the better and quality results. This stage consists of three phases which are data selection, data cleaning, and data normalization. In data selection phase, only seven parameters selected for the mining process.

Several parameters used to identify the factors that influence the students' achievement in academic. In this study, the parameters which used are Gender, Hometown (home), GPA, High School GPA, Section ID and Year. Then we performed data cleaning process to remove missing or incomplete data. That process led to only 346 of 358 data since 12 of missing data removed. The next phase was data normalization process. In this process the numerical values transformed into a nominal or categorical class such as GPA grades parameter that grouped into five categories; G1, G2, G3, G4, and G5.

3.2 Pattern Extraction

In this stage, WEKA open source tool is used to conduct the experiments. This stage consists of five phases: training, pattern building, testing, evaluation results and knowledge representation. The cleaned data divided into two equal parts; training set and testing set. In the training set, the model or pattern used to build from the classification techniques then testing set used to validate this model. After that, the result obtained would be evaluated and represented as a knowledge.

After all these steps, the model constructed by using three different classification methods. The three classification methods applied in this work to evaluate the features that may have an impact on the performance level of the students. Those classification techniques were Naïve Bayesian [NB] classifier, Decision Tree [DT], and Multi-Layer Perceptron (MLP).

DT is a very popular technique in EDM because of its intuitive and human-friendly interpretation for decision makers to do further action, so it is easy to understand and explain [15]. This technique was implemented to the students' data to classify them into successful and unsuccessful classes. So, the lecturers can provide more learning lessons to the students who are limited potential to be successful.

NB classifier is a technique which evaluates the probabilities of parameters values from training data then uses these probabilities to classify new classes. The outputs of the prediction model can be easily understandable to the human [1].

MLP is an application of artificial neural networks that concerns on training data inputs for producing the best accuracy. It constructs of three layers (an input layer, then hidden layer and an output layer). The input layer receives data from the user program and output layer send the result to user program also.

DM techniques aim to develop a predictive model for the students' performance in selected courses. The extracted model would help the lecturers to recognize the students' difficulties to develop the students' level of performance in academic [16, 17].

4. EXPERIMENTS AND RESULTS

4.1 Environment

The experiments were performed on a PC containing 4GB of RAM, Intel Core i3-2379M CPU (2.40 GHz each). For our experiments, WEKA [3, 18] is used to estimate the proposed classification models and comparisons. Furthermore, the dataset is divided into 50% : 50% training and testing partitions equally.

4.2 Evaluation Measures

In the experiments, four standard measures are used for the evaluation of the classification quality: Accuracy, Precision, Recall, and F-Measure which based on equations 1, 2, 3 and 4, respectively.

Accuracy is the proportion of the total number of predictions where correctly calculated. Precision is the ratio of the correctly classified cases to the total number of misclassified cases and correctly classified cases. The recall is the ratio of correctly classified samples to the total number of unclassified instances and correctly classified cases. Also, the F-measure used which combines the recall and precision[17].

$$\textbf{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \tag{1}$$

$$\textbf{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\textbf{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\textbf{F-Measure} = 2 * \frac{\textbf{Precision} * \textbf{Recall}}{\textbf{Precision} + \textbf{Recall}} \tag{4}$$

4.3 Results

In this section, the results of applying classification techniques on the data set are presented as follows. Fig.2 shows the accuracy of the three models for which DT gives 97.69% accuracy that means 169 of 173 are classified correctly to the right class labels. Also, MLP results in 92.49% and NB algorithm is 86.71% accuracy. The highest value for the Precision result is 95.6% obtained for DT algorithm as shown in figure 3. Recall results are plotted in Figure 4 which illustrate that DT model is 97.7% means that 169 students correctly classified to the total number of unclassified cases and correctly classified cases. Regarding the F-Measure results, as shown in figure 5, 96.6% is obtained for the DT techniques. Moreover, the KAPPA results are shown in figure 6. Furthermore, for the convenient of the reader, the results of using the different classification algorithms (MLP, NB and DT) are shown in figure 7 and listed in Table 1 together with the values of KAPPA. Finally, the confusion matrices are plotted in figure s 8-10 for further examination of the results that prove the accuracy of the used models in the current study.

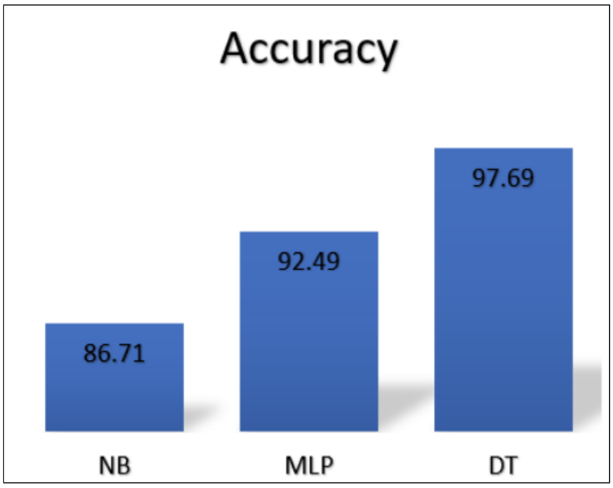


Fig. 2. Accuracy of the three algorithms (DT, MLP and NB)

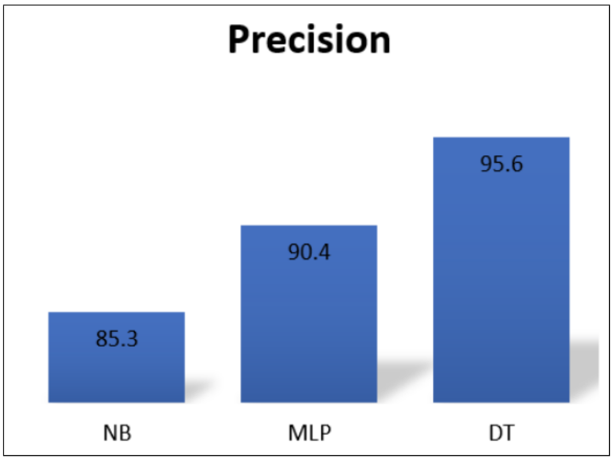


Fig. 3. Precision of the three algorithms (DT, MLP and NB)

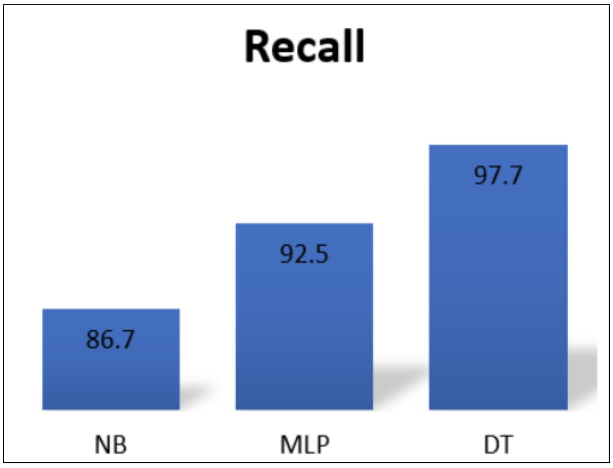


Fig. 4. Recall of the three algorithms (DT, MLP and NB)

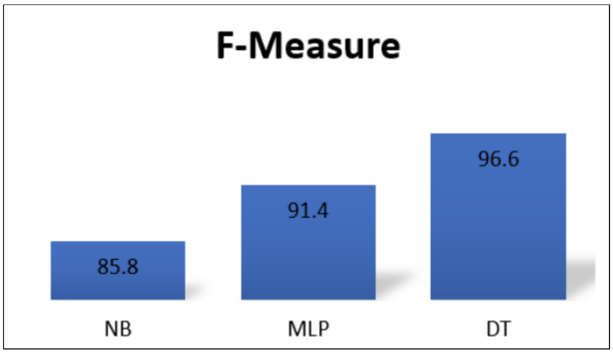


Fig. 5. F-Measure of the three algorithms (DT, MLP and NB)

5. CONCLUSION AND FUTURE WORK

Because of needed knowledge from the data that is rapidly increases every day, data mining techniques become an essential role Data mining techniques applied to discover this knowledge, by using the classification method that determine the contributed parameters of the students' performance. This study conducted a compar-

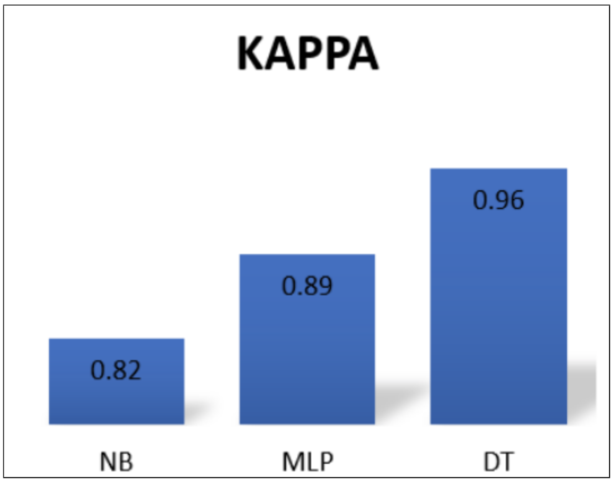


Fig. 6. KAPPA of the three algorithms (DT, MLP and NB)

Table 1. Evaluation Measures of (DT, MLP and NB)

Evaluation Measures	**DT**	**MLP**	**NB**
Accuracy	97.69	92.49	86.71
Recall	97.7	92.5	86.7
Precision	95.6	90.4	85.3
F-mesure	96.6	91.4	85.8
KAPPA	0.96	0.89	0.82

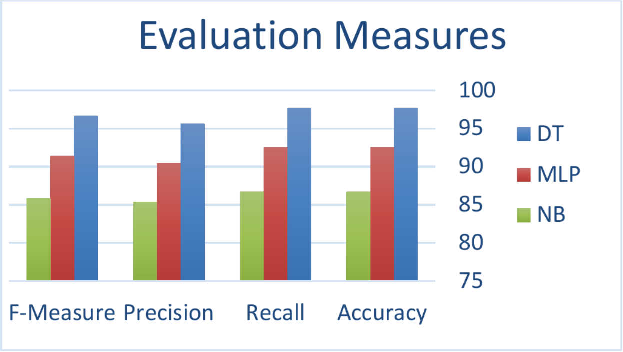
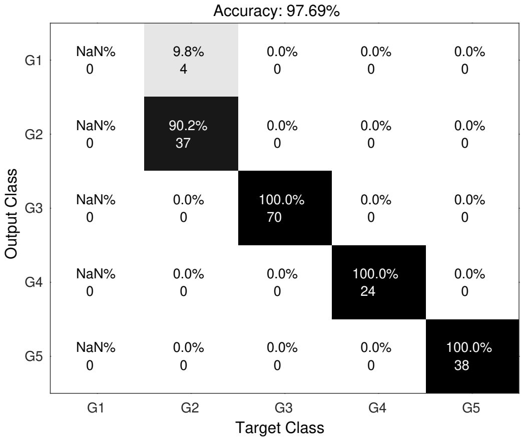


Fig. 7. Evolution Measures of (DT, MLP, NB) algorithms

ative analysis of three classification algorithms; DT, NB, and MLP using WEKA tool. The experimental results show that the DT has the best classification accuracy compared to NB and MLP. The extracted model will guide the lecturers to take early actions in order to help the poor and average students for improving their results. The limitation of this study is the small size of data due to incomplete and missing values in the collected data. In the future, this study could be expanded by adding more data in different years or by using more parameters in order to improve the model prediction. Also, more data mining methods such as genetic algorithms, SVM, K-Nearest Neighbor, and others could be applied.

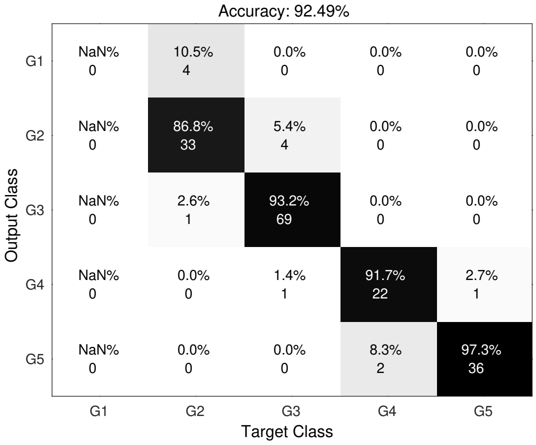
6. REFERENCES

[1] Han, J. 2012. Micheline Kamber, & Jian Pei."Data Mining: Concepts and Techniques".



	G1	G2	G3	G4	G5
G1	NaN% 0	9.8% 4	0.0% 0	0.0% 0	0.0% 0
G2	NaN% 0	90.2% 37	0.0% 0	0.0% 0	0.0% 0
G3	NaN% 0	0.0% 0	100.0% 70	0.0% 0	0.0% 0
G4	NaN% 0	0.0% 0	0.0% 0	100.0% 24	0.0% 0
G5	NaN% 0	0.0% 0	0.0% 0	0.0% 0	100.0% 38

Fig. 8. Confusion Matrix of Decision Tree (J48) The Best Model Obtained (Accuracy: 97.6879 %)



	G1	G2	G3	G4	G5
G1	NaN% 0	10.5% 4	0.0% 0	0.0% 0	0.0% 0
G2	NaN% 0	86.8% 33	5.4% 4	0.0% 0	0.0% 0
G3	NaN% 0	2.6% 1	93.2% 69	0.0% 0	0.0% 0
G4	NaN% 0	0.0% 0	1.4% 1	91.7% 22	2.7% 1
G5	NaN% 0	0.0% 0	0.0% 0	8.3% 2	97.3% 36

Fig. 9. Confusion Matrix of Multi-Layer Perceptron Model (MLP) (Accuracy: 92.4855 %)

[2] Aziz, A. A., Ismail, N. H., & Ahmad, F. 2013." Mining Students' Academic Performance" inJournal of Theoretical & Applied Information Technology,53(3).

[3] Ahmad, F., Ismail, N. H., & Aziz, A. A. 2015. "The Prediction of Students' Academic Performance Using Classification Data Mining Techniques." inApplied Mathematical Sciences,9(129), 6415-6426.

[4] El-Halees, A. 2009. Mining Student's Data to Analyze E-Learning Behavior: A Case Study.

[5] Tair, M. M. A., & El-Halees, A. M. 2012." Mining Educational Data to Improve Students' Performance: A Case Study" in International Journal of Information,2(2), 140-146.

[6] García, E. P. I., & Mora, P. M. 2011." Model Prediction of Academic Performance for First Year Students". InArtificial Intelligence (Micai), 2011 10th Mexican International Conference (pp. 169-174). IEEE.

[7] Bhardwaj, B. K., & Pal, S.2012." Data Mining: A Prediction for Performance Improvement Using Classification".Arxiv Preprint Arxiv:1201.3418.

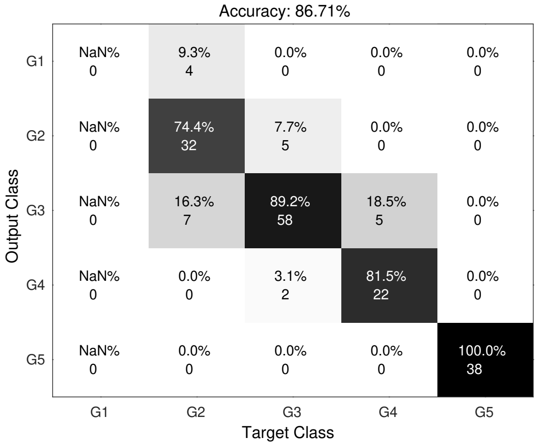


Fig. 10. Confusion Matrix of Naïve Bayes Model (NB) (Accuracy: 86.7052 %)

[8] Al-Radaideh, Q. A., Al-Shawakfa, E. M., & Al-Najjar, M. I. 2006.” Mining Student Data Using Decision Trees.” In-International Arab Conference on Information Technology, Yarmouk University, Jordan.

[9] Lakshmi, D., Arundathi, S., & Jagadeesh, D. 2014.” Data Mining: A Prediction for Student’s Performance Using Decision Tree Id3 Method”.

[10] Ali, M. M. 2013. ”Role of Data Mining in Education Sector” in International Journal of Computer Science and Mobile Computing,2(4), 374-383.

[11] Sembiring, S. 2012.”An Application of Predicting Student Performance Using Kernel K-Means and Smooth Support Vector Machine” in(Doctoral Dissertation, Ump).

[12] Bhullar, M. S., & Kaur, A. 2012.” Use of Data Mining in Education Sector.” InProceedings of The World Congress on Engineering and Computer Science(Vol. 1, pp. 24-26).

[13] Sumitha, R.,& Vinothkumar, E. S. 2016.” Prediction of Students Outcome Using Data Mining Techniques.” inInternational Journal of Scientific Engineering and Applied Science,2(6)8.

[14] Saa, A. A. 2016.” Educational Data Mining & Students’ Performance Prediction.” inInternational Journal of Advanced Computer Science and Applications,7(5), 212-220.

[15] Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. 2017. ”Analyzing Undergraduate Students’ Performance Using Educational Data Mining.” inComputers & Education,113, 177-194.

[16] Hussain, S., Dahan, N. A., Ba-Alwib, F. M., & Ribata, N. 2018.” Educational Data Mining and Analysis of Students’ Academic Performance Using Weka.” inIndonesian Journal of Electrical Engineering and Computer Science,9(2).

[17] Chen, T. Y., Kuo, F. C., & Merkel, R. 2004.” On the Statistical Properties of The F-Measure.” InQuality Software, 2004. Qsic 2004. Proceedings. Fourth International Conference on (pp. 146-153). IEEE.

[18] Arora, R. 2012.” Comparative Analysis of Classification Algorithms on Different Datasets Using Weka.” in International Journal of Computer Applications,54(13).

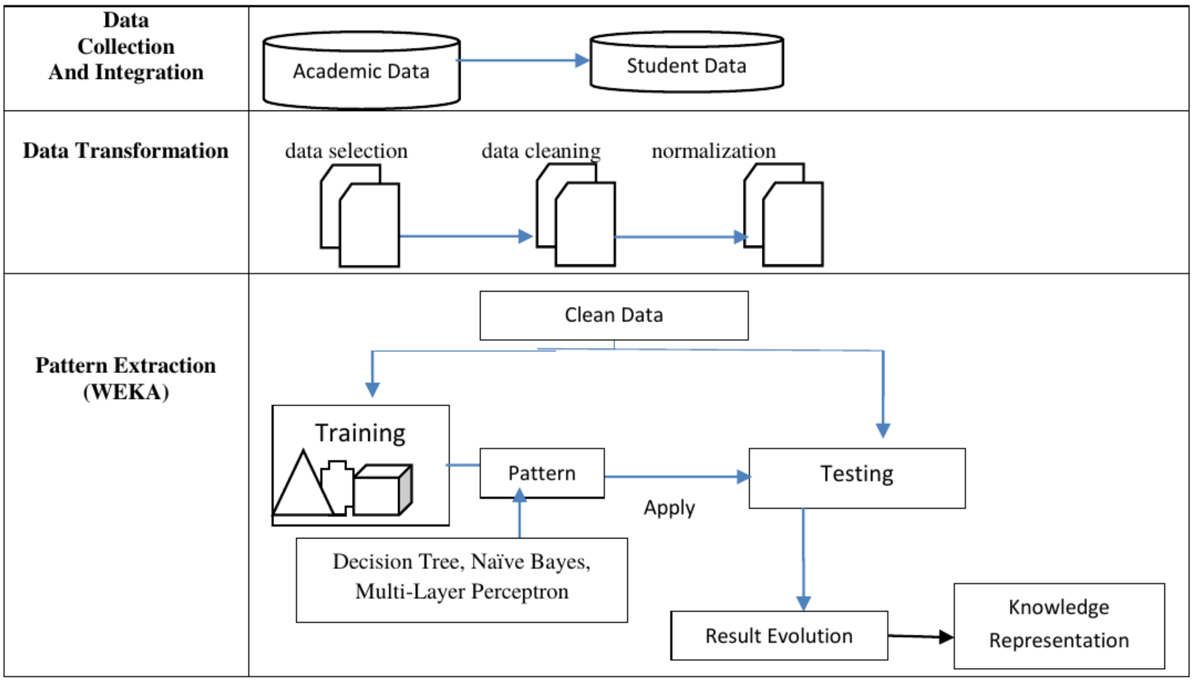


Fig. 1. Framework of Students' Academic Performance prediction