Prediction of Students' Performance in University Level using KNN Classification Algorithm

Delna Sebastian[1]Saravanan K.N.[2]
[1,2]Department of Computer Science & Engineering
[1,2]Christ University, Bengaluru, Karnataka, 560029, India

Abstract**—***The academic performance of a student depends on one's own potential which can be judged by various attributes and its respective numeral values associated with their parameters. In this research work, KNN classification algorithm has been applied to predict the category to which a student falls based on their skills, ability and behaviour of the students. With this research work it is possible to predict the capability of a student and there by the concerned educational institutions will be able to provide better feedback to various stake holders. We have collected the data set from Christ University faculty members to facilitate the prediction. Using KNN classification with thirteen identified features, the prediction shows an accuracy of 97.61%.*

Keywords—Predictive Analytics, K-Nearest Neighbour, Clustering Algorithm, Minimum Distance Classifier, Feature, Feature Extraction, Prediction

I. INTRODUCTION

Predictive analytics is the process of predicting the unknown future events by the use of historical data which can also be used for forecasting and modelling. For predicting the future pattern predictive analytics uses many techniques from data mining, statistical algorithm and machine learning, all based on the historical data. The industries use predictive analytics in different fields such as health Care, manufacturing, banking and finance etc. to reduce risk, enhance operations and increase profits. Recently predictive analytics had been used in Universities to evaluate the performance of students. Data mining in higher education is another rising field called Educational Data Mining(EDM). Using this technique, the universities will be able to identify which student is at the risk stage. If the prediction says that a student tends to fail in the examination, then the institutions can take extra effort to improve their studies and help them to get good score in the examination. Classification technique such as K- Nearest neighbour(KNN), Decision Trees, Naïve Bayes, Neural Networks and many others are used for the prediction of the student's performance.

The main aim of this study is to predict whether the particular student is eligible within the category of poor, average, good or very good with the help of collected features such as listening skills, reproducing skills, understanding skills, recollecting skills, subject interest, punctuality, regularity, continuity, involvement, health condition, time spend out of the class, English proficiency, and degree of intelligence. These features and its respective values were collected from the faculties of Christ University, Bangalore, India. The KNN classification algorithm is applied to the dataset and classify the student into their respective category.

II. LITERATURE REVIEW

K V Krishna Kishore et. al., [1] made a study to estimate the Grade Point Average (GPA) of students using the data with sample size of 134 that was collected from third Year Students of Computer Science Engineering Vignan University. The features used by the researcher was aggregate marks, attendance and backlogs up to the current semester, 10th & 12th percentages, Degree of Intelligence, Working Nature, Discipline, Social Activities and Grade [1]. Data pre-processing techniques are used to increase the efficiency of the data and to fill the null values. Once the data pre-processing technique is applied, the final dataset is divided into train and test dataset. The training dataset is used to build the basic classification model. The test dataset either tests the performance and efficiency of the classification model or compares the results with the known target data. The GPA of the students is predicted using a classification technique known as Multilayer Perceptron (MLP). Multilayer Perceptron is compared with other classification techniques like Naïve Bayes, CART, RBF Networks and J48. The Confusion matrix generated with Naïve Bayes, J48, CART, RBF Neural Network, MLP algorithm has accuracy 85.82%, 87.31%,93.28%, 95.52% and 97.37% respectively. The experimental results show that accuracy of the Multilayer Perceptron (MLP) is 97.37% on test dataset which is better than classification methods like J48, CART, Naïve Bayes and RBF Neural Network.

MuslihahWooket. al., [2] has compared the Artificial Neural Network (ANN) and the combination of clustering and decision tree classification techniques to predict the academic performance of the students. This study identifies the attributes that has influence on the performance of students after their first year degree examinations. The student data of Computer Science Department, Faculty of Science and Defense Technology, National Defense University of Malaysia (NDUM) is used as the dataset for this study. The author collected 85 students' primary and secondary data. The primary data was obtained from the questionnaire which consists of significant features from each student. The selected features from the primary data collection are Demographics, Education background and Personality. The secondary data is the students' previous results like CPA, CGPA and Grade Points. The preprocessing techniques have applied to fill the missing values using smoothing and mean value. Then the data was divided into training and testing data of which is used by the former to develop the model. Neural network and combination of clustering and decision tree techniques are applied to the train dataset and the obtained results from each of the techniques are compared and the best technique was chosen as the model for the research. This research defined CRISP-DM methodology and it involves six steps: The experimental result is a comparison between Artificial

Neural Network (ANN) and the combination of clustering and decision tree classification techniques. The best model is chosen as the technique that gave an accurate prediction and classification.

Surjeet Kumar Yadav et. al., [3] have conducted a study on the engineering student's data to predict their performance in the final exam. The aim of this experiment was to find the number of students who may have the chance to pass or fail. The dataset used for the study was collected from VBS Purvanchal University, Jaunpur (Uttar Pradesh) on the sampling method for Institute of Engineering and Technology for session 2010 with sample size of 90. The data was obtained during the time of admission from the students in the filled enrollment form. Some of the information which is related to the variable were extracted from the database. Students Branch, gender, grade in High School, grade in Senior Secondary, Admission Type, Medium of Teaching, Living Location, accommodation, family size, status of the family, annual income, parents' qualification and occupation, Result in B. tech Ist year are the selected features obtained from the data base. After the feature selection and the data transformation, the decision tree algorithms such as C4.5, ID3 and CART are applied to the normalized dataset to calculate the performance of the students in the final exam using Weka, an open source software widely used in data mining applications. 10-fold cross-validation technique was used for evaluating the accuracy of this prediction. From the accuracy it is clear that the rate of true positive for the fail class is 0.786 which indicates that the model identifies the students who are expected to fail with the help of decision tree such as ID3 and C4.5. Students who have more chances of failure can look for proper advising measures to improve their result.

MD. FahimSikderet. Al., [4] conducted a study on students' data to predict student's Cumulative Grade Point Average (CGPA) using the neural network and compared that with real CGPA. For predicting the students' yearly performance, 120 sample data were collected from the students through an online survey, which was collected from the Department of Computer Science and Engineering of Bangabandhu Sheikh Mujibur Rahman Science and Technology University. Class Test Marks, Class Performance, Class Attendance, Assignment, Lab Performance, Previous Semester Result, Study Time, Family Education, living area, Social Media Interaction, Extracurricular Activity, Drug Addiction, Affair, Year Final Result [4], is the extracted features for this study. The collected sample consists some unwanted data, MATLAB as the data mining tool which is used for filtering unnecessary data. After the pre-processing, the finalized data divided into three parts. The first part is training dataset. The second part is testing dataset and the third part is used for validating the result. To train the network the Levenberg-Marquardt algorithm added as a training algorithm. After applying the Neural Network technique, an accuracy of 97% was obtained for the prediction of the particular dataset.

Behrouz Minaei-Bidgoliet. Al., [5] have classified the students based on their predicted final grade with the use of some features collected from logging data in an Educational web-based system. The sample size of the dataset which was used for this study is 227 of the latest online educational systems developed at MSU, the Learning Online Network with Computer-Assisted Personalized Approach (LON-CAPA) [5]. Using six different classifiers the author compared the dataset using some Error estimate. 1-nearest neighbour (1-NN), Quadratic Bayesian classifier, k-nearest Neighbour (k-NN), Parzen-window, multi-layer perception (MLP), and Decision Tree are the six different classifier which were used for the comparison. Normalization has been used as a pre-processing technique, a combination of classifiers was used to remove the outliers and improve the efficiency of the data. There are three different ways, the author predicted the final grade of the students that is (1) Utilizing nine applicable class labels related to that of student grades. (2) Using three class labels that is low, middle and high, (3) using two class labels namely pass and fail [5]. With the use of Cross-Validation with 10 folds the accuracy of the prediction is evaluated. The result of the study indicates that KNN has the best performance in the case of 2-classes and CART has the best accuracy in the case of 3-classes and 9-classes.

III. METHODOLOGY

The model of the proposed system is shown in Figure 1. It describes each stage to build a predictive model. For the development of the proposed system, the first step is to identify the features and their respective values. The dataset which is used for this study has the sample size 106. The obtained data is pre-processed by applying statistical calculations in excel tool. The normalized data after being pre-processed is separated into train data set and test data set. In this proposed system, we predict student's performance by applying the K nearest neighbour (KNN) algorithm using MATLAB tool and then we find out the accuracy percentage.

A. Data Collection & Feature Description

The method used to collect the data is a primary data collection method. The dataset was collected from the faculties of Christ University, Bangalore through the questionnaire. Extracted data from the questionnaire consist 106 samples. Following are the selected features which are used for the prediction.

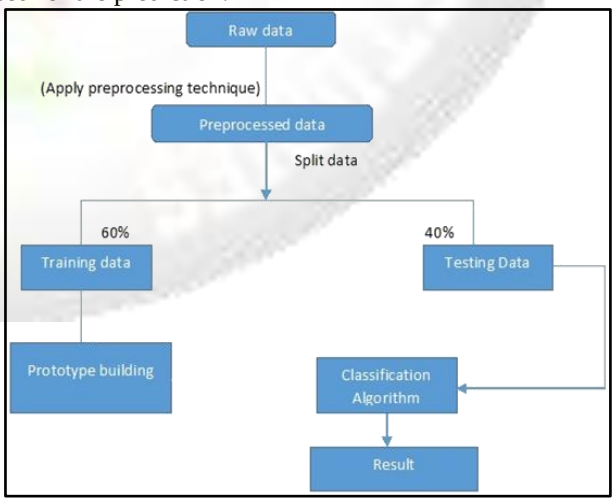

Fig.1:Model of Proposed System

Listenig skills	Reprod-ucing skill	Understan-ding skill	Recollect ing skill	Subject Interest	Punctuality	Regularity	Continuity	Involvement	Health condition of the student	Time spend out of the class	English profecence	Degree of Inteligence
3	4	2	4	2	2	2	2	2	5	1	1	2
5	5	3	3	3	4	4	5	5	9	4	7	6
4	5	6	7	5	7	8	6	5	10	6	8	7
3	5	6	6	5	6	7	6	5	10	6	8	7
8	8	8	7	8	8	8	7	8	7	8	7	8
7	7	7	6	7	7	8	7	7	7	7	7	7

Fig.2:Sample Data Collection Using Thirteen Features

1) Listening Skills

A teacher can evaluate a student listening skill trough the various approaches like questioning section, interaction with the students, making notes, way of attending the class etc.

2) Reproducing Skills

Clarification of a specific subject which was at that point examined amid the class, their capacity to clarify the queries of their companions about the discussed topic and preparation of notes after the class has been done are some of the factors which helps to measure the reproducing skill of a student.

3) Understanding Skills

It can be measured by observing the number of repetition a teacher takes to make a student clear about a specific topic.

4) Recollecting Skills

Capability of the student to recollect previously taught lessons, Students performance during the revision classes before examination and the way of answering a question in exam.

5) Subject Interest

Response towards teacher, making lecture notes in between the class, completion of class works and home work related to that particular subject before due date, mental and physical presence during the class hours, way of behaviour towards the faculty etc. are the some of the measurable factors for subject interest.

6) Punctuality

The Punctuality of a student can bedetermined with the help of aspects such as attending the class, submitting assignments and Projects on time, conducting seminars without any kind of delay.

7) Regularity

It can be determined by how a student regular in a particular subject. If a student is not regular in a particular subject's topic it leads to an incomplete understanding of that particular topic.

8) Involvement

Sharing their own practical or theoretical knowledge they had about a topic when that specific topic is discussed in class, how a student will be able to define a topic apart from the syllabus. All these shows the involvement of a student towards the subject.

9) English Proficiency

Due to the lack of English proficiency, some students are unable to perform well in exam even though they are good in subject. Apart from this, usage of suitable vocabularies during seminar presentation, publication of articles and way of interaction are some of the measurable factors.

10) Continuity

While describing a topic, if a student doesn't attend it completely they may lose their continuity in that specific topic. Alternative absence in class, diversion of concentration from the topic which is taking in class are also the reasons which affect a student's continuity.

11) Health Condition

The health condition of a student plays a vital role in theireducation. An unhealthy condition can cause a student to miss their class, they also may find it difficult to follow the lectures, incompletion of notes, unable to attend the exam due to the health problems, bad performance in exams due to hypertension and stressed up persons end up having an imbalance in their studies.

12) Time Spend Out of the Class

This attribute value is measured from a student's extra effort outside the class by utilizing Library hours, combine studies, attending online courses related to particular subjects.

13) Degree of Intelligence

Some students even if theydon't attend class or complete notes they will able to secure good marks in exams because of their IQ level.

B. Data Pre-Processing

The collected data set had some missing values and some outliers. The missing values can be replaced with the help of measures of central tendency such as mean, median and mode. Here Mean method was taken to fill the missing values. With the use of standard deviation some extreme values from the average value of each attribute were obtained. Such extreme values were removed directly from the data set. The cleared data was categorized into two sets for further analysis. The first 60 % data part comes under training dataset and the remaining 40% is in the test dataset.

C. Building Knowledge Model

K-nearest neighbour(KNN) classification is a supervised algorithm. It is also used in regression predictive problem. The advantages of KNN algorithm are easy interpretation of output, predictive power and less calculation time. These highlights prompts choose KNN to build the predictive model for this system. The main objective of this algorithm is to find the nearest neighbours of an unknown data point and it is works on the basis of k value. If K=n, then the 'n' nearest neighbour can be predicted. The final classification output can be decided by calculating the distances between the test data and each of the training data with help of KNN algorithm. The class label of the data set is poor, average, good and very good. Using this algorithm, we can predict the class to which a student belongs to or which is the nearest neighbour of a particular student. Euclidean distance function is the most commonly used distance equation in KNN. The Euclidean distance between two points (x_1, y_1) and (x_2, y_2) in the plane is given by the equation, [6]

$$\text{Dist}\,(x_1, y_1), (x_2, y_2) = \sqrt{(x_1-x_2)^2+(y_1-y_2)^2} \quad (1)$$

IV. INTERPRETATION

By the application of KNN algorithm, distinctive values of 42 students (test data) were entered in the training data set. Here the value of K is considering as 11. MATLAB tool was used for this process. Out these 42 values, 41 were true positive and remaining 1 was a false positive. Confusion matrix was used to find out the accuracy rate.

Total=42	Predicted Poor	Predicted Average	Predicted Good	Predicted Very Good
Actual poor	12	1	0	0

Actual average	0	8	0	0
Actual good	0	0	9	0
Actual very good	0	0	0	12

Table 1: Confusion Matrix

Classification accuracy
= Correct predictions / total prediction*100
= 41/(42*100) = 97.61
When KNN algorithm was applied an accuracy rate of 97.61% was obtained.

V. CONCLUSIONS

From this study it has been found that the K-nearest neighbour is the most efficient algorithm which can be used to predict the students' performance. This study will help the educational institutions to identify a student who required more effort and time for their advancement. It also enables institutions to identify bright students and nurture their future growth. We have used about 106 students' data in our experiment. KNN has attained 97.61% accuracy on test data set. We have found that KNN is the best solution for this type of prediction.

REFERENCES

[1] K V Krishna Kishore," Prediction of Student Academic Progression: A Case Study on Vignan University", International Conference on Computer Communication and Informatics, IEEE, 2014
[2] MuslihahWook," Predicting NDUM Student's Academic Performance Using Data Mining Techniques", Second International Conference on Computer and Electrical Engineering, IEE, 2009
[3] Surjeet Kumar Yadav," Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification", World of Computer Science and Information Technology Journal, IEE, 2012
[4] Md. FahimSikder, "Predicting Students Yearly Performance using Neural Network:A Case Study of BSMRSTU", 5th International Conference on Informatics, Electronics and Vision, IEE, 2016
[5] Behrouz Minaei-Bidgoli, "Predicting Student Performance: an application of data mining methods with an educational web-based system", IEE, 2003
[6] https://www.cut-the-knot.org/pythagoras/DistanceFormula.shtml