Predicting the performance of instructors using Machine learning algorithms

Article *in* High Technology Letters · December 2020

DOI: 10.37896/HTL26.12/2315

CITATIONS
3

READS
802

2 authors:

Panimalar Kathiroli
SRM Institute of Science and Technology
23 PUBLICATIONS **20** CITATIONS
SEE PROFILE

Vijayalakshmi v.
CMR Institute of Technology
45 PUBLICATIONS **48** CITATIONS
SEE PROFILE

Predicting the performance of instructors using Machine learning algorithms

V. Vijayalakshmi[1], K. Panimalar[2], Sakthivel Janarthanan[3]
[1]Associate Professor, Department of CSE, Sri Manakula Vinayagar Engineering College, Pondicherry
[2]Research Scholar, Department of CSE, Pondicherry Engineering College, Pondicherry
[3]Research Scholar, Department of CSE, Hindustan University, Chennai

ABSTRACT
Data mining applications are feasible tools for tackling different business applications such as financial, healthcare, telecommunication and e-learning, and so on. Most of the literature survey concentrates on predicting the student's performance. But the student improvement depends on the quality of the instructor. Hence this paper concentrates on predicting the teacher's performance and explores the factors influencing student accomplishment. The conventional technique to assess the educator's performance is an assessment survey considering the student's point of view. The real-time dataset was gathered from the students. In this paper, some of the machine learning algorithms was utilized such as Naïve Bayes, K-Nearest Neighbor, Random Forest, Support Vector Machine, and Decision Tree. Implementation was done in the R programming environment which is a reasonable dialect for data mining applications. Numerous execution measurements were used to evaluate the system. Results demonstrate that variable straightforwardly relies upon the instructor's performance such as accuracy, precision, recall, specificity, sensitivity using different machine learning algorithms. Support Vector Machine produces the highest accuracy than other models created with the same dataset.

Keywords: Educational Data Mining (EDM); R Programming; SVM; Machine Learning Algorithm; Prediction.

1 Introduction

Data Mining defined as extracting knowledge from massive amounts of data. It is one of the significant steps of the KDD process. Data Mining is the process of discovering interesting knowledge from massive amounts of information stored in data warehouses, databases or other information repositories [1]. It is otherwise called as Knowledge Discovery from Databases. Hidden information from large database is best analyzed with data mining techniques for making predictions or behaviors. Data Mining has many algorithms and techniques like classification, clustering, regression, artificial intelligence, neural network, association rule, decision tree, genetic algorithm, nearest neighbor etc, for discovering the knowledge from the databases. Data mining technique used in many domains like education, banking, fraud detection, Intrusion Detection System, marketing, medicine, real estate, customer relationship management, engineering, web mining. Applying data mining techniques in data from the educational sector is called Educational Data Mining [2-6]. Applications of EDM are predicting student's or instructor's performance, recommendation system, etc. Machine learning is related to data mining and statistics. Machine can be learning from collection of previous data and acting efficiently in future. It is divided into four types such as Supervised Learning, Unsupervised Learning, Semi Supervised Learning and Reinforcement Learning [7]. Classification and Regression is example for supervised learning, Clustering and Association is example for unsupervised learning, Clustering and Classification is example for semi supervised learning, Control and Classification is example for Reinforcement Learning. Some of the machines learning algorithms are Naïve Bayes, Decision Tree, K-Nearest Neighbor, Support Vector Machine, Random Forest, Logistic Regression, Linear Regression, etc. Some of the algorithms to be discussed in this paper are mentioned here. The decision tree [12,14,15,16] is a tree structure that is used to fabricate the classification models. It partitions a dataset into littler subsets. Leaf node symbolize to a decision. In view of highlight estimations of examples, the decision trees order the instances. Every decision show to a component in an occurrence in a decision tree which is to be grouped, and each branch represents to a worth. Grouping of Instances begins from the root node and arranged dependent on their feature values. Categorical and mathematical information can be taken care of by decision tress. C5.0 calculation is generally utilized as a decision tree strategy in AI. At first, we have ID3.0 algorithm. In view of ID3.0, people created C4.5 algorithm, lastly develop C5.0 algorithm. This kind of decision tree model depends on entropy and data gain. Similarly, Naive Bayesian classifier [12,13,16] is a method for assessing probabilities of individual variable qualities, given a class, from preparing information and afterward permits the utilization of these probabilities to characterize new elements. The Naïve Bayes Classifier procedure is especially fit when the dimensionality of the information sources is high. Notwithstanding its straightforwardness, Naive Bayes can frequently beat more refined arrangement techniques. Naive Bayes model recognizes the qualities of dropout students. It shows the likelihood of each input attribute for the anticipated state. A Naive Bayesian classifier

is a straightforward probabilistic classifier dependent on applying Bayesian hypothesis (from Bayesian measurements) with solid (naive) independence assumptions. The other machine learning algorithm, KNN algorithm [13] is a vigorous and adaptable classifier that is frequently utilized as a benchmark for more unpredictable classifiers, like Artificial Neural Networks (ANN) and Support Vector Machines (SVM). K nearest neighbors is an easy algorithm that stores every accessible case and arranges new cases dependent on a similitude measure (e.g., distance functions). KNN has been utilized in statistical assessment and pattern recognition as of now in the start of 1970's as a non-parametric strategy. The most significant algorithm is Support Vector Machine [15] which attempts to discover a hyperplane to isolate the classes while limiting the characterization mistake and expanding the edges. SVM is a decent order and regression procedure proposed by Vapnik at AT&T Bell Laboratories. Support Vector Machine (SVM) was first heard in 1992, presented by Boser, Guyon, and Vapnik in COLT-92. Support vector machines (SVM) are a set of related supervised learning methods used for classification and regression. They have a place with a group of summed up linear classifiers. In other terms, Support Vector Machine (SVM) is a classification and regression forecast apparatus that utilizations AI hypothesis to boost prescient precision while naturally staying away from over-fit to the information. Random Forest [12] is one of the most generally utilized machine learning algorithms for the purpose of classification. This technique was advocated by Leo Breiman and Adele Cutler, and joins the base standards of bagging with arbitrary feature selection to add extra assorted variety to the decision tree models. It can likewise be utilized for regression model (for example categorical target variable) however it primarily performs well on grouping model. The exhibition of decision trees can be improved by gathering them in a model known as a random forest (or decision tree forests) zeros in just on troupes of decision trees. After the outfit of trees, the forest is created, the model uses a vote to consolidate the trees' forecasts. Random alludes to principally two cycles: 1. Arbitrary perceptions to develop each tree 2. Random factors chose for parting at every node. Two boundaries are significant in the random forest algorithm: 1. Number of trees utilized in the forest (ntree) 2. Number of random factors utilized in each tree (mtry). The three fundamental packages utilized are "caret", "party", "randomForest". For this technique, there is no requirement for cross-validation. Among these methods, this paper predicts that SVM approach is better in terms of accuracy.

2 Related Works

It is mandatory to measure the performance of teacher to improve the effectiveness of teaching and improve the knowledge of students in the field of E-Learning. It is done by using feedback collected from the students. Bertan [8], examined the elements of educators training performance utilizing two diverse data mining techniques for example, stepwise regression and decision trees (CHAID and CART). Data was gathered from learners of MIS department during the time of 2004-2009. Factor investigation was applied to reveal free factors influencing the overall performance of teachers. Stepwise regression model was created using SPSS 19.0 and decision trees were built using Answer Tree 3.1. The after-effect of this study was used in curriculum design, doling out the educator to courses and granting the teachers. Ahmadi [9] examined the feedback from learners about the instructors to shape Teacher Evaluation framework using WEKA tool. Data gathered from 830 understudies around 104 records with five attributes. Decision tree (J48) algorithm applied and the outcomes were utilized by the educationalist to distinguish the specific educator is proceed to the following semester or not. Aranuwa Felix Ola, [10] utilized intelligent strategy for assessment of educators performance in higher institutions and proposed an optimal machine learning algorithm to design a system framework. Formative and Summative assessment methods applied to assess the educator’s performance to increase the quality. Formative evaluation used the estimation of assessing the educator's study hall exercises, appraisal of understudy report, quality and shortcoming of teachers. Summative assessment strategy thought about the meeting, assessment, review and wellspring of documentation. Anwar et al. [11] investigated the practicality of applying data mining procedures to dissect the viability of educators. Data was gathered, nine data mining methods were applied such KNN, NB, SVM, RA, J48, JPip, AdaBoost, BN, Random Forest among these Random Forest and SVM show striking execution. Ajay Kumar Pal, Saurabh Pal [12] framed a framework to anticipate the performance of educator utilizing their assessment, checking the classes and performance assessment of instructor. The data mining strategies utilized in this system was Naive Bayes, ID3, LAD tree and CART in WEKA tool. Three years of data gathered from post graduate students with 14 attributes. The precision created by ID3 was 65.14%, 72.32% by CART, 75.00% by LAD Tree and the most accuracy 80.35% produced by NB classifier. Hemaid, [13] found the path for improving the exhibition of educators utilizing data mining procedures, for example, Rule Induction, Decision Tree, Naive Bayes and K-Nearest Neighbor in the Rapid Miner tool. The dataset gathered about the educators form the Ministry of Education Gaza on training course from the length of 2010 to 2013. It consists of 813 records and 46 attributes in the form of survey. The accuracy delivered by the various methods is 76.23% by rule induction, 77.05% by decision tree, 77.46% by naive bayes and finally acceptable accuracy 79.92% by K-nearest neighbor. Asanbe M.O. et al, [14] assessed and anticipated the performance of educators in higher educational institutions utilizing Decision Tree (C4.5 and ID3), MLP in WEKA data mining tool. The dataset was gathered from the Academic Department of a University, South West Nigeria with 350 records of 2010 to 2015. The accuracy delivered by C4.5 was 83.5%, 82.5% by MLP and

71% by ID3.Agaoglu, [15] assessed the performances of teacher dependent on view of student using questionnaire of course evaluation using four distinctive data mining techniques, for example, Support Vector Machines, Decision Tree algorithms (C5.0, CART), Discriminant Analysis and Artificial Neural Networks (ANN-2QH, ANN-3QH, ANN-MH). The performance measurements utilized were recall, precision, accuracy, and specificity. Also, variable importance was done to eliminate the immaterial attributes. At last this work indicated the expressiveness and adequacy of models of data mining in higher education. Dataset gathered from Marmara University, Istanbul, Turkey. The information contains 2850 records, 25 attributes, and one class name. Among various strategies C5.0 delivered high accuracy as 92.3%.Ahmed Mohamed Ahmed et al, [16] examined the parts which are chiefly influencing the accomplishment of learners for foreseeing the performance of educator to increase the quality of the educational system using various methods like Multilayer Perception, J48 Decision Tree, Sequential Minimal Optimization and Naïve Bayes. Dataset collected from University of CaliforniaIrvine (UCI) Machine Learning Repository. There were 32 attributes gathered from Q1 to Q28 asked with reactions from 1 to 5. Attributes are assessed utilizing OneR, eight attributes are chosen as highly impact. Algorithms are executed with all attributes and with profoundly affected attributes only. J48 delivered 84.8% with all the attributes and SMO produces 85.8% for chosen attributes. Surjeet Kumar, [17] increased the effectiveness and dependability of evaluation system of educator's performance utilizing three diverse data mining strategies, for example, Decision Trees J48 and lazy IBK and Meta Bagging in WEKA tool. The dataset collected from UCI Machine Learning Repository Teaching Assistant for 2 summer semesters and 3 regular semesters of insights concerning 151 instructors of Department of Statistics of the University of Wisconsin-Madison. IBK delivered impressive performance 62.2% than the J48 and Bagging algorithms. According to Dabalen, [18] an enormous mismatch seems to exist between university output and labor market request lately. Their discoveries show that the performances of late alumni have unmistakably deteriorated, fundamentally in view of the operational approaches and deficient degree of skilled human resources, particularly the nature of university-trained bit of the work power. Decaying quality recognition is likewise upheld by the outcomes from empirical research. Glazerman et al, [19] centered the organization made state backing of thorough educators' assessment systems a pre-condition for rivalry in "Race to the Top", and have spread out a diagram for the reauthorization demonstration in which instructor viability characterized by evaluation of hands on execution is a significant feature. Self-declared education reformers such as Bill Gates, Davis Guggenheim and Michelle Rhee in their accommodation placed that instructor's evaluation ought to be at the cutting edge of the training change plan and that evaluation results be utilized as the reason for settling on decisions about employing, restraining, compensating, granting residency to and authorizing ineffective educators. One of the reasons for this may not be fantastical from the way that the quality of good education in any educational foundation relies upon the nature of the academic staff in that organization; and there is no acceptable substitute for skillful staff that has sound educational way of thinking and dynamic initiative. As the most valuable resource in schools, instructors are crucial to increase student results and improve education standards. Subsequently, improving the effectiveness and value of tutoring depends, in huge measure, on guaranteeing that educators are profoundly talented, well resourced, and spurred to perform at their best. From this viewpoint, teachers' performance evaluation is a crucial step in the drive to increase the viability of learning system and increase educational standards. DeNisi, [20], a focal purpose behind the work of performance evaluation is performance improvement (at first at the degree of the individual workforce, and at last at the degree of the establishment). Other principal reasons incorporate reason for employment decisions. Furthermore, performance evaluation can help in the plan of measures and determination of people who are most appropriate to perform required organizational tasks. It can be part of managing and observing worker career development and improvement. Abaidullah, et al, [21] as proposed system that uses k-means clustering for investigation of learner's feedback information for building effective decision by upper level board members liable for watching and making appraisal of educational quality and for increasing teaching quality and improving nature of students' learning experience. They have talked about model for utilizing clustering to improve students' learning experience that would bring about progress in nature of instructive condition of establishment. Baradwaj [22] proposed system that depicts performance of student in semester end examination which is useful in recognizing potential dropouts and students who need extraordinary consideration, so those educators can suitable preparing/encouraging to them. The decision tree strategy algorithm was applied by the creators on students' database to anticipate the division of individual student. The study made by them will help to both students and teachers to improve individual division of student just as to improve by and large aftereffect of a class. This study will likewise attempt to recognize those students who need extraordinary regard to diminish drop-out rate and making essential and legitimate move for next semester examination. Their study demonstrated that the grade attained from senior secondary school examination (SSCE) in mathematics is the uppermost determinant of learner's performance utilizing the C4.5 learning algorithm in building the model of the student's performance. The Table 1 depicts the work done by the existing systems that is types of techniques, tool was used, sample data taken and finally results yield.

Table 1. Literature Survey

Authors	Data Mining Techniques	Tool Used	Results (%)	Sample Data
[12]	• Naïve Bayes, • Decision Tree	WEKA	NB-80.35 ID3-65.17 CART-72.32 LAD-75.00	3 Years data 14 Attributes
[13]	• Rule Induction, • Nearest Neighbor, • Naïve Bayes	Rapid Miner	RI-76.23 KNN-79.92 NB-77.46	2010-2013 data, 813 records ,46 Attributes
[14]	• Neural Network, • Decision Tree	WEKA	ID3-71 C4.5-83.5 MLP-82.5	2010-2015 data, 350 records ,12 Attributes
[15]	• Decision Tree, • Support Vector Machine, • Neural Network, • Discriminant Analysis	IBM SPSS Modeler-Quick and Multiple Classifiers	C5.0-92.3 CART-89.9 SVM-91.3 ANN-Q2H-91.2 ANN-Q3H-90.8 ANN-M-90.5 DA-90.5	2850 data, 26 Attributes
[16]	• Decision Tree, • Neural Network, • Naïve Bayes, • Sequential Minimal Optimization	WEKA	J48 DT - 84.8 NB - 83.3 SMO - 84.5 MLP-82.5	5820 data, 33 Attributes
In this paper	• Naive Bayes, • Nearest Neighbor, • Decision Tree, • Random Forest, • Support Vector Machine	R	NB -87.7 KNN – 91.7 C5.0 – 94.2 RF- 98.09 SVM – 99.25	2220 data ,21 Attributes

3 Analysis of the Machine Learning Algorithms

Data collection is the first step from various sources; the data is preprocessed and fills the empty columns. For implementation data is loaded or imported into R Programming using library functions read(). Then the data is divided into training and testing data set in 70% and 30% ratio. The machine learning algorithms are applied in training data set to create the models and the created model is tested in testing data to measure the performance of the model. Improve the accuracy by fine tuning the model by changing the parameters. Finally, the predicted outcome will be produced. The model learns from the data if new data comes the model predict the result. The process starts at data collection and ends at prediction of knowledge. The student evaluation data have 21 attributes in which 20 independent variables (from Q1 to Q20) and 1 class variable (Q21). The response values of these questions are in the form {1,2,3,4,5} where 1,2,3,4,5 represents the "Poor", "Fair", "Good", "Very Good", and "Excellent" respectively for Q1 to Q20, and the values of Q21 in the form {1,2,3} where 1 stands for "excellent", 2 stands for "Good" and 3 stands for "Poor". The questionnaire is divided into five parts. Q1 to Q5 belongs to planning and organization of teacher. Q6 to Q10 deals with the presentation and communication of concepts. Q11 to Q15 explains about student's active participation in the class. Q16 to Q20 deals with the class management assessment of students. The last class variable Q21 predicts the performance of the instructor. The possible values for Q21 are 'Excellent' (1456 Observations) 'Good' (492 observations) 'Poor' (272 observations). These descriptions of the dataset and its possible values are shown in Table 2. The sample data is in figure 1 and plot of the data using pairs. panels() are shown in figure 2.

Table 2. Description of the data set

Variable	Description	Possible Values
	PLANNING AND ORGANIZATION	
Q1	Teacher comes to the class on time	{1,2,3,4,5}
Q2	Teaching is well planned	{1,2,3,4,5}
Q3	Aims / Objectives made clear	{1,2,3,4,5}
Q4	Subject organized in logical sequence	{1,2,3,4,5}
Q5	Teacher well prepared in the subject	{1,2,3,4,5}
	PRESENTATION AND COMMUNICATION	
Q6	Teacher speaks clearly and audibly	{1,2,3,4,5}
Q7	Teacher writes and draws legibly	{1,2,3,4,5}
Q8	Teacher provides explanations are clear and effective	{1,2,3,4,5}
Q9	Teacher's level of Instruction is suited to the ability of students	{1,2,3,4,5}
Q10	Teacher helps and counseling to the needy students	{1,2,3,4,5}
	STUDENTS PARTICIPATION	
Q11	Teacher asks questions to promote interaction and reflective thinking	{1,2,3,4,5}
Q12	Teacher encourages questioning raising doubts by students	{1,2,3,4,5}
Q13	Teacher ensures learner activity and problem-solving ability	{1,2,3,4,5}
Q14	Teacher encourages, compliments originality and creativity of students	{1,2,3,4,5}
Q15	Teacher is courteous and Impartial in dealing with the students	{1,2,3,4,5}
	CLASS MANAGEMENT / ASSESSMENT OF STUDENTS	
Q16	Teacher engages classes regularly and maintains discipline	{1,2,3,4,5}
Q17	Teacher covers the syllabus completely and at appropriate pace	{1,2,3,4,5}
Q18	Teacher holds tests regularly which are helpful to students	{1,2,3,4,5}
Q19	Teacher's marking of scripts is fair impartial	{1,2,3,4,5}
Q20	Teacher is valuing and returning the answer scripts	{1,2,3,4,5}
	PERFORMANCE	
Q21	The Performance of the Instructor	{1,2,3}

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

randomforest.r DATA2

Filter

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20	performance
147	5	5	5	4	5	5	4	5	4	5	5	4	5	5	5	5	1	4	4	1	EXCELLENT
148	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	EXCELLENT
149	3	3	3	3	2	2	3	3	3	2	2	2	3	3	2	4	4	3	2	3	POOR
150	2	3	2	2	1	1	2	1	2	2	2	2	3	1	2	2	1	2	2	1	POOR
151	4	4	4	4	4	4	5	4	5	5	5	5	5	5	5	4	4	4	4	4	EXCELLENT
152	5	5	1	2	3	4	2	3	2	3	4	3	4	4	5	5	4	4	5	5	GOOD
153	5	5	5	4	5	5	5	4	5	5	5	5	5	5	5	5	5	5	5	5	EXCELLENT
154	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	4	5	EXCELLENT
155	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	EXCELLENT
156	4	4	4	4	4	4	4	4	4	4	3	3	4	3	3	4	4	4	4	4	GOOD
157	5	4	4	4	4	3	3	4	4	4	3	3	4	4	4	4	4	4	4	4	GOOD
158	4	4	5	4	4	4	3	4	4	4	3	3	4	4	4	3	3	4	3	5	GOOD

Showing 147 to 159 of 2,220 entries

Figure 1. Sample Data

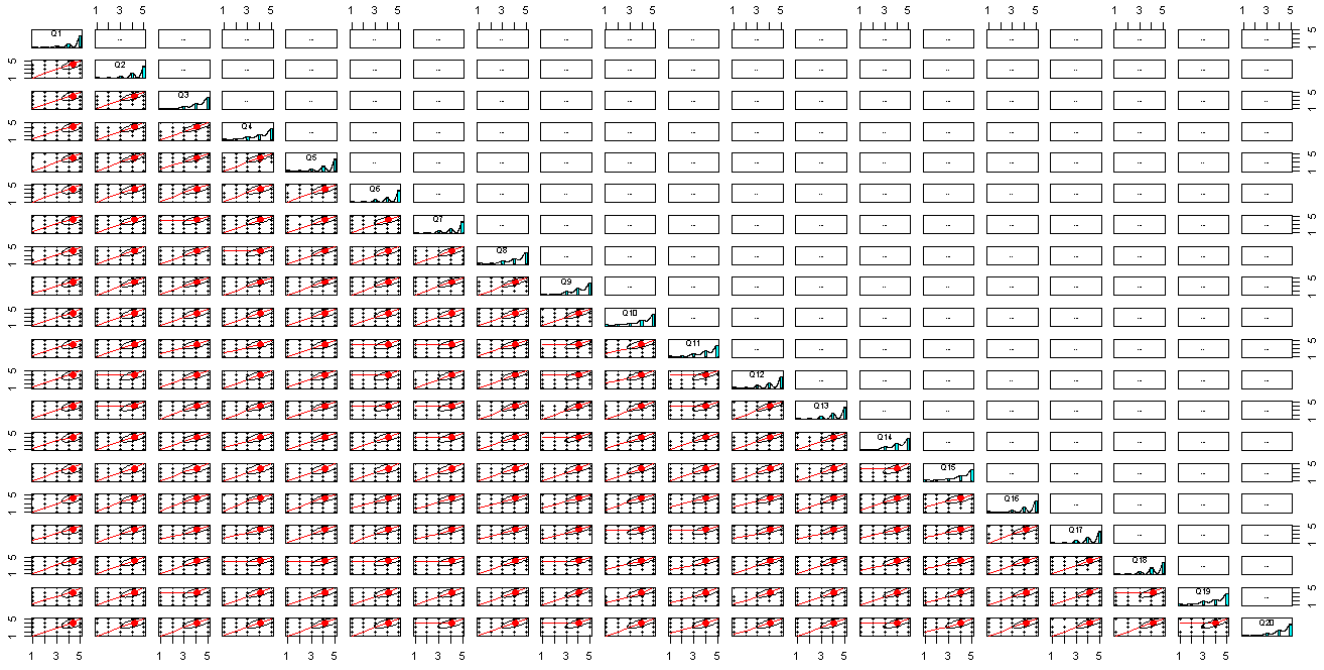

Figure 2. Plot of the sample data

4 Performance measures

There are some performance measures to evaluate the proposed model for the correctness of the system. The values of class variables are positive (+), Negative (-), Actual Positive that is correctly classified as positive is named as True Positive (TP) while Actual Positive that is incorrectly classified as negative is False Negative (FN). Likewise, Actual Negative that is correctly classified as positive is named as False Positive (FP) whereas Actual Negative that is incorrectly classified as negative is named as True Negative (TN). The confusion matrix is shown in Table 3. The calculations of performance measures such as Accuracy, Precision, Recall, Specificity, and Sensitivity are done by the terms of Confusion Matrix.

Table 3. Confusion Matrix

	Predicted +	**Predicted -**	**Total**
Actual +	True Positive (A)	False Positive (B)	A+B
Actual -	False Negative (C)	True Negative (D)	C+D
Total	A+C	B+D	N

Accuracy measures the percentage of test set samples that proportion of correctly classified instances for all instances. As mentioned in equation (1), it is relied upon to be more like 1 if performance of our model is admirably.

$$\text{Accuracy} = \frac{A + D}{(A + B + C + D)} \quad (1)$$

The incorrectness of predicted output values is called as error of the method. If target values are categorical, the error is expressed as an error rate.

Error rate is measuring the proportion of incorrectly classified instances for all instances or it is calculated by 1-Accuracy and it is given in equation (2).

$$\text{Error rate} = \frac{B + C}{(A + B + C + D)} \quad (2)$$

Kappa measures the problems on imbalances class of dataset as in equation (3).

Observed Accuracy - Add the number of instances with the truth label and divide by the total number of instances.

Expected Accuracy - Related to the number of instances of each class, along with the number of instances with the truth label.

$$\text{Kappa} = \frac{\text{Observed Accuracy} + \text{Expected Accuracy}}{1 - \text{Expected Accuracy}} \quad (3)$$

Recall or **Sensitivity** as in equation (4) is the measure of correctly classified positive instances to a total number of positive instances.

$$\text{Sensitivity or Recall} = \frac{A}{A + C} \quad (4)$$

Specificity is the measure of correctly classified negative instances to a total number of negative instances. It is calculated using equation (5).

.

$$\text{Specificity} = \frac{D}{B + D} \quad (5)$$

Balanced accuracy uses specificity and sensitivity using equation (6).

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \quad (6)$$

Precision measures the proportion of correctly classified instances for those instances that are classified as positive. It is mentioned in equation (7).

$$\text{Precision} = \frac{A}{A + B} \quad (7)$$

Prevalence as in equation (8).is a single number value or the rate of the matrix of the positive class of the data.

$$\text{Prevelance} = \frac{A + C}{(A + B + C + D)} \quad (8)$$

Detection Rate as depicted in the equation (9), is the number of correctly predicted positive class finished as an amount of all the predictions made.

$$\text{Detection rate} = \frac{A}{(A + B + C + D)} \quad (9)$$

Detection Prevalence as portrayed in the equation (10), is the how many numbers of predictions of positive classes made as a fraction of all predictions.

$$\text{Detection Prevelance} = \frac{A + B}{(A + B + C + D)} \quad (10)$$

Positive Predictive Value (PPV), Specificity directly proportional to PPV. It is mentioned in the equation (11),

$$\text{PPV} = \frac{\text{Sensitivity} * \text{Prevelance}}{((\text{Sensitivity} * \text{Prevelance}) + ((1 - \text{Specificity}) * (1 - \text{Prevelance}))} \quad (11)$$

Negative Predictive Value (NPV), Sensitivity directly proportional to NPV. It is mentioned in the equation (12),

$$\text{NPV} = \frac{\text{Sensitivity} * (1 - \text{Prevelance})}{((\text{Sensitivity} * \text{Prevelance}) + ((1 - \text{Specificity}) * (1 - \text{Prevelance}))} \quad (12)$$

OOBError – Out-Of-Bag Error (Misclassification error is otherwise called as OOB error which is mentioned as a percentage).

5 Results and discussion

The proposed system was implemented in R programming dialect or basically R was outlined [25] by Ross Ihaka and Robert Gentleman in 1993 as a successor to S, however, is directly made by the R Development Core Team since 1997. The name R is from the principal primary names of both R makers. It is basically used for data analysis, statistical computing, and graphics and has transformed into the reference working programming apparatus in many fields of creative work. The RStudio will be used as the GUI to build the R code for decision tree. The RStudio involves four basic window sheets Scripts, Workspace, Plots, and Console. Scripts fills in as an area to compose and save R code, Workspace records the datasets and variables utilized as a part of R environment, Plots window is utilized to demonstrate the plots made by the code and gives a system to send out the plots, Console exhibits the historical backdrop of the executed RCode and the output [26]. R is wholeheartedly available under the GNU General Public License. The core of R is a translated code which permits branching and looping and in addition modular programming using functions. R grants blend with the procedures written in the C, C++, .Net, Python or FORTRAN dialects for viability [27].The variables are assigned with R-Objects and the data type of the R-object turns into the data type of the variable. There are many sorts of R-objects. The occasionally used are Vectors, Lists, Matrices, Arrays, Factors, and Data Frames. In R, a variable itself is not proclaimed of any data type, rather it gets the data type of the R -object consigned to it. So R is known as a dynamically typed language. It is rich in built-in operators, looping and decision-making statements. R has an extensive number of in-built functions and the client can make their own functions. For any R work, you can get to an assistance record by basically writing a question mark, trailed by the name of the function or the help function. R packages are amassing of R functions, complied code and sample data. They are secured under a directory called "library" in the R environment. The R essential programming incorporates around 30 packages; however there are more than 4000 additional packages available on the CRAN page [28].IEEE Spectrum has distributed its fourth yearly positioning of best programming languages [29], and the R language is again included in the Top 10. In the year 2017 R positions at #6 rank, in 2016 at #5 rank, in 2015 at #6 rank and in 2014 at #9 rank. They have mentioned that R was designed for programming statistical analysis and data mining applications. In this paper, we primarily concentrate on classification problems using machine learning algorithms. The R packages are used to implement the proposed work. In the unsupervised mode the clusters are designed, by trying some clusters (nominally 2) and measuring the vicinity between data points. There are many tasks such as Data Collection, Data Preprocessing, Data Selection, Model Construction involved. In step one; Data were collected from the students of Christ College of Engineering and technology. Sample contains details about college identity, course, year and subject of 2220 observations with 21 attributes. In step two, missing values were replaced with the mean value instead of 'NA'. In step three, we have selected only relevant attributes by omitting others which are not affecting the performance. Imported the selected data into R Environment and converted predicted variable 'performance' numeric into factor using as.factor() with three classes. Then divided the dataset (2220 samples) into 70% (1593) as training set and 30% (627) as testing set. In step four, Load the necessary packages and implemented Machine learning algorithms such as Decision Tree, Naïve Bayes, K-Nearest Neighbor, Random Forest and Support Vector Machine on training set and evaluate the performance of the algorithms using testing data.

5.1 K-Nearest Neighbor

The e1071 package is available to implement the KNN model. Figure 3 shows the KNN model with accuracy as y axis and number of neighbors as x axis.The table 4 shows the description of KNN model in which the highest accuracy and kappa value produced when the K value is 5.

Table 4. Accuracy and Kappa of KNN.

K	Accuracy	Kappa
5	0.9353196	0.8687417
7	0.9346648	0.8668327
9	0.9301597	0.8569912
11	0.9209523	0.8373292
13	0.9115271	0.8160879
15	0.9050962	0.8017174
17	0.8993228	0.7887571
19	0.8961204	0.7812665
21	0.8941836	0.7759753
23	0.8894771	0.7652453

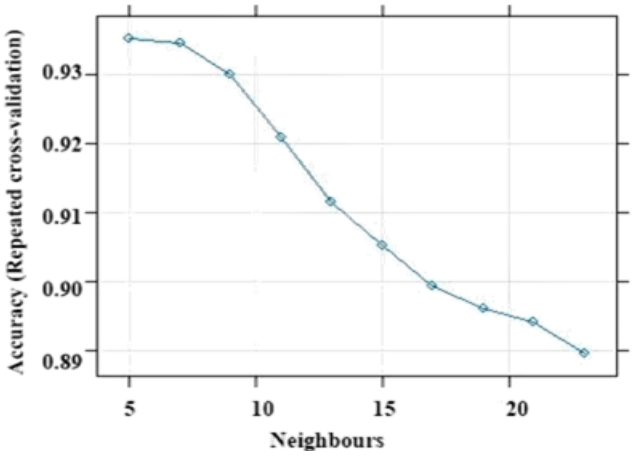

Figure 3. K-Nearest Neighbor Model.

5.2 Support Vector Machine

The parameters for SVM model are createDataPartition() for partition the dataset into testing and training set. trainControl() with Method="repeatedcv", number=10, repeat=3. train() with Method="svmLinear". Predict() is used to predict the performance of SVM model on testing data. ConfusionMatrix() generate the Confusion Matrix of SVM model. Finally plot() is used to display the plot of SVM model. The performance plot of Support Vector Machine model denotes the darker region represents high accuracy and the lighter region represents low accuracy. The overall statistics of Support Vector Machine model with the accuracy of 99.25% and kappa value of 0.985. The plot of support vector machine with cost as x axis and accuracy as y axis shown in Figure 4 When the cost value is 5 the accuracy is highest .The dotted circle represents the cost values are 0.01, 0.05, 0.10, 0.25, 0.50, 0.75, 1.00, 1.25, 1.50, 1.75, 2.00, and 5.00 in the Figure 5, the performance of the SVM is shown in the graphical representation.

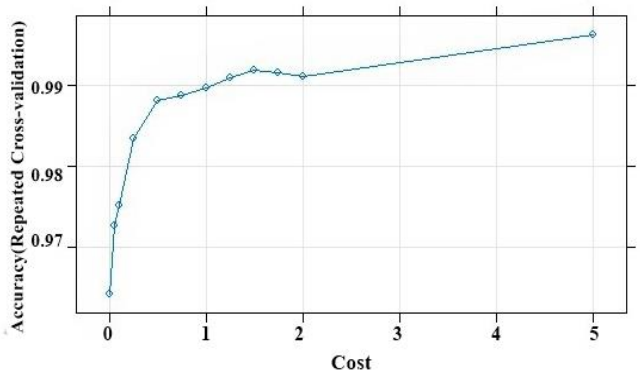

Figure 4 Support vector machine Model.

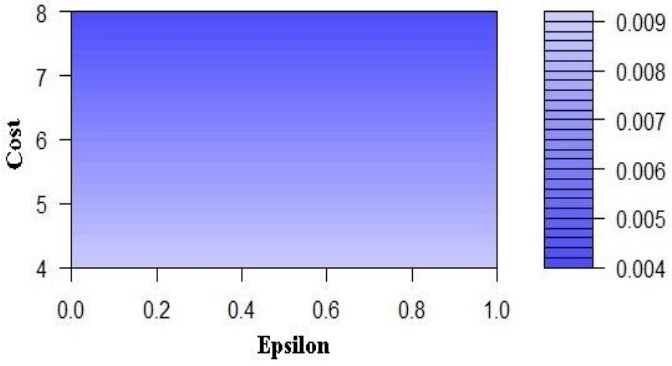

Figure 5 Performance of support vector machine

5.3 Random Forest

The needed parameter to implement Random forest model are createDataPartition() method used to partition the overall dataset in to training and testing dataset. randomForest() method create random forest for number of trees 500. The predict() and plot() method used to predict and display the plot of random forest model. confusionMatrix() method create the confusion matrix and importance() calculates the variable importance value and varImpPlot() method display the variable importance plot. Figure 6 denotes the plot of variable importance. The highest importance variable is Q13 and the least importance variable is Q14. It is represented by MeanDecreaseGini

Table 5. MeanDecreaseGini of variables

Variable	MeanDecreaseGini
Q1	14.56934
Q2	15.77667
Q3	41.59601
Q4	43.16799
Q5	48.65297
Q6	20.2727
Q7	27.50709
Q8	43.82578
Q9	70.71042
Q10	70.06157
Q11	38.06109
Q12	47.28222
Q13	**92.9755**
Q14	**14.42197**
Q15	28.6155
Q16	34.07262
Q17	28.86567
Q18	25.32717
Q19	40.48268
Q20	53.38237

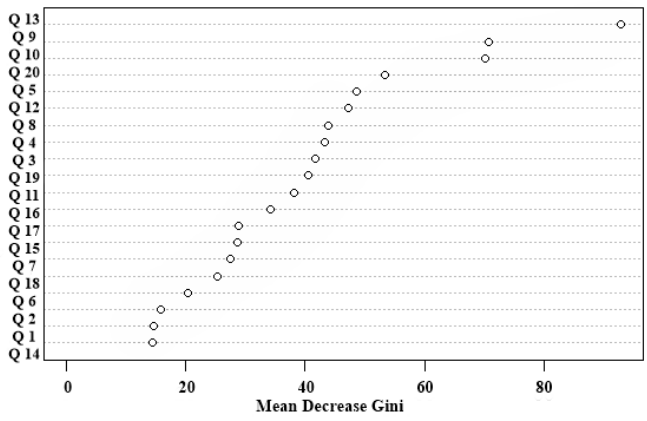

Figure 6 Plot of MeanDecreaseGini

Table 5 shows the MeanDecreaseGini value of all the attributes. Q13 have highest value 92.9755 and Q14 have lowest value 14.42197. The Plot indicates the error for the different classes with color and out-of-bag error over number of trees shown in figure 7. The colors are EXCELENT (Blue), GOOD (Green) POOR (Red) and OOB Error (Black).

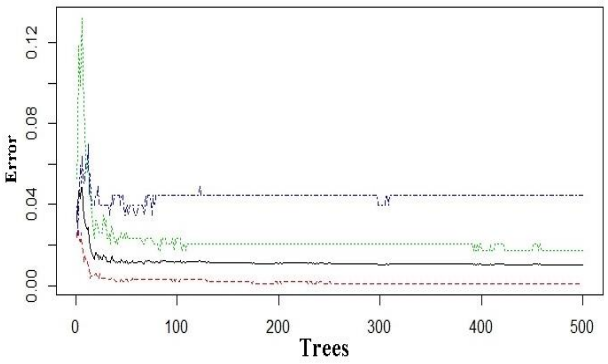

Figure 7 Plot of random forest

5.4 Decision Tree (C5.0)

The packages C5.0, caret installed using install.packages() method. The methods used are createDataPartition() for partition the data into training and test set, C5.0() for Create the Decision Tree, predict() for predict the performance of Decision Tree, confusionMatrix() for generate Confusion Matrix, and plot() for display the decision tree. Figure 9 shows the model of the decision tree.

5.5 Naive Bayesian

The dataset is converted into frequency table. Based on probability likelihood table was created. Calculate posterior probability using the Naive Bayes equation for each class. The highest value of posterior probability leads to outcome of prediction. The e1071 package is used to implement the Naive Bayes algorithm. By the utilization of Bayesian hypothesis, we can compose.

$$P(A/B) = \frac{\mathrm{P(B/A)P(A)}}{\mathrm{P(B)}}$$

This algorithm produces 87.7% as accuracy. Figure 10(a) and 10(b) shows the plot of Naive Bayes model for all the independent attributes. Red line indicates excellent, green line for good and blue line for poor class. Libraries used are naivebayes, dplyr, ggplot2 and psych to implement Naive Bayes algorithm and to plot the model.

Finally, the confusion matrix for different Machine learning algorithms is shown in table 6. It is constructed on the testing data. Mark the Accuracy, Kappa values in Table 7. The high accuracy and kappa values are for SVM. The minimum accuracy and kappa values are for Naive Bayes. The figure 8 clearly depicts the comparison chart of performances of algorithms. We calculated the measurements such as Specificity, Sensitivity, Pos Pred Value, Neg Pred Value, Prevalence, Detection Rate, Detection Prevalence, and Balanced Accuracy for different algorithms. They are shown in Table 8.

Table 6. Comparison of confusion matrix of all classifiers

Classifiers		Excellent	Good	Poor	Accuracy
	Excellent	393	17	0	
DT	Good	14	130	3	94.26%
	Poor	0	2	68	
	Excellent	399	7	0	
RF	Good	8	140	0	97.29%
	Poor	0	2	71	
	Excellent	427	26	0	
KNN	Good	9	115	14	91.72%
	Poor	0	6	67	
	Excellent	372	0	0	
NB	Good	66	139	15	87.82%
	Poor	0	0	73	
	Excellent	436	5	0	
SVM	Good	0	142	0	**99.25%**
	Poor	0	0	81	

Table 7. Comparison of Accuracy and KAPA

METHODS	Accuracy	Kappa
NB	87.82%	0.777
KNN	91.72%	0.832
DT (C5.0)	94.26%	0.887
RF	97.29%	0.963
SVM	**99.25%**	**0.985**

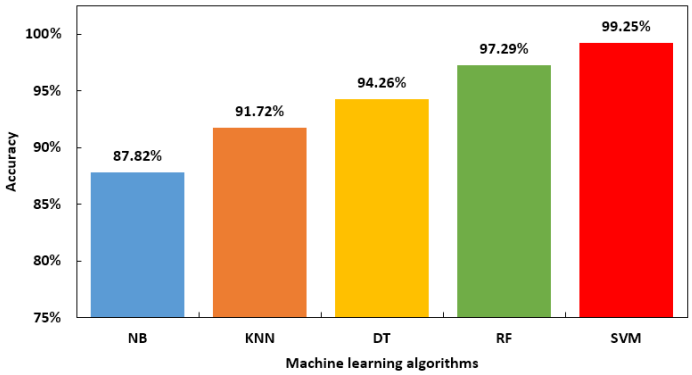

Figure 8 Comparison of accuracy of classifier

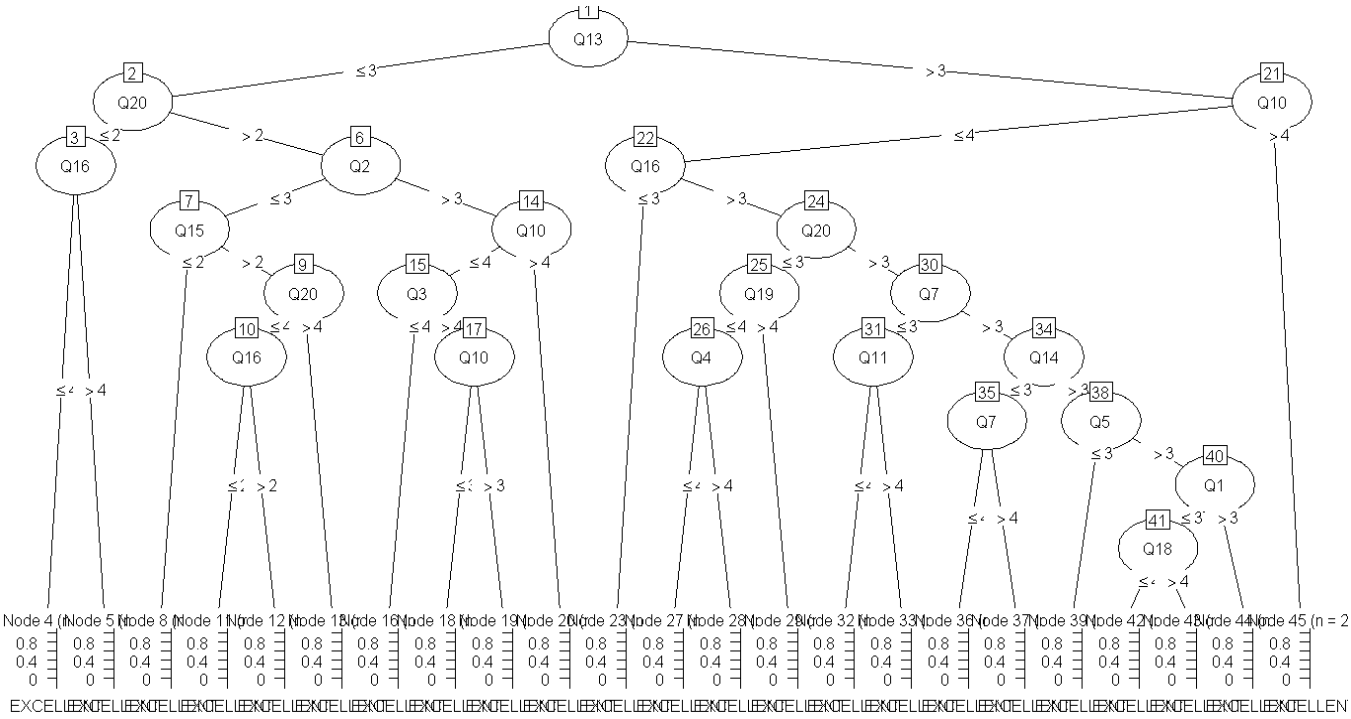

Figure 9. Plot of Decision Tree Model

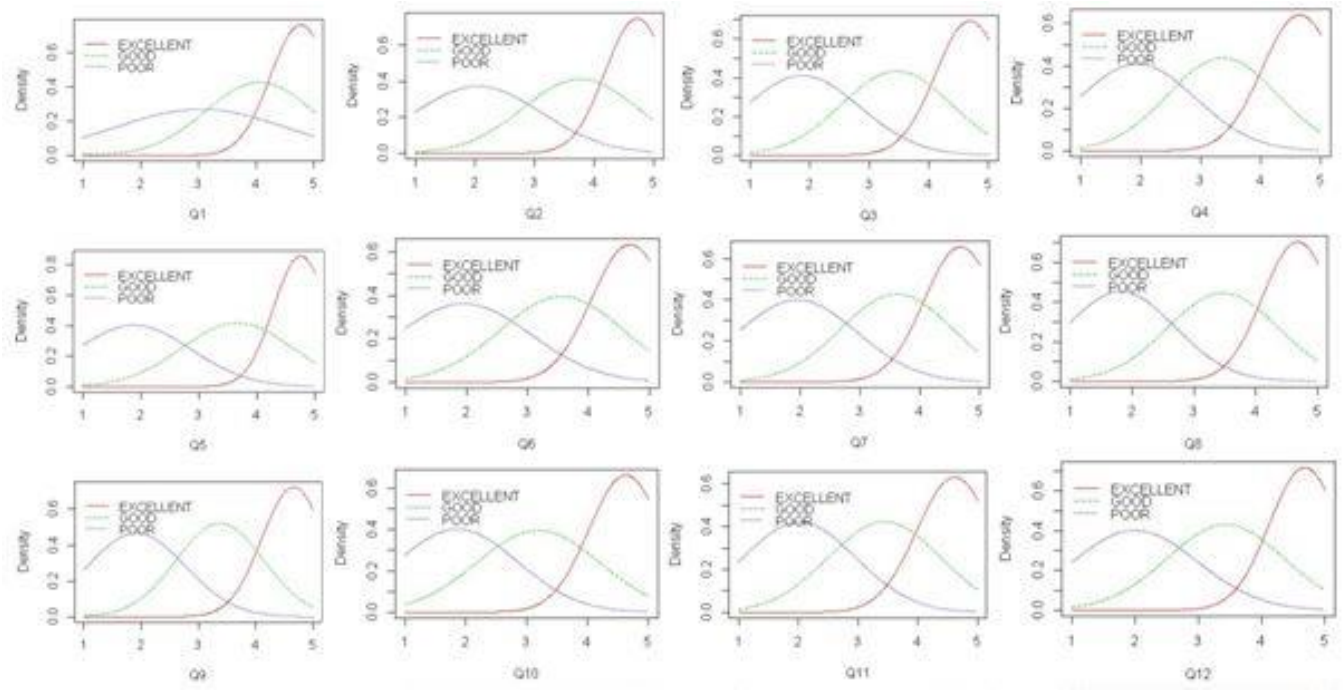

Figure 10(a). Naive Bayes Model for Q1-Q12

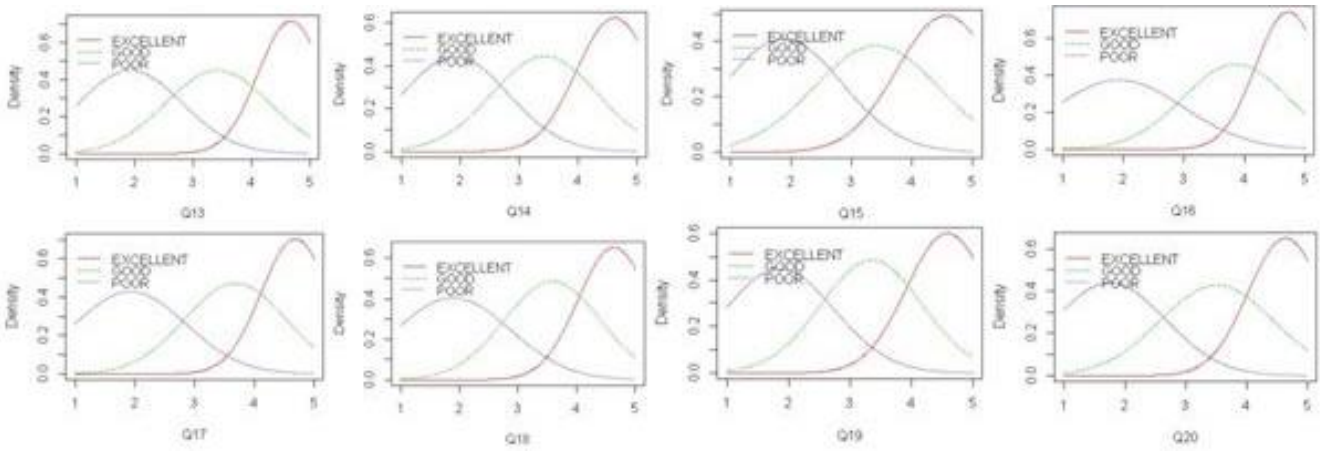

Figure 10(b). Naive Bayes Model for Q13-Q20

Table 8. Measurement with different Machine Learning Algorithms.

METHODS	Class	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Prevalence	Detection Rate	Detection Prevalence	Balanced Accuracy
NAIVE BAYES	EXCELLENT	0.8443	1.0000	1.0000	0.7899	0.6307	0.5325	0.5325	0.9221
	GOOD	1.0000	0.8244	0.6274	1.0000	0.2282	0.2282	0.3638	0.9122
	POOR	0.7353	1.0000	1.0000	0.9583	0.1411	0.1037	0.1037	0.8676
K NEAREST NEIGHBOR	EXCELLENT	0.9794	0.8860	0.9426	0.9573	0.6566	0.6431	0.6822	0.9327
	GOOD	0.7823	0.9555	0.8333	0.9392	0.2214	0.1732	0.2078	0.8689
	POOR	0.8272	0.9897	0.9178	0.9763	0.1220	0.1009	0.1099	0.9084
C5.0	EXCELLENT	0.9656	0.9227	0.9585	0.9355	0.6491	0.6268	0.6539	0.9442
	GOOD	0.8725	0.9644	0.8844	0.9604	0.2376	0.2073	0.2344	0.9185
	POOR	0.9577	0.9964	0.9714	0.9946	0.1132	0.1085	0.1116	0.9771
RANDOM FOREST	EXCELLENT	0.9853	0.9818	0.9901	0.9730	0.6491	0.6396	0.6459	0.9835
	GOOD	0.9732	0.9833	0.9477	0.9916	0.2376	0.2313	0.2440	0.9782
	POOR	0.9718	1.0000	1.0000	0.9964	0.1132	0.1100	0.1100	0.9859
SUPPORT VECTOR MACHINE	EXCELLENT	1.0000	0.9781	0.9887	1.0000	0.6566	0.6566	0.6642	0.9890
	GOOD	0.9660	1.0000	1.0000	0.9904	0.2214	0.2139	0.2139	0.9830
	POOR	1.0000	1.0000	1.0000	1.0000	0.1220	0.1220	0.1220	1.0000

Conclusion

The instructor performance prediction is obligatory in the field of educational sector to show signs of improvement the nature of training and knowledge of students. This process is a lot of accommodating to any educational institution to think about the performance of instructors and to take decision for performance appraisal. This is finished with the assistance of feedback collection from the learners about the mentor's introduction, study materials, taking class for getting level, and so on there are twenty questions in the feedback form. Students need to give score for instructor's presentation from 1 to 5 for all the questions. This dataset gathered from the students of Christ College of Engineering and Technology. The dataset is grouped into 70% as training dataset and 30% as testing dataset. Machine Learning Algorithms are utilized to anticipate the presentation via preparing the system with training dataset and testing the system with testing dataset. The executions are done in R Programming which contains colossal number of packages and methods. Subsequent to preparing, the machine learning classifiers are equipped for creating the outcome for new data. Supervised machine learning algorithms utilized to make the model, for example, K-Nearest Neighbor, Decision Tree, Naive Bayes, Support Vector Machines, and Random Forest with the same dataset. The classifiers tried with testing dataset and discover the accuracy of the models utilizing confusion matrix. The accuracy comparison obviously shows that Support Vector Machine classifier produce most noteworthy accuracy than other machine learning models. These discoveries are helpful to educationalist to improve their performances.

REFERENCES

[1] Han, J. and Kamber, M. Data Mining: Concepts and Techniques, 2nd edition. The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor, 2006.

[2] Romero, C., and Ventura, S. (2007), "Educational Data Mining: A survey from 1995 to 2005" Expert Systems with Applications. Vol. 33, pp.135-146

[3] C Romero, S Ventura, E García. (2008) "Data mining in course management systems: Moodle case study and tutorial" Computers & Education 51 (1), 368-384.

[4] Enrique García, Cristóbal Romero, Sebastián Ventura, Carlos De Castro. (2011). "A collaborative educational association rule mining tool" The Internet and Higher Education, 14(2) 77-88.

[5] Alejandro Pena-Ayala (2014). "Educational Data Mining: A Survey and a Data Mining-Based Analysis of Recent Works" Expert Systems with Applications, 41, pp. 1432–1462

[6] Cristobal Romero, Sebastian Ventura. (2013), "Data mining in education" Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 3(1) 12-27.

[7] V. Vijayalakshmi and K. Venkatachalapathy, "Machine Learning Algorithms and Open Source Tools for Data Mining", Journal of Computational and Theoretical Nanoscience, Vol. 16, 1350–1355, 2019.

[8] Mardikyan, Sona, and Bertan Badur. "Analyzing teaching performance of instructors using data mining techniques" Informatics in Education 10.2 (2011): 245.

[9] Ahmadi, Fateh, and ME Shiri Ahmad. "Data Mining in Teacher Evaluation System using WEKA." *International Journal of Computer Applications* 63.10 (2013).

[10] Ola, A., and Sellapan Pallaniappan. "A data mining model for evaluation of instructors' performance in higher institutions of learning using machine learning algorithms." *International Journal of Conceptions on Computing and Information Technology* 1.1 (2013).

[11] Yahya, Anwar Ali, Addin Osman, and Mohamed Khairi. "MINING EDUCATIONAL DATA TO ANALYZE TEACHING EFFECTIVENESS." *Journal of Theoretical and Applied Information Technology* 89.1 (2016): 267.

[12] Pal, Ajay Kumar, and Saurabh Pal. "Evaluation of teacher's performance: a data mining approach." *IJCSMC* 2.12 (2013): 359-369.

[13] Hemaid, R., and Alaa M. El-Halees. "Improving Teacher Performance using Data Mining." *International Journal of Advanced Research in Computer and Communication Engineering* 4.2 (2015).

[14] Asanbe, M. O., A. O. Osofisan, and W. F. William. "Teachers' Performance Evaluation in Higher Educational Institution using Data Mining Technique." *International Journal of Applied Information Systems (IJAIS)–ISSN*: 2249-0868.

[15] Agaoglu, Mustafa. "Predicting Instructor Performance Using Data Mining Techniques in Higher Education." *IEEE Access* 4 (2016): 2379-2387.

[16] Ahmed, Ahmed Mohamed, Ahmet Rizaner, and Ali Hakan Ulusoy. "Using data mining to predict instructor performance." *Procedia Computer Science* 102 (2016): 137-142.

[17] Surjeet Kumar, (2017). A Modern Data Mining Method for Assessment of Teaching Assistant in Higher Educational Institutions. International Journal of Computer Science and Information Technologies Vol. 8(3), pp. 424-429.

[18] Dabalen, Andrew, Bankole Oni, and Olatunde A. Adekola. (2001). Labor market prospects for university graduates in Nigeria. Higher Education Policy, Vol. 14(2), pp. 141-159.

[19] Glazerman, S., Loeb, S., Goldhaber, D. D., Raudenbush, S., & Whitehurst, G. J. (2010). Evaluating teachers: The important role of value-added (Vol. 201, No. 0). Washington, DC: Brown Center on Education Policy at Brookings.

[20] DeNisi, Angelo S., and Robert D. Pritchard. (2006). Performance appraisal, performance management and improving individual performance: A motivational framework. Management and organization review, Vol. 2(2), pp. 253-277.

[21] Abaidullah, Anwar Muhammad, Naseer Ahmed, and Edriss Ali. (2015). Identifying Hidden Patterns in Students" Feedback through Cluster Analysis. International Journal of Computer Theory and Engineering, 7(1), 16.

[22] Baradwaj, Brijesh Kumar, and Saurabh Pal (2011). Mining educational data to analyze students' performance. International Journal of Advanced Computer Science and Applications Vol. 2(6), pp. 63-69.

[23] Chin Chia Hsu and Tao Huang. (2006). The use of Data Mining Technology to Evaluate Student's Academic Achievement via multiple Channels of Enrolment. An empirical analysis of St. John's University of Technology.

[24] Olamiti, A. O., and Osofisan, A. O. (2009). Academic Background of Students and performance in a computer science programme in a Nigerian university. in European Journal of Social Science, 9(4).

[25] Flor´eal Morandat, Brandon Hill, Leo Osvald, and Jan Vitek. "Evaluating the Design of the R Language Objects and Functions for Data Analysis" (2016).

[26] The R Programming Language and Software

[27] Sharma, Akshat. "Understanding Decision Tree Algorithm by using R Programming Language." (2016).

[28] http://www.rproject.org

[29] https://www.r-bloggers.com/ieee-spectrum-2017-top-programming-languages/