A Comparative Study of Machine Learning Approaches on Prediction System of Students Academic Performance

K.Nithya,M Phil Research Scholar[1], Dr.V.Narayani, Assistant Professor2

[1-2]Department of Computer Science, St. Xavier's College, Tirunelveli, Tamilnadu, India.

Abstract - *Performance prediction of students is essential to check the feasibility of improvement. Regular evaluation not only improves the performance of the student but also it helps in understanding where the student is lacking. It takes a lot of manual effort to complete the evaluation process as even one college may contain thousands of students. This paper compared an automated solution for the performance prediction of the students using machine learning. Predicting students performance becomes more challenging due to the large volume of data in educational databases. There are two main reasons of why this is happening. First, the study on existing prediction methods is still insufficient to identify the most suitable methods for predicting the performance of students. Second is due to the lack of investigations on the factors affecting students achievements in particular courses. Therefore, a comparative study on predicting student performance by using machine learning approaches is compared to improve students achievements. The main objective of this paper is to find the best machine learning techniques that have been used to predict students performance. This paper also focuses on how the prediction algorithm can be used to identify the most important attributes in a student's data.*

Keywords— Student performance, Educational Data Mining; Learning Analytics model; FPSO; SVM; KNN; Navie Bayes

I. INTRODUCTION

Students performance is an essential part in higher learning institutions. This is because one of the criteria for a high quality university is based on its excellent record of academic achievements [1]. There are a lot of definitions on students performance based on the previous literature. Usamah et al. (2013) stated that students performance can be obtained by measuring the learning assessment and co-curriculum [2]. However, most of the studies mentioned about graduation being the measure of students success. Generally, most of higher learning institutions in Malaysia used the final grades to evaluate students performance. Final grades are based on course structure, assessment mark, final exam score and also extracurricular activities [2] . The evaluation is important to maintain students performances and the effectiveness of learning process.

Currently, there are many techniques being compare to evaluate students performance. Data mining is one of the most popular techniques to analyze students performance. Data mining has been widely applied in educational area recently [10]. It is called educational data mining. Educational data mining is a process used to extract useful information and patterns from a huge educational database [11]. The useful information and patterns can be used in predicting students performance. As a result, it would assist the educators in providing an effective teaching approach. Besides, educators could also monitor their students achievements. Students could improve their learning activities, allowing the administration to improve the systems performance. Thus, the application of data mining techniques can be focused on specific needs with different entities. In order to encounter the problems, a systematically review is compared. The main objective of this work are:

1. To study and identify the gaps in existing prediction methods.
2. To study and identify the variables used in analyzing students performance.
3. To study the existing prediction methods for predicting students performance.

II. RELATED WORK

Machine learning is part of the Artificial Intelligence (AI), where computer can teach themselves to learn the data. While data mining is a technique to find pattern in large amount of data[3]. Machine learning used in education have much attention lately [4] For predicting student's performance, the widely used model are instance-based learning, Naïve Bayes, Decision Tree, Artificial Neural Network, Support Vector Machine, Classification Tree [5].

The idea computing a hyper plane to minimize the loss function [6]. Support vector machine is widely used for data mining and classification to predict membership of a data. It is based on the geometrical interpretation. Decision tree is one of the most common used techniques in predicting student's performance[7].

A publication done in 2016, published about Predicting student's final year GPA by the result of important courses that distributed in 8 semesters of studies[8]. Data mining in education field known as Educational Data Mining(EDM). It happen because of the increase of

educational resources and data that can be explored to learn how a student learned [9].

III. METHODOLOGY

In this work, three classifiers which are used to predict student performance are compared. The outline of student performance prediction work is shown in Fig.1. This work consists of four modules. They are

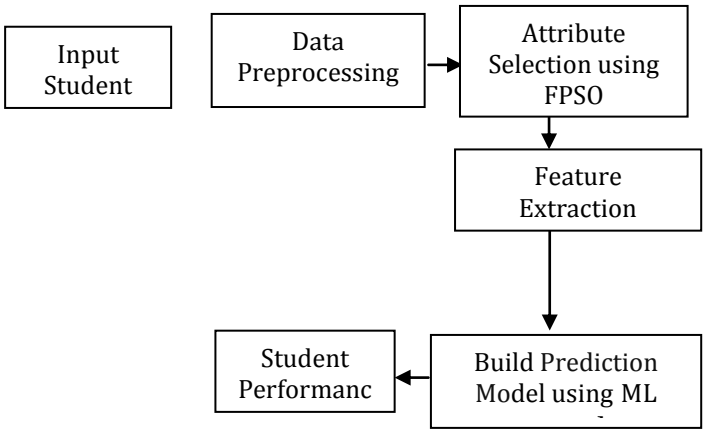


Fig.1 Outline of Student Performance Prediction Model

1. Data Preprocessing
2. Attribute Selection
3. Feature Extraction
4. Prediction Model Generation

1.Data Preprocessing

Initially, data on the student records are collected from the university enterprise database. The data is then re-formatted in the stage of data transformation in order to prepare for processing by subsequent algorithms. In the data cleaning process, the parameters used in the data analysis are identified and the missing data are either eliminated or filled with null values.

2.Attribute Selection

In attribute selection, the most important attributes of the student database are only selected by using the novel FPSO attribute selection approach.

In order to achieve good prediction results, generally several types of features are applied at the same time. Since the different types of features may contain complementary information, it could bring better prediction performance through selecting discriminative features from various feature spaces. The advantage of feature selection is to determine the importance of original feature set.

For feature selection, Fuzzy Particle Swam Optimization (FPSO) is applied. A FPSO [20] is composed of a knowledge base, that includes the information given by the expert in the form of linguistic control fuzzy rules, a fuzzification interface, which has the effect of transforming crisp data into fuzzy sets, an inference system, that uses them together with the knowledge base to make inference by means of a reasoning method, and a defuzzification interface, that translates the fuzzy control action thus obtained to a real control action using a defuzzification method.

3.Feature Extraction

After selected the important attributes, the next step is to generate the feature matrix. In this step three feature matrices are generated namely grade matrix, performance matrix, interest matrix based on the students mark, performance and interests.

4.Prediction Model Generation

In this module the prediction model is created by using various several machine learning approaches. Among several machine learning approaches, this work uses K Nearest Neighbour, Naive Bayes and Support Vector Machine as prediction model generation.

K-Nearest Neighbour

The k-Nearest Neighbors algorithm is a non-parametric method used for classification and regression. The input consists of the k closest training examples in the feature space. In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k-nearest neighbors. If k = 1, the object is simply assigned to the class of that single nearest neighbour.

Knn Algorithm Pseudocode:

1. Calculate "d(x, x_i)" i =1, 2, ….., **n**; where **d** denotes the Euclidean distance between the points.
2. Arrange the calculated **n** Euclidean distances in non-decreasing order.
3. Let **k** be a +ve integer, take the first **k** distances from this sorted list.
4. Find those **k**-points corresponding to these **k**-distances.
5. Let $\mathbf{k}_i$ denotes the number of points belonging to the i^{th} class among **k** points i.e. $k \geq 0$
6. If $k_i > k_j \; \forall \; i \neq j$ then put x in class i.

Naive Bayes

Naive Bayes classifier depends on a probability model and allocates the specific class, which has the maximum estimated posterior probability to the feature vector. The posterior probability $P(C_v/FV)$ of a specific class C_v is

given by a feature vector FV is determined using Bayes' theorem is given in equation below,

$$P(C_v/FV) = \frac{P(FV/C_v)\,P(C_v)}{P(FV)}$$

Naive Bayes Classifier technique is based on the Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods.

Naive Bayes steps
Step 1: Separate By Class.
Step 2: Summarize Dataset.
Step 3: Summarize Data By Class.
Step 4: Gaussian Probability Density Function.
Step 5: Class Probabilities.

Support Vector Machine

SVM is a binary classification method that takes as input labelled data from two classes and outputs a model file for classifying new unlabeled/labeled data into one of two classes. It group items that have similar feature maps into groups. SVM constructs a hyperplane that maximizes the margin between negative and positive samples. Finally, classification is performed by the decision based on the value of the linear combination of the features.

SVM is trained by feeding known data with previously known decision values, by forming a finite training set. It is from the training set that an SVM gets its intelligence to classify unknown data. In SVM, for two class classification problem, input data is mapped into higher dimensional space using RBF kernel. Here, a hyper plane linear classifier is applied in this transformed space utilizing those patterns vectors that are closest to the decision boundary.

The estimation for the classification using SVM with N support vectors $g_1, g_2, \cdots g_n$ and weights $\tau_1, \tau_2, \ldots \tau_n$ is given by:

$$SVM = \sum_{i=1}^{n} \tau_i\,(g_i, x) + b$$

Where x represents a feature vector and b represents a bias.

Pseudocode for SVM:
Training a SVM can be illustrated with the following Pseudocode:
Require: X and Y loaded with training labeled data, $\propto \leftarrow 0$ *or* $\propto \leftarrow$ partially trained SVM:
1: C← some value
2: repeat
3: for all $(xi, yi), (xj, yj)$, do
4: optimize $\propto_i$ and $\propto_j$
5: end for
6: until no changes in $\propto$ or other resource constraint criteria met
Ensure: Retain only the Support Vector ($\propto_i > 0$)

IV. RESULT AND ANALYSIS

A. *Efficiency Parameters*

To assess the efficiency of the machine learning approaches, several efficiency metrics are available. This paper employs the Detection Accuracy to analyses the efficiency.

Detection Accuracy

Detection Accuracy is the measurement system, which measure the degree of closeness of measurement between the original results and the correctly prediction results.

Accuracy= (TP+TN)/(TP+FP+TN+FN)

B. *Experimental Results*

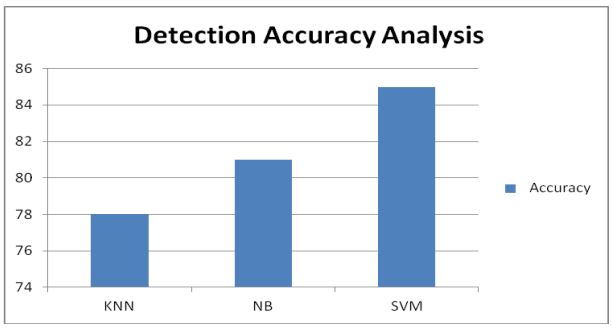
Experiment : Accuracy Analysis of Machine Learning Approaches

In this experiment, this work will assess the contribution of each classifier approaches which are employed in the work. Ideally, a excellent machine learning scheme is accepted to have a high Accuracy value. Table 1 lists the accuracy analysis of FPSO with SVM.

Table 1: Detection Accuracy Analysis of Machine Learning Approaches

Metrics	Accuracy
KNN	78
NB	81
SVM	85

As observed from Table 1, the Accuracy of the FPSO with SVM in range 85, which is superior than other methods. So the FPSO with SVM classifier is considered to be the best for sentiment analysis. Fig.1 depicted the Detection Accuracy of classifier approaches.



V. CONCLUSION

Predicting students performance is mostly useful to help the educators and learners improving their learning and teaching process. This paper has compared various

machine learning approaches and feature selection approaches on predicting students performance with various analytical methods. Classification is done in order to predict students in different class categories like High, medium and low. The results of both feature selection approaches and machine learning were compared on the basis of accuracy and precision. It was found and detected that classification implemented by SVM with FPSO is more efficient compare to other classifiers as seen in the accuracy and precision. Based on the results, SVM with FPSO technique is more efficient compared to other technique in prediction of students' performance.

REFERENCES

[1] M. of Education Malaysia, National higher education strategic plan (2015).
[2] U. bin Mat, N. Buniyamin, P. M. Arsad, R. Kassim, An overview of using academic analytics to predict and improve students' achievement: A proposed proactive intelligent intervention, in: Engineering Education (ICEED), 2013 IEEE 5th Conference on, IEEE, 2013, pp. 126–130.
[3] I. No, S. Anam, and S. Gupta, "A research Review on Comparative Analysis of Data Mining Tools, Techniques and Parameters," Int. J. Adv. Res. i, vol. 8, no. 7, pp. 523–529, 2017.
[4] J. Xu, K. H. Moon, and M. van der Schaar, "A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs," IEEE J. Sel. Top. Signal Process., vol. 11, no. 5, pp. 742–753, 2017.
[5] I. Đurđević Babić, "Machine learning methods in predicting the student academic motivation," Croat. Oper. Res. Rev., vol. 8, no. 2, pp. 443–461, 2017.
[6] A. Kumar, J. Naughton, and J. M. Patel, "Learning Generalized Linear Models Over Normalized Data," Proc. 2015 ACM SIGMOD Int. Conf. Manag. Data - SIGMOD '15, pp. 1969–1984, 2015.
[7] M. Pandey and V. K. Sharma, "A Decision Tree Algorithm Pertaining to the Student Performance Analysis and Prediction," Int. J. Comput. Appl., vol. 61, no. 13, pp. 2–6, 2013.
[8] M. A. Al-Barrak and M. Al-Razgan, "Predicting Students Final GPA Using Decision Trees: A Case Study," Int. J. Inf. Educ. Technol., vol. 6, no. 7, pp. 528–533, 2016.
[9] C. Romero and S. Ventura, "Data mining in education," Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 3, no. 1, pp. 12–27, 2013.
[10] C. Romero, S. Ventura, Educational data mining: A review of the state of the art, Trans. Sys. Man Cyber Part C 40 (6) (2010) 601–618.
[11] D. M. D. Angeline, Association rule generation for student performance analysis using apriori algorithm, The SIJ Transactions on Computer Science Engineering & its Applications (CSEA) 1 (1) (2013) p12–16.