International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 10, No.1 (Jan-2021)

http://dx.doi.org/10.12785/ijcds/100107

Student Classification Based on Cognitive Abilities and Predicting Learning Performances Using Machine Learning Models

T. PanduRanga Vital[1, *], K. Sangeeta[2]and Kalyana Kiran Kumar[3]

[1]Associate Professor, Dept. of CSE, Aditya Institute of Technology and Management, Tekkali-532 201, Srikakulam (Dt.), A.P., India,
[2] Sr.Asst. Professor, Dept. of CSE, Aditya Institute of Technology and Management, Tekkali-532 201, Srikakulam (Dt.), A.P., India,
[3]Professor, Dept. of EEE, Aditya Institute of Technology and Management, Tekkali-532 201, Srikakulam (Dt.), A.P., India,

Received 9 Mar.2020, Revised 30 Jul. 2020, Accepted 13 Nov. 2020, Published 1 Jan. 2021

Abstract: Education is the vital parameter of the country for development in divergent areas like cultivation, economic, political, health, and so on. Any educational Institute's (universities, colleges, schools) main goal is to increase the student's learning capabilities and their skills for their full contribution towards society. ~~In~~ These days, "student's learning process and skill development" research topic requires much-needed attention for the betterment of society. The student's performance depends on his/her learning ability and is influenced by many factors. In this paper, we analyze the different categories of student's leanings that are very fast, fast, moderate, and slow. For this, we conducted the training and tests and use the features likeability, knowledge level, reasoning, and core subject abilities for the 313 engineering students in AITAM, Tekkali, affiliated to JNTUK, India of 2017 to 2019. We gathered information about the personal, academic, cognitive level, and demographic data of students. In this experiment, we are conducting statistical analysis as well as classification of students into 4 types of learners and applying the different Machine Learning (ML) techniques, and choose the best ML algorithm for predicting students learning rates. This leads to conducting remedial classes with new teaching methods for moderate and slow learning students. The proposed paper accommodates the individual differences of the learners in terms of knowledge level, learning preferences, cognitive abilities, etc. For this, we apply 5 ML algorithms that are Naive Bayes, Classification Trees (CTs), k-NN, C4.5, and SVM. As per ML analysis, the k-Nearest Neighborhood (k-NN) algorithm is more efficient than other algorithms where the accuracy and prediction values are nearer to 100%.

Keywords: Education, Machine Learning, Student's Learning, student's performance

1. Introduction

Education is the backbone of any country or society [16]. It is one of the measurements of the development of the country socially, economically, and politically. This paper uses real-time student data of the computer science engineering department, Aditya Institute of Technology and Management, Tekkali in Srikakulam district. The study involves experiments to understand the influence of cognitive attributes on academic performance. Students are classified into very-fast learners, fast learners, average learners, and slow learners using classification algorithms and finding out the best prediction model. The proposed paper accommodates the individual differences of the learners in terms of knowledge level, learning preferences, cognitive abilities, etc.

Prediction of student's performance is a challenging task as it depends on many factors such as grades, class performance, demographic data, and emotional features. The teachers need to forecast the future performance of a student based on his past performances, identifying weak students at an early stage so that additional material and special attention can be facilitated to avoid the risk of failure.

Further analysis, we compare to other works in deeply. Section 2 gives other research works descriptions in detail relevant to student performance analysis with machine

E-mail: vital2927@gmail.com, ksangeeta.10@gmail.com, kirankalyana1@gmail.com

learning algorithms with different data sets with various benchmarks MLs. Section 3 provides the proposed model and materials that different ML algorithms are analyzed. Section 4 and section 5 provides detailed comparative experimental result analysis and conclusion of the work and future work proposals.

2. Literature Review

In this section, we reviewed 120 papers from reputed journals like IEEE, Elsevier, and Springer, and so on. Some of the papers are presented that is much related to this experiment. Fok et al. [1] applied Deep Learning (DL) analytic engine to evaluates the student's performance. The study involves the analysis of the influence of academic performance and their extra-curricular activities like services, arts, and their conduct. The experiment achieved accuracy ranged from 80% to 91%. A total of 2000 sized dataset is split into 75% training data and 25% test data. The Tensor flow deep learning model is optimally configured to achieve the highest prediction accuracy. Ma et al. [2] considered the dependency among student's attributes to initialize coefficients of machine learning algorithms using initialization coefficients rules. This helps in faster convergence of algorithms. In this, the grid search algorithm is tuned to DT and SVM algorithms and gets the optimized model.

Sekeroglu et al. [3] applied Support Vector Regression (SVR), Backpropagation, LSTM for classification and prediction of student's performances, and also Gradient Boosting Classifier is imposed in this classification. For this, two different datasets SPD [14] and SAPD [15] are used, one for prediction and another one for classification. Cortez et al. [4] used a real-time student dataset comprising of students of leading academic institutions in India for prediction of student performance. They used student's features like CGPA, Lab performance, etc., for classification into four groups such as poor, average, good, and excellent. The conventional decision tree has improved functionality through association functions and normalized factors.

Yu et al. [7] applied sentiment analysis on text- based self- evaluated comments given by students for the prediction of student's performance. Experimental outcomes show that sentiment information from these remarks provides accuracy in the prediction. They regarded structured data, such as homework completion, attendance, and exam grades for their experiment. Sorour et al. [8] also applied text mining techniques for predicting student's performance. They used comments given by students and applied K-means and LSA (latent semantic analysis) methods for prediction. Overlap and similarity measuring methods are used along with LSA and k-means for improvement. The accuracy observed was 66.4% for k-means and 78.5% for the overlap method relatively. Aziz et al. [9] researched educational databases utilizing DM techniques to identify patterns. They collected real-time data of I- B.Tech. students in CSE and conceived features related to their academic records, family history, and demographics. They applied ML techniques like rule-based, naïve Bayes, and DTs for the prediction of student's performance. The rule-Based classification technique achieved high accuracy compared to the other two models with an accuracy value of 71.3%. Zhang et al. [10] focused on identifying students at risk so that early measures can be taken to increase student retention. They applied DM and NLP algorithms to observe a student's academic performance.

Mohsin et al. [11] applied data mining algorithms on student's programming performance datasets. The training dataset consists of the concert profile of Utara Malaysia University undergraduates from 4 distinctive programs that are bachelor in IT, Multimedia, Decision Science, and Education in IT of the year 2004/2005. There were 419 records with 70 attributes on which the Apriori association rule mining algorithm was applied. The author applied DM techniques on a student dataset of Bulgarian University for predicting student performance. Huang et al. [13] studied the comparative analysis of regression in multiple linear, the MLP network, RBF network, and the SVM algorithms to discover the best model for the forecast of academic performance of students. The student's attributes considered are CGPA, mid-test marks, and so on. The outturn of the models is the undergraduate's scores on the final exam. Parack et al. [14] proposed ML techniques for student grouping and profiling. They applied the Apriori algorithm for student profiling and finding co-relations among a set of items. K-means is employed for clustering students. Student profiling is completed utilizing the performance of academic records, the marks secured in term exams and mid exams, some of the papers are described in table I in detail.

TABLE I. CONTRIBUTION OF AUTHORS IN DIFFERENT PAPERS

Ref. No.	Author	Year	Techniques used	Highlighted contributions
[12]	Kabakchieva et al.,	2013	DM Techniques	DM research project for university management intended to reveal the most elevated potential of DM applications.
[5]	Hussain et al.,	2018	Bayes Network, , PART, RF, J48	Students' Academic Performance using ML methods
[6]	Sivakumar et al.,	2018	Supervised classifiers, NN, SVM, K-NN, NB DT and Improved DT	Study on student's academic performance dataset with different ML methods and concludes Improved DT is the best.
[19]	Z. Raihana et al.,	2018	SVM	The study shows the importance of quality of life on academic performance.
[20]	Fahad Razaque et al.,	2018	Naïve Bayes algorithm	The study is on performance in academics with ML techniques. It helps students for improving their performance.
[21]	Parneet Kaur et al.,	2015	Data mining algorithms	The study focuses on identifying slow learners using classification algorithms
[22]	Hasan et al.,	2018	Random Forest Tree algorithm	The study focuses in improving student's performance by early prediction.
[23]	Thomas et al., .	2017	Descriptive and correlation analysis	The study shows the significance of emotional intelligence, coping, and cognitive test anxiety on academic performance.
[24]	Hamaideh et al.,	2014	Descriptive and correlation analysis	This study shows student's cognitive, Psychological and personal aspect for academic achievement with using ML predication algorithms
[25]	Ying Lin et al.,	2017	Descriptive and correlation analysis	The study shows the importance of mental toughness on academic performance
[26]	Amirah et al.,	2015	DM techniques	Student's performance prediction
[27]	Tjioe Marvin Christian et al.,	2014	NB, Tree classification technique	In this study the authors analyzed personal education, admission, and academic data of students and provide the predications on these features.
[28]	Mayilvaganan et al.,	2014	DM algorithms	The study focuses on comparing classification algorithms used to predict student's performance based on semester exams.

García et al., [15] considered socio-demographic and academic performance for predicting the performance of 1st-semester Engineering students. The students were classified into three categories: low, middle, and high. Low means students who passed none or up to two courses, middle refers to three to four courses passed students and high refers to passed in all. They applied the Naïve Bayes classifier and the Rapid miner software that lead to 60% accuracy. The authors collected data from three colleges of Assam, India which consists of socio-economic, demographic as well as academic information of three hundred students with twenty-four attributes. They applied ML algorithms like Bayes Network, PART, RF, and SVM. The attributes that influence the most are considered using the tool. Although, the Apriori algorithm was implemented mining with association rules for all attributes. Ogunde et al. [17] analyzed the impact of university entrance examinations on student's graduation grades. They applied the ID3 decision tree algorithm for the prediction of final grades. The IF-THEN rules are framed out to represent knowledge extracted from decision trees. Hamoud et al., [18] studied the performance of Portuguese students. The applied decision tree algorithms and compared results of DTs, Hoeffding (VFDT), C4.5, and RP trees. The results showed that the J48 algorithm outperforms the other classifiers and gives the accurate prediction of students who can complete higher education courses successfully.

3. PROPOSED MODEL AND MATERIALS

In this research, we conduct and analyze the different categories of student learning levels. As per data, we classify the student's learning in 4 ways that are slow, moderate, or average, fast, and very fast learners. For this, we conduct the pieces of training and tests for the CSE students of AITAM, Tekkali, India, and gather the information about student's personal, academic, class performance, learning preferences and so on, from 2017 to 2019. The figure shows the proposed model for predicting student learning rates with the respective data. In the cognitive level, a set of questionnaires that observes cognitive abilities such as analytical thinking, error identification, and misconceptions, decision making, knowledge level, etc of a student is prepared to collect the raw data. The collected raw data and information regarding their academic performance, gender, demographic data, their learning preferences, and class performance are added to the data table and converted into *.CSV format for preprocessing. The data preprocessing is a crucial step practiced to improve the accuracy of the algorithms. The information needs to be transformed to the form used by specific algorithms. The preprocessed data is input to various machine learning algorithms and statistical analysis. This study compares the results of experiments carried out using Classification Tree, C4.5, SVM, K-NN, Naïve Bayes algorithms. The results are compared using parameters precision, recall, F1-score, and accuracy. Performance analysis is displayed using

ROC curves. The outcomes of the statistical analysis and ML model reports are communicated to the analysts for further planning and action.

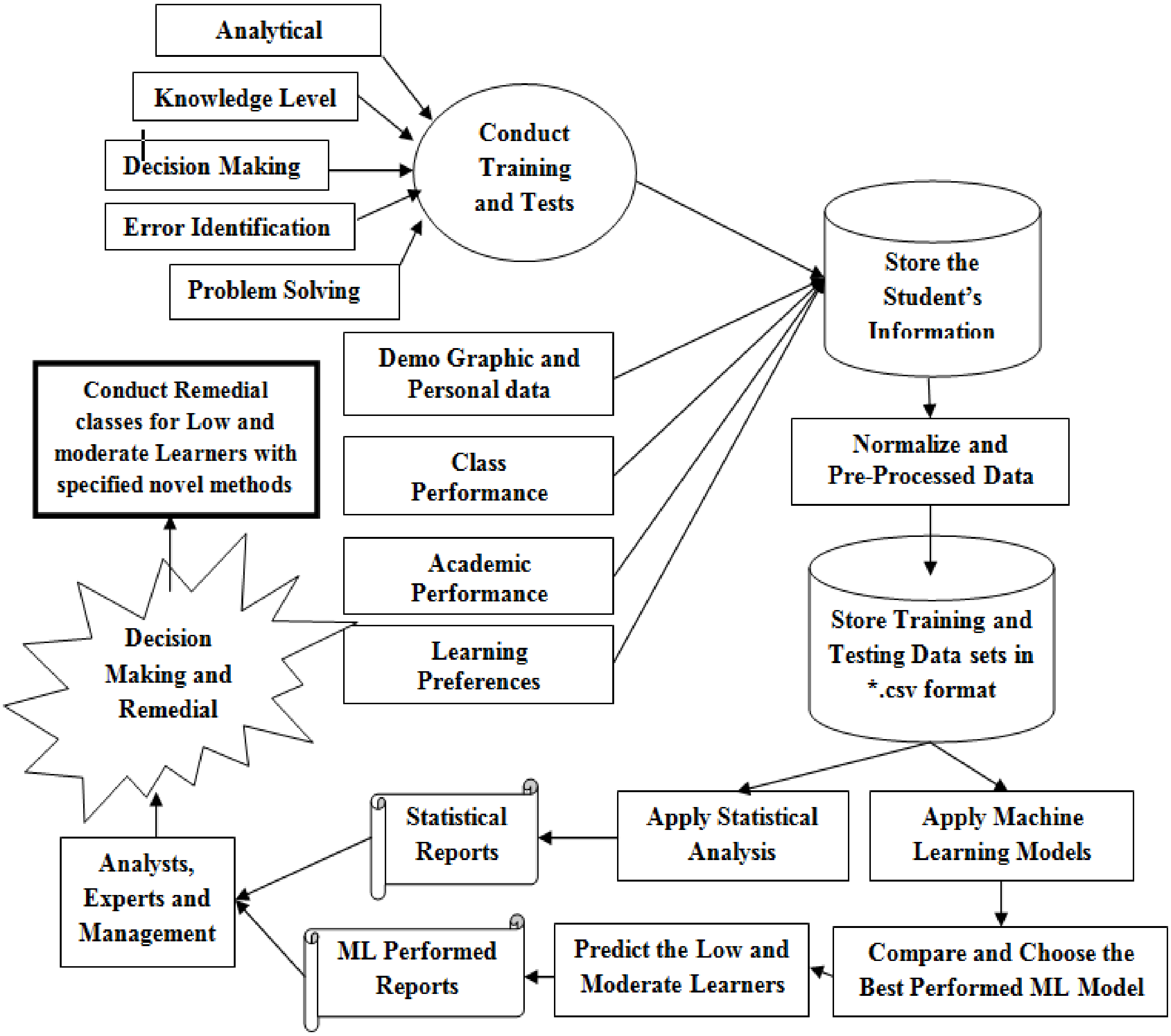
Analytical
Knowledge Level
Decision Making
Error Identification
Problem Solving
Conduct Training and Tests
Store the Student's Information
Demo Graphic and Personal data
Class Performance
Academic Performance
Learning Preferences
Conduct Remedial classes for Low and moderate Learners with specified novel methods
Normalize and Pre-Processed Data
Store Training and Testing Data sets in *.csv format
Decision Making and Remedial
Statistical Reports
Apply Statistical Analysis
Apply Machine Learning Models
Analysts, Experts and Management
ML Performed Reports
Predict the Low and Moderate Learners
Compare and Choose the Best Performed ML Model

Figure 1. Proposal Model for the Predicting Learning Levels of Students

Dataset collection and description:

Dataset collection and description: The Data is collected from AITAM College, Tekkali, A.P., India. For this experiment, 313 engineering students are involved that 143 male students and 170 female students. The total dataset learning capabilities of student class categories (fast, very fast, moderate, and slow). Table II shows the Dataset Attributes description in detail. Every attribute is described with data type nominal or discrete values. Mainly Sex, Area, learning preferences feature values are Videos, PDF, or PPT. the remains are in discrete values that are 1 to 5 or 1 to 10 values.

TABLE II . DATASET ATTRIBUTES DESCRIPTION

Features (Attributes)	Data type	Description
Sex	Nominal	Male or Female
Area	Nominal	Rural or Urban
Learning Preferences	Nominal	Videos or PDF or PPT
Class Performance	Discrete	1 to 5 values
CGPA	Discrete	1 to 10 values
Analytical Thinking	Discrete	1 to 10 values
Knowledge Level	Discrete	1 to 10 values
Problem Solving Skills	Discrete	1 to 10 values
Decision Making	Discrete	1 to 10 values
Errors Identification	Discrete	1 to 10 values
Class Attribute	Nominal	Fast, Very Fast, Moderate and Slow

Naïve Bayes Classification:

It expects that the presence of an unambiguous aspect of a class is autonomous of every other aspect. As per Bayes theorem, the contingent probability is given by the Equations (1) and (2). It is the most successful algorithm for many applications such as text document classification, spam filtering, Recommender system, etc.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{1}$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2}$$

Support Vector Machine (SVM):

Another incredible supervised ML model is SVM that can be used for both regression and classification issues. The numbers of characteristics 'n' are spoken to on the n-dimensional space with each component depicted by the estimation of a specific coordinate. An information component comprising of n characteristics is plotted on this n-dimensional space. The point is to find a hyper plane that classifies and increases the edge in an n-dimensional space.

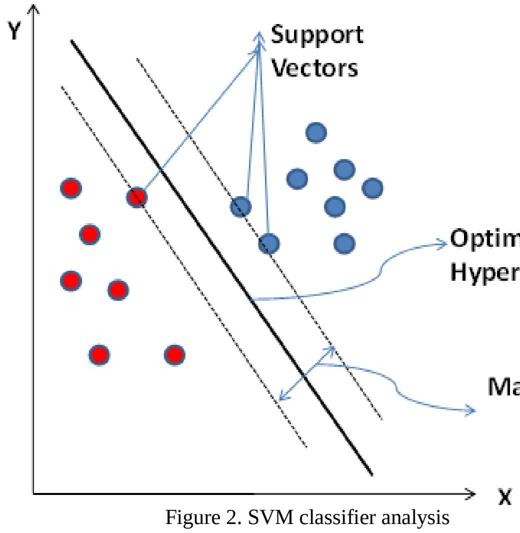

Figure 2. SVM classifier analysis

K-Nearest Neighbors' (k-NN) Classification:

The k-NN is a non-parametric supervised algorithm method suitable for both classification and regression. It considers the k closest data points in the training examples [12]. The output differs based on the fact that K-NN is used for classification or regression. The output predicts the class to which a data point belongs based on how closely it matches with the k nearest neighbors. This is one of the instance-based learning or lazy learning algorithm, because the function considers the local data points and all computation is deferred until classification. This algorithm uses a distance function to calculate the close approximate with the K Nearest Neighbors. For continuous variables, Euclidean, Manhattan, and Minkowski distance measures are used and hamming distance for categorical variables shown in equations (3-5).

$$Euclidean\ Distance\ = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \tag{3}$$

$$Mahattan\ Distance\ = \sum_{i=1}^{k}|x_i - y_i| \tag{4}$$

$$Minkowski\ Distance\ = \left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q} \tag{5}$$

Confusion Matrix and Performance Parameters:

In this, we represent the 4-class problem that are Fast, Very-Fast, Average and Slow. The table III shows the confusion matrix for the student leaning data set with 4 class problem. The accuracy is calculated by the diagonal of the confusion matrix. The confusion matrix is constructed with using actual or true values and Predicted values.

TABLE III. CONFUSION MATRIX

Classifier		Actual or True Values				
	Class	Fast (F)	Very-Fast (V)	Average (A)	Slow (S)	Total
Predicted Values	Fast (F)	F-F	F-V	F-A	F-S	T5
	Very-Fast(V)	V-F	V-V	V-A	V-S	T6
	Average (A)	A-F	A-V	A-A	A-S	T7
	Slow (S)	S-F	S-V	S-A	S-S	T8
	Total	T1	T2	T3	T4	T

We calculated the performance parameters like TPR-True Positive Rate-Recall-Sensitivity, Probability of Detection, Power, FNR-False Negative Rate, Miss Rate, FPR-False Positive Rate, Fall Out, Probability of False Alarm, SPC-Specificity, Selectivity, True Negative Rate(TNR), PPV- Positive Predictive Value, Precision, FOR-False Omission Rate, LR+-Positive Likelihood Ratio, LR—Negative Likelihood Ratio, ACC-Accuracy, FDR-False Discovery Rate, NPV-Negative Predictive Value, DOR-Diagnostic Odds Ratio, F1Score 4 to 17 respectively.

$$TPR = \frac{\sum True\ Positive}{\sum Condition\ Positive} \tag{6}$$

$$FNR = \frac{\sum False\ Nagative}{\sum Condition\ Positive} \tag{7}$$

$$FPR = \frac{\sum False\ Positive}{\sum Condition\ Negative} \tag{8}$$

$$SPC\ or\ TNR = \frac{\sum True\ Nagative}{\sum Condition\ Negative} \quad (9)$$

$$Prevalence\ = \frac{\sum Condition\ Positive}{\sum Total\ Population} \quad (10)$$

$$PRC = \frac{\sum TP}{\sum Predicted\ Condition\ Positive} \quad (11)$$

$$FOR\ = \frac{\sum False\ Negative}{\sum Predicted\ Condition\ Negative} \quad (12)$$

$$Accuracy(ACC)\ = \frac{\sum TP + \sum TN}{\sum TotalPopulation} \quad (13)$$

$$FDR\ = \frac{\sum False\ Positive}{\sum Predicted\ Condition\ Positive} \quad (14)$$

$$NPV\ = \frac{\sum True\ Negative}{\sum Predicted\ Condition\ Negative} \quad (15)$$

$$DOR = \frac{LR+}{LR-} \quad (16)$$

$$F_1 score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (17)$$

4. Results And Discussions

In this, we carried out statistical analysis of the total data set as well as observing ML algorithms performances of this dataset. In statistical analysis, the data set is observed concerning class category and attributes. The statistical analysis results are very useful to the analysts, managers, and teachers for further decision. The ML algorithms are very useful for predicting the student's category concerning the target class (fast, very fast, average (moderate), slow) learners.

4.1 Statistical Analysis

Table IV shows the Statistical values of all attributes of the Student's Learning data set that are Sex, Urban and Rural, Learning Preferences, Class Performance, CGPA, Analytical Thinking, Knowledge Level, Problem Solving Skills, Decision Making, and Errors Identification. In the experiment, we used a total of 313 individuals consisting of 143 male students and 170 female students. The total dataset learning capabilities of student class categories are fast, very fast, moderate and slow that the values are 115 (36.7%), 21 (6.7%), 127 (40.6%) and 50 (16.0%) respectively. Comparatively, female students are very fast learners than male pupils. The very fast learning students are 21, among which 61.9% are female students and 38.1% of pupils are male as well 27.8% are male, and 72.2% are female students in fast learners out of 115 fast learners.

TABLE IV . ATTRIBUTE STATISTICAL ANALYSIS OF STUDENT DATA SET

Attributes	Category	Very fast	fast	slow	Average	Total
Sex	Male	8 (38.1%)	32 (27.8%)	33 (66.0%)	70 (55.1%)	143 (45.2%)
	Female	13 (61.9%)	83 (72.2%)	17 (34.0%)	57 (44.9%)	170 (54.3%)
Urban Rural	Urban	13 (61.9%)	70 (60.9%)	34 (68.0%)	67 (52.8%)	184 (58.8%)
	Rural	8 (38.1%)	45 (39.1%)	16 (32.0%)	60 (47.2%)	129 (41.2%)
Learning Preferences	Videos	13(61.9%)	53 (46.1%)	25 (50.0%)	67 (52.8%)	158 (50.5)
	PDF	8(38.1%)	43 (37.4%)	11 (22.0%)	42 (33.1%)	104 (33.2%)
	PPT	0(0.0%)	19 (16.5%)	14 (28.0%)	18 (14.2%)	51 (16.3%)
Class Performance	Min	1	1	1	1	1
	Median	2	2	3	3	2
	Mean	1.90±0.97	2.26±0.92	2.88±1.07	2.69±1.08	2.51±1.05
	Max	4	5	5	5	5
CGPA	Min	7	6	5	6	5
	Median	8.3	7.9	6.52	7.45	7.56
	Mean	8.25±0.58	7.89±0.67	6.56±0.72	7.42±0.68	7.51±0.83
	Max	9.5	9.5	8.54	9.17	9.5
Analytical Thinking	Min	4	3	1	1	1
	Median	8	7	4	5	6
	Mean	7.62±1.36	6.33±1.69	3.96±1.71	5.06±1.67	5.52±1.94
	Max	10	9	7	8	10
Knowledge Level	Min	4	2	0	0	0
	Median	8	6	3	4	4
	Mean	7.14±1.70	6.02±2.05	2.96±2.01	4.24±2.03	4.88±2.37
	Max	10	10	6	10	10
Problem Solving Skills	Min	2	2	0	0	0
	Median	6	6	2	4	4
	Mean	6.48±1.84	5.06±1.65	2.36±1.82	3.65±1.61	4.15±2.02
	Max	10	8	6	6	10
Decision Making	Min	2	0	0	0	0
	Median	6	4	2	2	4
	Mean	6.29±2.33	4.49±1.91	2.16±1.83	2.94±1.92	3.61±2.24

	Max	10	8	6	10	10
Errors Identification	**Min**	0	0	0	0	0
	Median	6	4	0	2	2
	Mean	5.14±2.87	3.29±1.91	0.96±1.34	2.35±1.88	2.66±2.16
	Max	10	8	4	8	10

Figure 3 shows the ‘Analytical Thinking’ for slow, average, fast, and very fast, and Total Student learning’s. The total data set Analytical Thinking level mean value is 5.52±1.94; this is the higher value than slow learners and average or moderate learners and that the mean values are 3.96±1.71 and 5.06±1.67 respectively. The very fast and fast learner's mean values are higher than the total data set mean value that the mean values are 7.14±1.70 and 6.02±2.05 respectively. The other statistical values are described in table 3 as well as in figure 2 in detail. As per the analysis, Knowledge Level, Problem Solving Skills and Decision Making are also important attributes classifying the learning levels of the students. The mean values of these attributes are the mean values concerning the slow to very fast learners. The total data set mean values of Knowledge Level, Problem Solving Skills, and Decision Making are 4.88±2.37, 4.15±2.02, and 3.61±2.24 respectively.

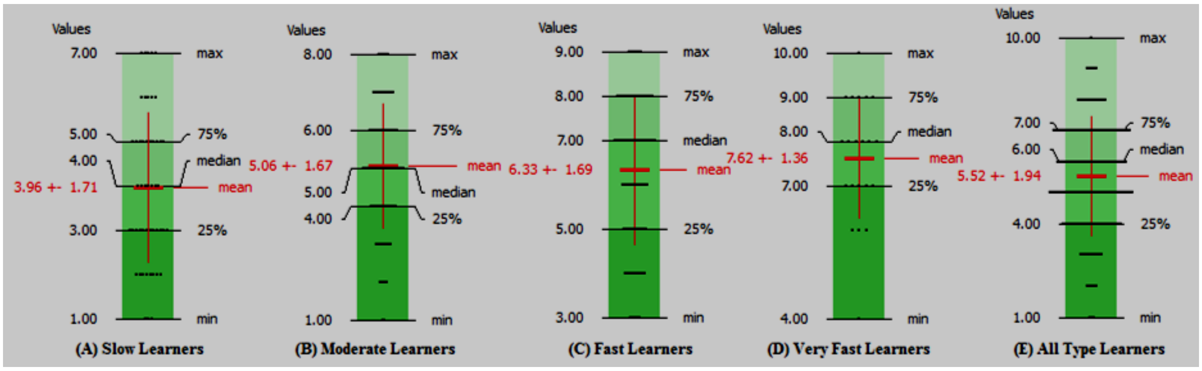

Figure 3. Analysis of Analytical Thinking for Student learners

Figure 4 shows the Analysis of some of the attributes of Total data set such as the class category, Learning preferences, gender, and area categories (Rural or Urban). Out of 313 students, fast and average or moderate learning (fast 115 (36.7%), moderate or average 127 (40.6%)) students are more than very fast and slow learners (very fast 21 (6.7), slow 50 (16.0%)) shown in figure 4(A). Most of the students choose videos as their learning preference and the count is 158 (50.5 %), the next choosing learning material is pdfs. In the category of gender 170 (54.3%) female persons and 143(45.7%) male persons are involved in the experiment and under area 184 are urban and 129 from rural.

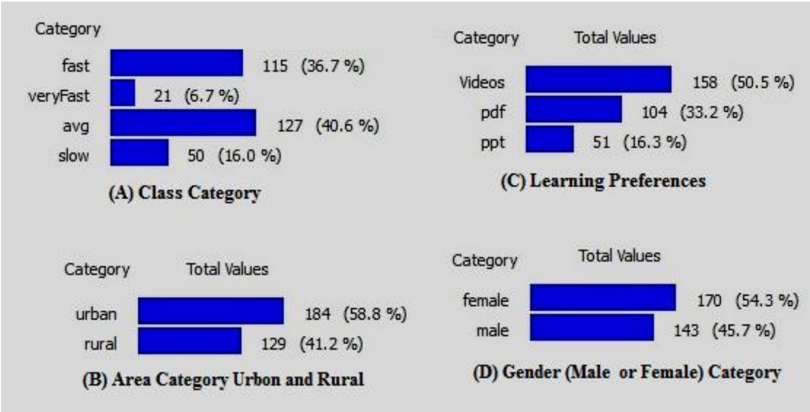

Figure 4. Analysis of Attributes from Total Data Set

4.2 Machine Learning Algorithms

After statistical analysis, we apply the ML algorithms for predicting the students learning rates that are fast, very fast, average (moderate), and slow. In this, we analyze specific ML algorithms accuracies and compare them to each other. Mainly, focused algorithms are Naive Bayes, C4.5, Classification Trees (CTs), k-NN, and SVM. For comparison, we calculate the performance parameters like CA, Sens., Spec, AUC, IS, F1, Prec., Recall, Brier, MCC, and so on with utilizing Confusion Matrix. The ROC curve also proves to be a crucial measurement for the performance of the dataset. Finally, we compare the accuracy, ROC values, and time is taken for the building of the model for each algorithm.

4.2.1. Naïve Bayes (NB) Model Analysis for Student's learning:

In the NB model, the total classified instances are 114 out of 313 within 0.06 seconds. Table V shows the NB classifier confusion Matrix. The values are assigned through predicted and true values. The class fast learning classified by this algorithm is 35 out of 115 and remaining instances (75+3+2) are classified incorrectly as well the very fast class classified 20 instances out of 21, miss classified instance goes to the fast class. The average (moderate) class instances are incorrectly classified more that are 10 in fast,60 is very fast, and 48 in slow, so correctly classified instances are only 9. The class slow learner instances are classified very accurately that 50 out of 50 classified correctly. Table V shows the detailed analysis of the classification of NB with a confusion matrix for calculating performance parameters like CA, Sens, Spec, AUC, IS, F1, Perc, Recall, Brier, and MCC.

TABLE V. NAÏVE BAYES CLASSIFIER CONFUSION MATRIX ANALYSIS

NB Classifier		**True Values**				
	Class	**fast**	**Very-Fast**	**average**	**slow**	**Total**
Predicted Values	**fast**	35	75	3	2	**115**
	Very-Fast	1	20	0	0	**21**
	average	10	60	9	48	**127**
	slow	0	0	0	50	**50**
	Total	**46**	**155**	**12**	**100**	**313**

Table VI shows the NB performance parameters analysis. The average accuracy value of all classes is 0.3642 below 0.5, so this algorithm is not used for predicting the student's learning measures. The class slow is very accurate where sensitivity or recall values are 1. The average accuracy is very low where sensitivity is 0.0709 and the MCC value is 0.14. Overall Naïve Bayes is the failure model for this data set.

TABLE VI . NAÏVE BAYES MODEL PERFORMANCE ANALYSIS

Class	**CA**	**Sens**	**Specificity**	**AUC**	**IS**	**F1**	**Precision**	**Recall**	**Brier**	**MCC**
VeryFast	0.3642	0.9524	0.5377	0.9391	0.4325	0.2273	0.129	0.9524	0.9447	0.2452
Fast	0.3642	0.3043	0.9444	0.9391	0.4325	0.4348	**0.7609**	0.3043	0.9447	0.3388
Average	0.3642	0.0709	**0.9839**	0.9391	0.4325	**0.1295**	0.75	0.0709	0.9447	**0.14**
Slow	0.3642	**1**	0.8099	0.9391	0.4325	0.6667	0.5	**1**	0.9447	0.6364
Avg	**0.3642**	**0.5819**	**0.818975**	**0.9391**	**0.4325**	**0.364575**	**0.534975**	**0.5819**	**0.9447**	**0.3401**

4.2.2. k-NN Model Analysis for Student's learning:

TABLE VII . K-NN CLASSIFIER CONFUSION MATRIX ANALYSIS

K-NN Classifier		**True Values**				
	Class	**fast**	**Very-Fast**	**average**	**slow**	**Total**
Predicted Values	**fast**	115	0	0	0	**115**
	Very-Fast	0	21	0	0	**21**
	average	0	0	127	0	**127**
	slow	0	0	0	50	**50**
	Total	**115**	**21**	**127**	**50**	**313**

In the k-Nearest Neighbor model, the total classified instances are 313 out of 313 within 0.11 seconds. Table VII shows the k-NN classifier confusion Matrix. The values are assigned through predicted and true values. All fast classes, very fast, average (moderate), and slow are classified correctly were the instances 115, 21,127, and 50 in true positive in respect order. So, ~~CNN~~ K-NN is a very accurate and predictable model for the student learning dataset.

Table VIII shows the k-NN performance parameters analysis. The average accuracy value of all classes is 1, which means the accuracy is 100%, so this algorithm is used for predicting the student's learning measures accurately. All the classes are very accurate where sensitivity or recall values are 1. The average accuracy is very high where sensitivity is 1, MCC value is 1 and the brier values are very low that is 0.00047. Overall k-NN is the greatest model for this data set. Table VIII shows the detailed analysis of this algorithm.

TABLE VIII. K-NN MODEL PERFORMANCE ANALYSIS

Class	CA	Sens	Spec	AUC	IS	F1	Precision	Recall	Brier	MCC
Very-Fast	1	1	1	1	1.7112	1	1	1	0.0047	**1**
Fast	1	1	1	1	1.7112	1	1	1	0.0047	**1**
Average	1	1	1	1	1.7112	1	1	1	0.0047	**1**
Slow	1	1	1	1	1.7112	1	1	1	0.0047	**1**
Avg	**1**	**1**	**1**	**1**	**1.7112**	**1**	**1**	**1**	**0.0047**	**1**

4.2.3 Classification Trees Model Analysis for Student's learning: In the CTs model, the total correctly classified instances are 295 out of 313 within 0.15 seconds. Table IX shows the CTs classifier confusion Matrix. The values are assigned through predicted and true values. The class fast learning classified correctly 110 out of 115 remaining instances (0+5+0) are classified incorrectly as well the very fast class classified 15 instances out of 21, miss classified 6 instance goes to the fast class. Table X shows the k-NN classifier confusion Matrix. The values are assigned through predicted and true values.

TABLE IX. CTS CLASSIFIER CONFUSION MATRIX ANALYSIS

CTs Classifier		True Values				
	Class	**fast**	**Very-Fast**	**average**	**slow**	**Total**
Predicted Values	**fast**	110	0	5	0	**115**
	Very-Fast	6	15	0	0	**21**
	average	4	1	122	0	**127**
	slow	0	0	2	48	**50**
	Total	**120**	**16**	**129**	**48**	**313**

The average (moderate) class correctly classified instances are 122 and incorrectly classified instances are 4, 1, and 0 in fast, very fast, and slow respectively. Out of 50 slow learner instances classified 48 correctly and 2 miss classified instances are gone to the average. Table IX shows the detailed analysis of the classification of CTs with the confusion matrix for calculating performance parameters. Table X shows the CTs performance parameters analysis. The average accuracy value of all classes is 0.9425 above 94%, so this algorithm is good to predict the student's learning measures. The class slow is classified correctly compare to other classes with this algorithm where sensitivity or recall values are 0.96 and MCC values are 0.9761. As well as specificity and precision values are very high in slow learner class that the value is 1. The brier value is equal to all the classes that the value is 0.0905. The detailed performance analysis described in table X.

TABLE X. CTS MODEL PERFORMANCE ANALYSIS

Class	CA	Sens	Spec	AUC	IS	F1	Prec	Recall	Brier	MCC
Very-Fast	0.9425	0.9565	0.9495	0.9937	1.548	0.9362	0.9167	0.9565	0.0905	0.8984
Fast	0.9425	0.7143	0.9966	0.9937	1.548	0.8108	0.9375	0.7143	0.0905	0.8075
Average	0.9425	**0.9606**	0.9624	0.9937	1.548	0.9531	0.9457	**0.9606**	0.0905	0.9208
Slow	0.9425	0.96	**1**	0.9937	1.548	**0.9796**	**1**	0.96	0.0905	**0.9761**
Avg	**0.9425**	**0.89785**	**0.9772**	**0.9937**	**1.548**	**0.91993**	**0.949975**	**0.89785**	**0.0905**	**0.9007**

4.2.4. C4.5 Model Analysis for Student's learning:

In the C4.5 model, the total correctly classified instances are 288 out of 313 within 0.13 seconds. Table XI shows the CTs classifier confusion Matrix. The values are assigned through predicted and true values. The class fast learning classified correctly 108 out of 115, remaining instances (2+5+0) are classified incorrectly and also the very fast class classified 14 instances out of 21, miss classified 7 instance goes to the fast class. The average (moderate) class correctly classified instances are 119 and incorrectly classified instances are 5, 0, and 3 in fast, very fast, and slow respectively. Out of 50 slow learner instances classified 47 correctly and 3 miss classified instances are gone to the nearest class average. Table XI shows the detailed analysis of the classification of CTs with a confusion matrix for calculating performance parameters.

Table XII shows the C4.5's performance parameters analysis. The average accuracy value of all classes is 0.9201 above 0.9, so this algorithm is good for predicting the student's learning measures. The sensitive or Recall value is very high for slow learner class that the value is 0.94 and very low in class fast that the value is 0.6667. The F1 and MCC values are also very high for class slow learners compare to other classes and low values in Fast class. The brier value is equal to all the classes that the value is 0.131. The detailed performance analysis described in table XII.

TABLE XI. C4.5 CLASSIFIER CONFUSION MATRIX ANALYSIS

C 4.5 Classifier		True Values				
	Class	**fast**	**Very-Fast**	**average**	**slow**	**Total**
Predicted Values	**fast**	108	2	5	0	**115**
	Very-Fast	7	14	0	0	**21**
	average	5	0	119	3	**127**
	slow	0	0	3	47	**50**
	Total	**120**	**16**	**127**	**50**	**313**

TABLE XII. C4.5 MODEL PERFORMANCE ANALYSIS

Class	CA	Sens.	Spec	AUC	IS	F1	Prec.	Recall	Brier	MCC
Very Fast	0.9201	0.9391	0.9394	0.9893	1.4674	0.9191	0.9	0.9391	0.131	0.8711
Fast	0.9201	0.6667	**0.9932**	0.9893	1.4674	0.7568	0.875	0.6667	0.131	0.7495
Average	0.9201	0.937	0.957	0.9893	1.4674	0.937	0.937	0.937	0.131	0.894
Slow	0.9201	**0.94**	0.9886	0.9893	1.4674	**0.94**	**0.94**	**0.94**	0.131	**0.9286**
Avg	**0.9201**	**0.8707**	**0.96955**	**0.9893**	**1.4674**	**0.888225**	**0.913**	**0.8707**	**0.131**	**0.8608**

4.2.5 Support Vector Machine Model Analysis for Student's learning:

In the SVM model, the total correctly classified instances are 267 out of 313 within 0.29 seconds. Table XIII shows the CTs classifier confusion Matrix. The values are assigned through predicted and true values. The class fast learning classified correctly 94 out of 115, remaining instances (2+19+0) are classified incorrectly as well the very fast class classified 13 instances out of 21, miss classified instances 8 are gone to the nearest class fast. The average (moderate) class correctly classified instances are 113 and incorrectly classified instances are 9, 0, and 5 in fast, very fast, and slow respectively. Out of 50 slow learner instances, 47 are correctly classified and 3 miss classified instances are gone to the average. Table XIII shows a detailed analysis of the classification of SVM with the confusion matrix for calculating SVM performance parameters.

TABLE XIII. SVM CLASSIFIER CONFUSION MATRIX ANALYSIS

SVM Classifier		**True Values**				
	Class	**fast**	**Very-Fast**	**average**	**slow**	**Total**
Predicted Values	**fast**	94	2	19	0	**115**
	Very-Fast	8	13	0	0	**21**
	average	9	0	113	5	**127**
	slow	0	0	3	47	**50**
	Total	**111**	**15**	**135**	**52**	**313**

Table XIV shows the SVM performance parameters analysis. The average accuracy value of all classes is 0.853 above 0.8, so this algorithm is moderately used for predicting the student's learning measures. The class slow is classified correctly compare to other classes with this algorithm where sensitivity or recall values are 0.94 and MCC values are 0.9066. The specificity value is very high in a very fast class that the value is 0.9932. The brier value is equal to all the classes that the value is 0.1995. The detailed performance analysis described in the table.

TABLE XIV . SVM MODEL PERFORMANCE ANALYSIS

Class	CA	Sens	Spec	AUC	IS	F1	Prec.	Recall	Brier	MCC
Very Fast	0.853	0.619	**0.9932**	0.9761	1.2827	0.7222	0.8667	0.619	0.1995	0.717
Fast	0.853	0.8174	0.9141	0.9761	1.2827	0.8319	0.8468	0.8174	0.1995	0.7372
Average	0.853	0.8898	0.8817	0.9761	1.2827	0.8626	0.837	0.8898	0.1995	0.7649
Slow	0.853	**0.94**	0.981	0.9761	1.2827	**0.9216**	**0.9038**	**0.94**	0.1995	**0.9066**
Avg	**0.853**	**0.8166**	**0.9425**	**0.9761**	**1.2827**	**0.83458**	**0.86358**	**0.8166**	**0.1995**	**0.78143**

4.2.6 ML Models Comparative Analysis

As per comparison, all algorithms' performances are good except for the Naïve Bayes model where the accuracy is only 36%. As per the time to build the model, NBC is best. As per the analysis, the k-NN is the best model where the accuracy for predicting the target class is 100% with 0.11 seconds of model construction. The next predictors are CTs, C4.5, and SVM in order where performance accuracies 94%, 92%, and 85% respectively. Table XV shows the detailed description of every specified ML Model.

TABLE XV. ALL SPECIFIED ML MODELS PERFORMANCES COMPARATIVE ANALYSIS

ML Algorithm	Time in Seconds	CA	Sens.	Spec	AUC	F1	Prec.	Recall	Brier	MCC
NBC	**0.06**	0.3642	0.5819	0.81898	0.9391	0.364575	0.534975	0.5819	0.9447	0.3401
k-NN	0.11	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0.0047**	**1**
C4.5	0.13	0.9201	0.8707	0.96955	0.9893	0.888225	0.913	0.8707	0.131	0.8608
CTS	0.15	0.9425	0.89785	0.9772	0.9937	0.91993	0.949975	0.89785	0.0905	0.9007
SVM	0.29	0.853	0.8166	0.9425	0.9761	0.83458	0.86358	0.8166	0.1995	0.78143

The figure 5 shows the Predicted class fast by utilizing the ROC curves. In this analysis, the ROC curve constructed with specificity (FP Rate) and Sensitivity (TP Rate) measures with 0 to 1 value. In order, the k-NN predicted class fast value is one specified with the red line. The yellow line indicates the CTs model ROC that the AUC value is 0.99. The C4.5 AUC value is 0.98 indicated with the green line.

Figure 6 shows the Predicted class very fast with utilizing the ROC curves. In this analysis, the ROC curve constructed with specificity (FP Rate) and Sensitivity (TP Rate) measures with 0 to 1 value. In order, the k-NN predicted class fast value is one specified with the red line. The yellow line indicates the CTs model ROC that the AUC value is 0.99. The C4.5 AUC value is 0.98 indicated with the green line.

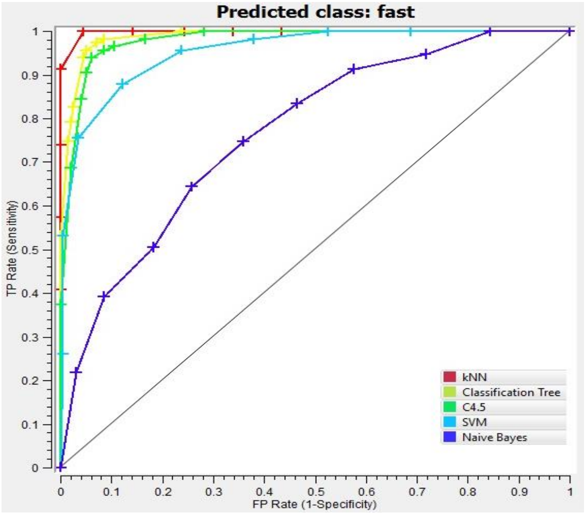

Figure 5. ROCs analysis Specified ML's for Predicted class fast

.

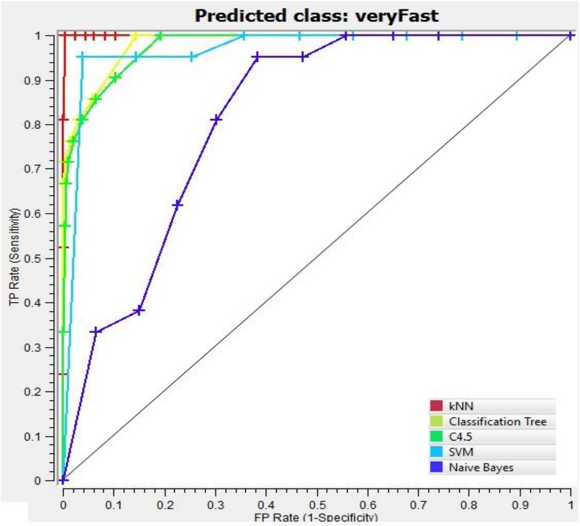

Figure 6. ROCs analysis Specified ML's for Predicted class very fast

Figure 7 shows the Predicted class moderate or average learner with utilizing the ROC curves. In this analysis, the ROC curve constructed with specificity (FP Rate) and Sensitivity (TP Rate) measures with 0 to 1 value. In order, the k-NN predicted class moderate or average value is one specified with the red line. The yellow line indicates the CTs model ROC that the AUC value is 0.99. The C4.5 AUC value is 0.98 indicated with the green line

Figure 8 shows the Predicted class, slow learner, by utilizing the ROC curves. In this analysis, the ROC curve constructed with specificity (FP Rate) and Sensitivity (TP Rate) measures with 0 to 1 values. All the MLS are nearly predicted class, slow learner, nearer to 1.

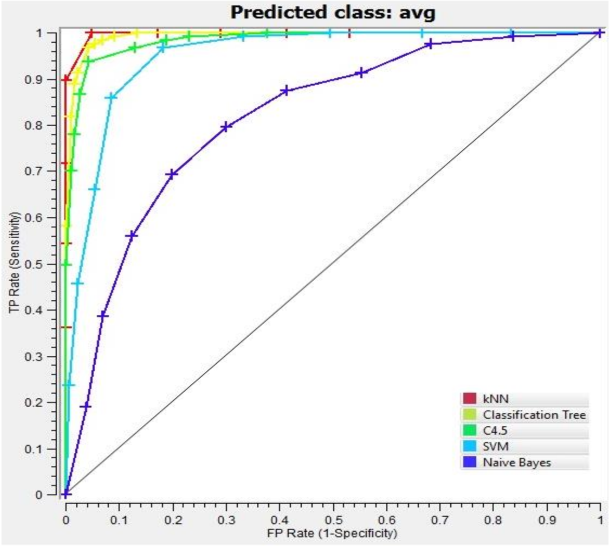

Figure 7. ROCs analysis Specified ML's for predicted class Average

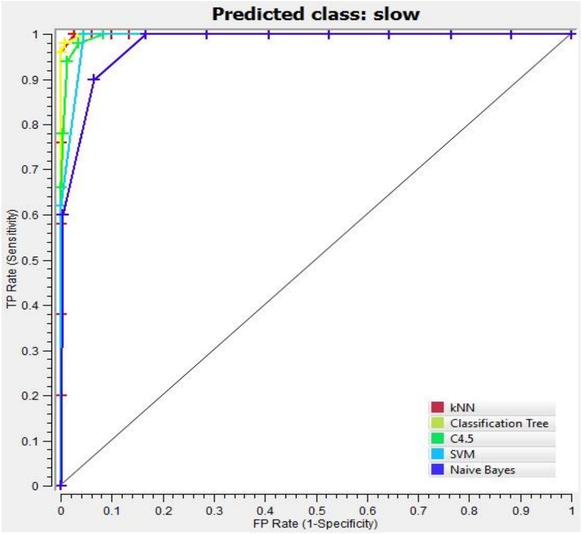

Figure 8. ROCs analysis Specified ML's for Predicted class slow

The figure 9 shows the time taken for building the ML models. In this analysis, the SVM model takes the 0.29 seconds highest time compare to other models. The least time 0.06 seconds is taken by the NBC. The k-NN, C4.5 and CTs take the times 0.11, 0.13 and 0.15 seconds respectively.

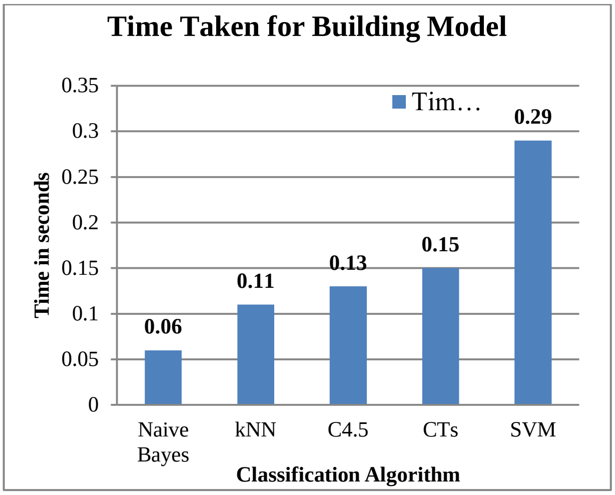

Figure 9. Time taken for Building the ML models

Figure 10 shows the comparative analysis for classification accuracy ML models. In this analysis, the k-NN model CA value is 1 that accuracy in 100%. The next accurate model is CTs with 0.9425 accuracy value. The ML algorithms C4.5 and SVM accuracy values are 0.9201 and 0.853 respectively. The algorithm NBC is not used for predicting the target class where it performs only 0.3642 (36%) only.

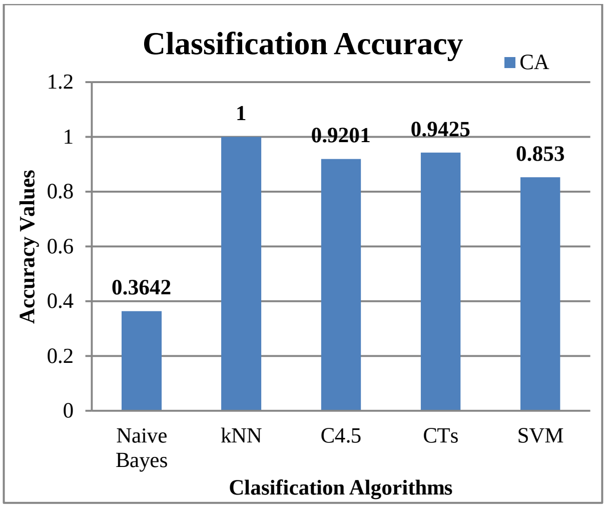

Figure 10. Comparative Analysis with respect to classification Accuracy

5. Conclusion

Education is a crucial factor for improvement for personal, economically, and socially. The student's performances and skill developments depend on their learning rates related to teaching methods, trainer capabilities, and methodologies. In this empirical educational research on engineering students, we studied nearly two years in-depth of student's learning capabilities. As per data and statistical analysis, we observed that the fewer members of the very fast and slow learners in engineering streams. The slow learners also increase their learning rate with self-learning, modern teaching methods, and remedial training sections. Data Mining plays a very crucial role in this research that we predict learners with ML models. The k-NN model is very accurate compared to other specified algorithms. Further, we will elaborate on this study with big data for all branches of engineering students and compare to science and arts degree student's learning capabilities, and predictions with neural networks and deep learning methods.

Acknowledgment

We would like to thank the Director Prof. V.V Nageswara Rao and Dr. A. Srinivas Rao Principal, AITAM College, Tekkali, HOD, and Staff of CSE Dept., for encouraging and supporting us. We are thinking to students of AITAM for cooperation to collect their data from them. The data that support the findings of this study are available in the college of AITAM, Tekkali, Andhra Pradesh, India.

References

[1] Fok, W. W., He, Y. S., Yeung, H. A., Law, K. Y., Cheung, K. H., Ai, Y. Y., & Ho, P. (2018, May). Prediction model for students' future development by deep learning and tensorflow artificial intelligence engine. In 2018 4th International Conference on Information Management (ICIM) (pp. 103-106). IEEE.

[2] Ma, X., & Zhou, Z. (2018, January). Student pass rates prediction using optimized support vector machine and decision tree. In 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC) (pp. 209-215). IEEE.

[3] Sekeroglu, B., Dimililer, K., & Tuncal, K. (2019, March). Student performance prediction and classification using machine learning algorithms. In Proceedings of the 2019 8th International Conference on Educational and Information Technology (pp. 7-11). ACM.

[4] Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance.

[5] Hussain, S., Dahan, N. A., Ba-Alwib, F. M., & Ribata, N. (2018). Educational data mining and analysis of students' academic performance using WEKA. Indonesian Journal of Electrical Engineering and Computer Science, 9(2), 447-459.

[6] Sivakumar, S., & Selvaraj, R. (2018). Predictive modeling of students performance through the enhanced decision tree. In Advances in Electronics, Communication and Computing (pp. 21-36). Springer, Singapore.

[7] Yu, L. C., Lee, C. W., Pan, H. I., Chou, C. Y., Chao, P. Y., Chen, Z. H., ... & Lai, K. R. (2018). Improving early prediction of academic failure using sentiment analysis on self-evaluated comments. Journal of Computer Assisted Learning, 34(4), 358-365.

[8] Sorour, S. E., Mine, T., Goda, K., & Hirokawa, S. (2015). A predictive model to evaluate student performance. Journal of Information Processing, 23(2), 192-201.

[9] Aziz, A. A., Ismail, N. H., & Ahmad, F. (2013). MINING STUDENTS'ACADEMIC PERFORMANCE. Journal of Theoretical & Applied Information Technology, 53(3).

[10] Zhang, Y., Oussena, S., Clark, T., & Kim, H. (2010, June). Use Data Mining to Improve Student Retention in Higher Education-A Case Study. In ICEIS (1) (pp. 190-197).

[11] Mohsin, M. F. M., Wahab, M. H. A., Zaiyadi, M. F., Norwawi, N. M., & Hibadullah, C. F. (2010, June). An investigation into influence factor of student programming grade using association rule mining. In International Journal of Advances in Information Sciences and Service Sciences. Vol.

[12] Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. Cybernetics and information technologies, 13(1), 61-72.

[13] Huang, S., & Fang, N. (2011, October). Work in progress—Prediction of students' academic performance in an introductory engineering course. In 2011 Frontiers in Education Conference (FIE) (pp. S4D-1). IEEE.

[14] Parack, S., Zahid, Z., & Merchant, F. (2012, January). Application of data mining in educational databases for predicting academic trends and patterns. In 2012 IEEE International Conference on Technology Enhanced Education (ICTEE) (pp. 1-4). IEEE.

[15] García, E. P. I., & Mora, P. M. (2011, November). Model prediction of academic performance for first year students. In 2011 10th Mexican International Conference on Artificial Intelligence (pp. 169-174). IEEE.

[16] Vital, T. P., Lakshmi, B. G., Rekha, H. S., & DhanaLakshmi, M. (2019). Student Performance Analysis with Using Statistical and Cluster Studies. In Soft Computing in Data Analytics (pp. 743-757). Springer, Singapore.

[17] Ogunde, A. O., & Ajibade, D. A. (2014). A data mining system for predicting university students' graduation grades using ID3 decision tree algorithm. Journal of Computer Science and Information Technology, 2(1), 21-46.

[18] Hamoud, A. (2016). Selection of best decision tree algorithm for prediction and classification of students' action.

[19] Raihana, Z., and Farah Nabilah, A. M., Classification of students based on quality of life and academic performance by using support vector machine. Journal of Academia UiTM Negeri Sembilan 6(1):45–52, 2018.

[20] Razaque, F., Soomro, N., Shaikh, S. A., Soomro, S., Samo, J. A.,Kumar, N., and Dharejo, H. Using naïve bayes algorithm to students'bachelor academic performances analysis. In: EngineeringTechnologies and Applied Sciences (ICETAS), 2017 4th IEEE International Conference. p 1-5, 2017.

[21] Parneet Kaur,Manpreet Singh,Gurpreet Singh Josan,Classification and prediction based data mining algorithms to predict slow learners in education sector . 3rd International Conference on Recent Trends in Computing 2015(ICRTC-2015)

[22] Hasan, R., Palaniappan, S., Raziff, A. R. A., Mahmood, S., and Sarker, K. U. Student Academic Performance Prediction by using Decision Tree Algorithm, 4th International Conference on Computer and Information Sciences. 2018.

[23] Thomas, C.L., Cassady, J.C. and Heller, M.L. (2017). "The influence of emotional intelligence, cognitive test anxiety, and coping strategies on undergraduate academic performance," Learning and Individual Differences, 55, 40 – 48.

[24] Hamaideh, S.H. and Hamdan-Mansour, A.M. (2014). "Psychological, cognitive, and personal variables that predict college achievement among health sciences students," Nurse Education Today, 34, 703 – 708.

[25] Lin, Y., Clough, P.J., Welch, J. and Papageorgiou, K.A. (2017). "Individual differences in mental toughness associate with academic performance and income," Personality and Individual Differences, 113, 178 – 183.

[26] A Review on Predicting Student's Performance using Data Mining Techniques. Amirah Mohamed Shahiria,*, Wahidah Husaina, Nur'aini Abdul Rashida,a School of Computer Sciences Universiti Sains Malayisa 11800 USM, Penang, Malaysia

[27] T. M. Christian, M. Ayub, Exploration of classification using nbtree for predicting students' performance, in: Data and Software Engineering (ICODSE), 2014 International Conference on, IEEE, 2014, pp. 1–6.

[28] Mayilvaganan .M., D. Kalpanadevi, Comparison of classification techniques for predicting the performance of students academic environment,in: Communication and Network Technologies (ICCNT), 2014 International Conference on, IEEE, 2014, pp. 113–118.

Dr. PanduRanga Vital Terlapu pursed Bachelor of Science in Computer Science from Andhra University of A.P, India in 1995 and Master of computer Application from Andhra University in year 1998. He completed his M. Tech in Computer Science and Engineering from Acharya Nagarjuna University of A.P, India and he completed his Ph.D. in Computer Science and Engineering from GITAM University of A.P, India. He has 19 years of teaching and 13 years of research experience. He is currently working as Associate Professor in Department of Computer Science and Engineering, Aditya Institute of Technology and Management (AITAM), India. He is a member of ACM, Life Time Membership from International Computer Science and Engineering Society (ICSES), USA and Life Time Membership from Indian Society for Technical Education (ISTE), New Delhi, India. He has published more than 30 research papers in reputed international journals including SCOPUS indexed and a conference including Sringer, Elsevier and it's also available online. He is reviewer of reputed journals like Springer, Elsevier and IEEE. His main research work focuses on Machine Learning, Deep Learning and Data Mining, Data and Big Data Analytics, IoT and Computational Intelligence, Voice Analysis and Voice Processing, Bioinformatics.

Smt. K. Sangeeta, Sr. Assistant Professor, CSE Department, AITAM, Tekkali. She did her MTech. in Computer science and Engineering from IIT Madras. Her research interest includes machine Learning, Artificial Intelligence, Deep Learning and Cryptography. She has 13 yrs of experience in teaching field and one year in industrial area. She has 6 publications in International journals,2 in national journals and presented 2 papers in national and 3 papers in international conferences

Dr. Kalyana Kiran Kumar was awareded Ph.d and ME degrees 2015 and 2004 respectively from Andhra University, Visakhapatnam, India. Currently he is working as a Professor in the Department of Electrical and Electronics Engineering, AITAM College of Engineering, TEKKALI, Andhra Pradesh. He has vast teaching experience of 16 years. He has published more than 20 research papers in reputed international journals .He is a member of IETE, ISTE and IAENG (HK). His Research interests include large scale systems, designing of controllers, Interval systems, Fractional systems and nonlinear systems.