ANALYSIS ON STUDENT PERFORMANCE
DATASET USING MINING TECHNIQUES (SVM)

ISSN(Online): 2581-5296

ANALYSIS ON STUDENT PERFORMANCE DATASET USING MINING TECHNIQUES (SVM)

Harshita Khangarot , Kavita Choudhary

ABSTRACT

Education is a key factor for achieving a long-term economic progress. For analyzing factors which are affecting the result of students, data mining has uncovered a lot of knowledge related to it and making decisions for their future. The purpose of this research is to analyze student performance's data and verify whether regression, normalization methods and models would work effectively. The data is taken from uci repository. The performance of students is evaluated using four distinct classifiers named as J48, naive bayes, SVM and regression. These algorithms are used to predict the final grade of students.

KEYWORDS: Higher

INTRODUCTION

Knowledge is a way to get success, achieved by getting the basic education, which in return gives economic benefit to our society as well as person, is benefitted in many ways giving positive effects. Over the decades, researchers are trying to predict how the student performance can be predicted using the past data. The dependency between various features is needed to be determined which affects the performance of student. Many reformers had claimed that performance of student can be improved by improving the various parameters related to school [1]. Predicting the performance of student can help instructors in taking various measures to improve their performance.

LITERATURE SURVEY

Table 1: Literature survey on different student dataset, methods used and parameters considered with appropriate accuracy and other details.

Authors	Techniques	Parameters considered	Results	Future scope or limitations
M.L. Anderson (2017) [9]	Regression	School Lunch Quality and Academic Performance.It depends on quality of food not calories.	0.031 standard deviation increase in test scores	Finding the relation between other parameters that may affect the performance.
Gina A. N. Chowa (2013) [10]	Structural equation modeling (SEM)	Parental involvement on performance	Home-based parental involvement positive, while school-based	parental involvement has a negative association.
			Parental involvement benefits students' academic	performance, but further investigation using longitudinal research in other developing countries is necessary.

Department of Computer Science, Institute of Engineering and Technology, JKLU Jaipur

Authors	Techniques	Parameters considered	Results	Future scope or limitations
S. P. Singh (2016) [11]	Mean, standard deviation and regression analysis	Learning facilities, communication skills and proper guidance from parents on the students' performance.	Positive impact,learning facilities (â=.514), communication skills (â=.303) and proper guidance from parents (â=.208)	For further research, family income, parent's education and educator can be analyzed.
E.M. Ganyaupfu (2013) [12]	Linear Model based univariate ANOVA	To investigate the differential effectiveness of teaching methods	The Tukey post hoc tests results indicated that student performance assessment scores of the teacher-centered approach differed significantly from other - (F(2, 106) statistic (= 10.12) at 0.05)	Teachers should also increase their knowledge of various instructional strategies in order to keep students engaged and motivatedthroughout the learning process.
G. A. Fayombo (2012) [13]	Regression, correlations	Investigates emotional intelligence	The emotional intelligence components also jointly contributed 48% of the variance in academic achievement. Attending	Further studies are required to further investigate the contributions of the variables that did not contribute significantly to the variance in academic achievement in this study.
I.Fialho (2010) [14]	CART (Classification and Regression Trees)	Quality indicators is needed in order to understand how students recognize quality in teaching. Teacher commitment, teaching and evaluation methodologies.	TeachingMethodologies (á = .839), Evaluation methodologies domain (á = .640)Teacher Commitment domain (á = .640).	
M.K.Dhaqane (2016) [15]	Correlation statistics (Pearson correlation)	Examines the role of satisfaction of students'	The Grand mean implies (2.70) and standard deviation of (1.07) on four-point likert scale.Positive response	The study recommends employing larger sample size to get various findings.
E.Tomul(2013) [16]	Regression	Effects of familial variables (education of the parents and family income) in various regions.	Familial variables on mathematics achievement- highest effect in Aegean Region (r2= ,257)In average, 16.6% of variation in	

Table 2: Literature survey on student performance dataset(uci repository), methods used and parameters considered with appropriate accuracy and other details.

Authors	Techniques	Results	Future scope or limitation
E.Osmanbegovic (2012) [8]	Rules-based, Trees-based, Functions-based, Bayes-based	Random Forestand J48 accuracy higher than 71%.	The degree of understanding and recognition of the most important attributeswith the aim of improvement in future research
P. Kavipriya (2016)[17]	Data Mining Methods	Support Vector Machine – 95%Naive Bayes – 96%	Various classifications and clustering applications to enhance the prediction speed and accuracy in the field of education.
D. Kabakchieva (2012) [18]	a rule learner,a decision tree , a neural network,Nearest Neighbour classifier	Neural Network algorithm – 73.59%.Kappa Statistic-0.473ROC Area- 0.82	The presented results will be compared to previous results, achieved for the same dataset but for a different format of the predicted target variable. Achieve better prediction. Recommendations will also be provided

A variety of mathematical techniques, such as multivariate linear regression[1], neural networks, Bayesian networks [2], decision trees, and genetic algorithm, have been employed to develop various models to predict student academic performance.

H. Shaoboet. al[1] hadstudied the multivariate linear regression to understand the performance of students from engineering domain. Multiple criteria are employed to evaluate and validate the predictive models, including R-square, shrinkage, the average prediction accuracy, and the percentage of good predictions.

I. Ianus[2] had improved prediction of freshman GPA based on college admission data to better inform the decision as to who to admit to Carnegie Mellon. Classical and Bayesian approaches are used by the author to better understand the previous criterion of acceptance and to investigate the significance of a difference between students who were admitted and enrolled and the students who were admitted and did not come to CMU. A Bayesian predictive approach was used to identify the cutoff based on admission data for the predictive probability that a student's first semester GP A is greater than 2.0.

D. Nguyenet. al [3] had predicted age from text using regression algorithm. The author had also employed a technique from domain adaptation that allows us to train a joint model involving all three corpora (blogs, telephone conversations, and online forum posts) together as well as separately and analyze differences in predictive features across joint and corpus- specific aspects of the model. Using a linear regression model based on shallow text features, he had obtained correlations up to 0.74 and mean absolute errors between 4.1 and 6.8 years.

S. Sunthornjittanon [4] had analyzed the ABC Company's data and verified whether the regression analysis methods and models would work effectively. However, final model is selected by him using Stepwise Regression Methods.

G. Narasinga Raoet. al [5] had presented a study on the academic performance of the students of a reputed college. For this purpose, he had applied multiple linear regression models on data and then evaluated the dependent variables for two given predictor variables.

W. Pyke et. al [6] had predicted the retention of students from masters as well as doctoral using logistic regression analysis.He had determined various factors that can help in increasing the chances of graduating the students with good marks.

After considering the literature review, there is a need to determine the relation between various demographic, social/emotional and school related variable that are expected to affect the performance of student. For this purpose, we had taken the dataset of Students' Performance from UCI Machine Learning Repository. The data set consists of 32 attributes consisting of the details of students from math's stream. Here, 20 point grading scale is used considering 20 as a perfect score and students are evaluated in three stages where last is the final score.

1. Methodology

The methodology for the processing of student performance dataset is discussed in below sections.It includes giving the description of the dataset that is being collected from uci repository. Regression, normal distribution, SVM, Naïve bayes, decision tree and ROC are applied on data to explore it.

1.1. Dataset

The data of student's performance had taken from UCI repository. It has provided information regarding the performance of students in mathematics subject. The dataset consists of 33 attributes and 395 tuples. It contains demographic features, parent's details and other activities details of students in which students are involved.

Table 3: Details of attributes of student performance dataset

Attribute	Min	Max	Mean	Median
Age	15	22	16.7	17
number of school absences (numeric: from 0 to 93)	0	75	5.709	4

School	GP: 349	MS: 46
Sex	F: 208	M: 187
Address	R: 88	U: 307
Famsize	GT3: 281	LE3: 114
Pstatus	I(together): 354	A(apart): 41

Attributes	**No**	**Yes**
Schoolsup	344	51
Famsup	153	242
Paid	214	181
Activities	194	201
Nursery	81	314
Higher	20	375
Internet	66	329
Romantic	263	132

Attributes	**1:Very bad**	**2**	**3**	**4**	**5: very high**
quality of family relationships	8	18	68	195	106
free time after school	19	64	157	115	40
going out with friends	23	103	130	86	53
workday alcohol consumption	276	75	26	9	9
weekend alcohol consumption	151	85	80	51	28
current health status	47	45	91	66	146

Medu	None: 3	primary education: 59	5th to 9th grade: 103	secondary education: 99	higher education: 131
Fedu	None: 2	primary e ducation: 82	5th to 9th grade: 115	secondary education: 100	higher e ducation: 96
Mjob	At_home:59	Health: 34	Other: 141	Services: 103	Teacher: 58
fjob	At_home:20	Health: 18	Other: 217	Services: 111	Teacher: 29
reason	Course: 145	Home: 109	Other: 36	Reputation:105	
guardian	Father:90	Mother: 273	Other:32		
traveltime	1 - <15 min.: 257	2 - 15 to 30 min.: 107	3 - 30 min. to 1 hour: 23	4 - >1 hour : 8	
studytime	1 - <2 hours: 105	2 - 2 to 5 hours: 198	3 - 5 to 10 hours: 65	4 - >10 hours: 27	
failures	0: 312	1: 50	2: 17	3: 16	

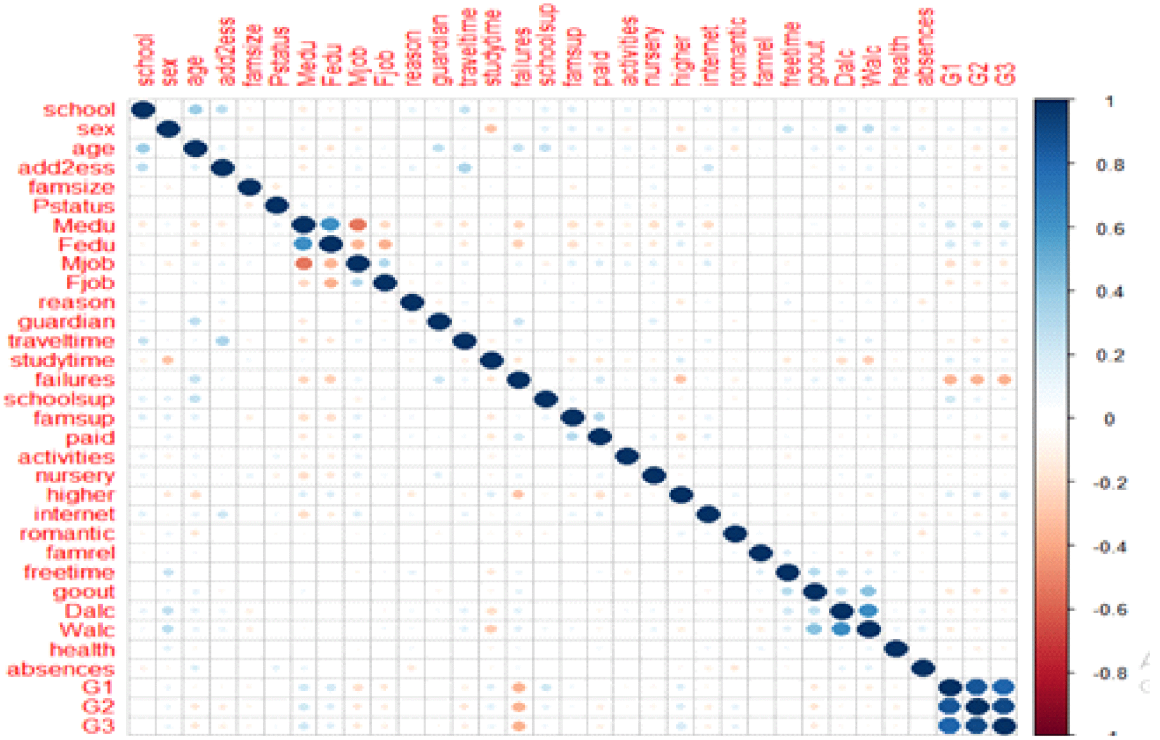
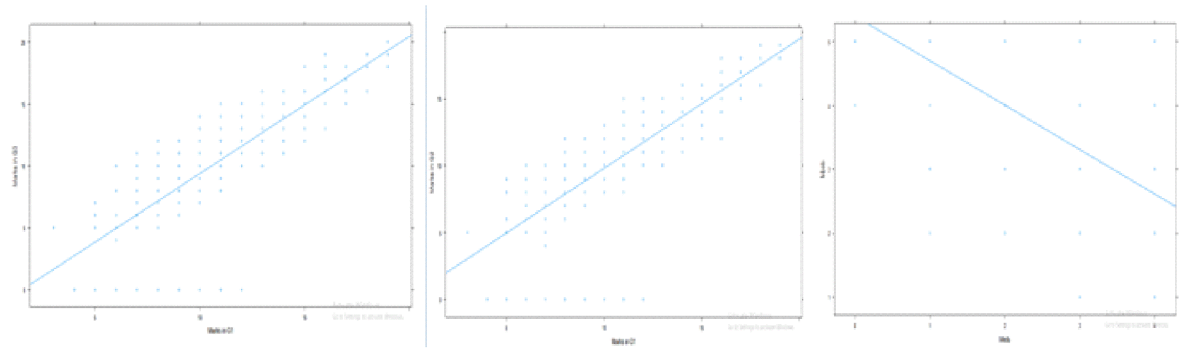
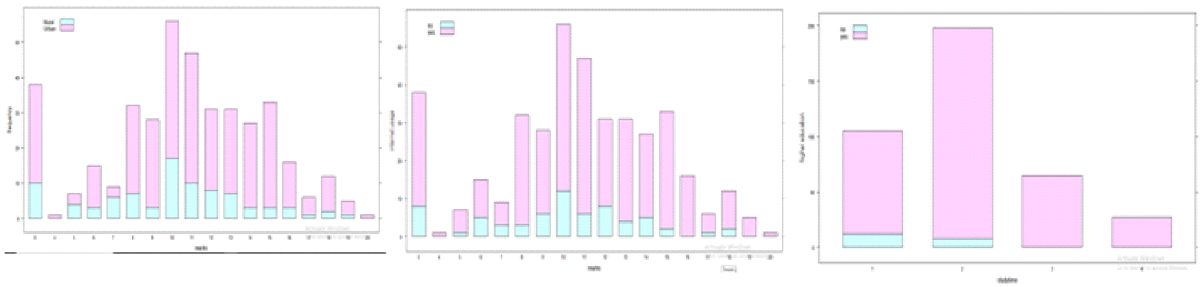


Fig 1: Correlation between all the features



(a) **(b)** **(c)**

Fig2 : Positive correlation between (a) G1 ,G2 and (b) G1,G3 ; (c) Negative correlation between Mjob,Medu.



(a) **(b)** **(c)**

Fig 3: graph between (a) G3 and residence (b) G3 and internet usage (c) Study-time and higher education.

3.2. Regression

Regression analysis is a predictive modeling technique. It estimates the relation between a dependent (target) and an independent (predictor) variable.

We use linear regression, when we have y values in range, but if value is discrete, we can't use it. When the dependent variable is categorical in nature in such cases logistic regression is used. Now since our value of y will be between 0 and 1, the linear line has to be clipped at 0 and 1. Before classifying the output we get a probability and based on that probability, we decide whether it will be yes or no. With this, our resulting curve cannot be formulated into a single formula. We need a new way to solve this kind of problem. Hence, we came up with Logistic Regression. Probability is being calculated based on training data, and we will have to decide a threshold value for 0 and 1. On average 0.5 is considered as threshold value, above which the value becomes 1 else 0.

$$Y=C+\beta_1 X_1+\beta_2 X_2+....$$

Where Range of Y is $-(\infty)$ to (∞) and in logistic regression Y value between 0 and 1.

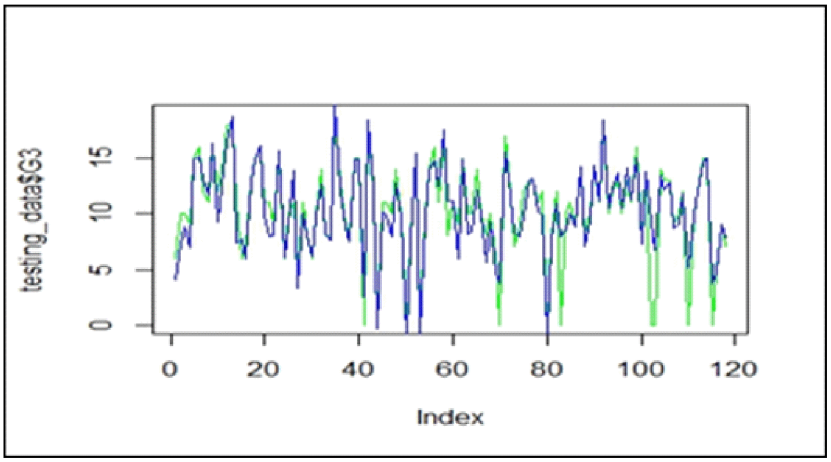
Fig 5 depicts summary of linear regression and determines significant attributes for the deciding parameter. Applying linear regression on data considering final grade (0-fail and 1-pass) as deciding parameter, 88% accuracy is obtained (fig 6).

```
> vif(model)
    school        sex        age    add2ess    famsize    Pstatus       Medu       Fedu       Mjob
  1.528861   1.478914   1.606834   1.455623   1.149941   1.170552   2.466781   2.031037   1.819452
      Fjob     reason   guardian travelтime  studytime   failures  schoolsup     famsup       paid
  1.402487   1.169594   1.266388   1.253574   1.401737   1.557951   1.330131   1.369857   1.436496
activities    nursery     higher   internet   romantic     famrel   freetime      goout       Dalc
  1.134893   1.134232   1.433087   1.194988   1.180519   1.232834   1.265789   1.545362   1.851187
    Walc     health   absences         G1         G2
  2.183341   1.167491   1.265527   4.515660   4.190080
```

Fig 4: Variance Inflation factor (multi-collinearity) of attributes

```
> summary(model)

Call:
lm(formula = G3 ~ famrel + absences + G1 + G2, data = training_data)

Residuals:
    Min      1Q  Median      3Q     Max
-9.1112 -0.4338  0.2518  1.0309  3.3633

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.73924    0.63233  -5.913    1e-08 ***
famrel       0.37381    0.12540   2.981  0.00313 **
absences     0.04677    0.01455   3.214  0.00147 **
G1           0.16964    0.06516   2.604  0.00973 **
G2           0.98284    0.05809  16.919  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.906 on 272 degrees of freedom
Multiple R-squared:  0.8315,    Adjusted R-squared:  0.829
F-statistic: 335.5 on 4 and 272 DF,  p-value: < 2.2e-16
```

Fig5 : Summary of linear regression model on data.




```
            Predictedvalue
Actualvalue FALSE TRUE
          0    21    5
          1     4   51
> (21+51)/(21+5+4+51)
[1] 0.8888889
```

Fig 6: Linear Model validation and accuracy

3.3 Normal Distribution and MLE

The goal of ML is to find the optimal way to fit a distribution to the data. There are a lot of different types of distribution for different types of data. The reason we want to fit a distribution to our data is it can be easier to work with and it is more general. It applies to every experiment of same type.

We expect most of the measurements to be closer to the mean. Expect the measurements to be relatively symmetrical around the mean. Once we settle on the shape, we have to figure out where to centre the thing. Now, we want the location that "maximizes the location" of observing the weights we measured.

In this case, we are specifically talking about "mean of the distribution" not the mean of data. We have found the value for mean on standard deviation that maximizes the likelihood that we observe the things. "Likelihood" specifically refers to this situation, we have covered here. When we are trying to find the optimal value for the mean or standard deviation for a distribution given a bunch of observed measurements.Likelihood is the probability of observations with given model parameters. The subsequent (i) indicates one particular observation x(i) among multiple observations of x.

We have the data, and what is to be determined are parameters. In Gaussian model, μ and ó are the parameters. Obtaining the parameters of our model that maximizes the likelihood of given set of observations.

$$\mu\hat{}, \acute{o}\hat{} = \arg\max p \{ (\{x_i\} | \mu, \acute{o}) \}$$

ˆ indicates the estimate to μ or ó. It is the joint probability of all data, which can be intractable if each instance of x is dependable on other instance. The joint likelihood is simply expressed as product of individual likelihood. Applying the MLE on G3 attributes, it gives the mean as 10.42 (fig 7).

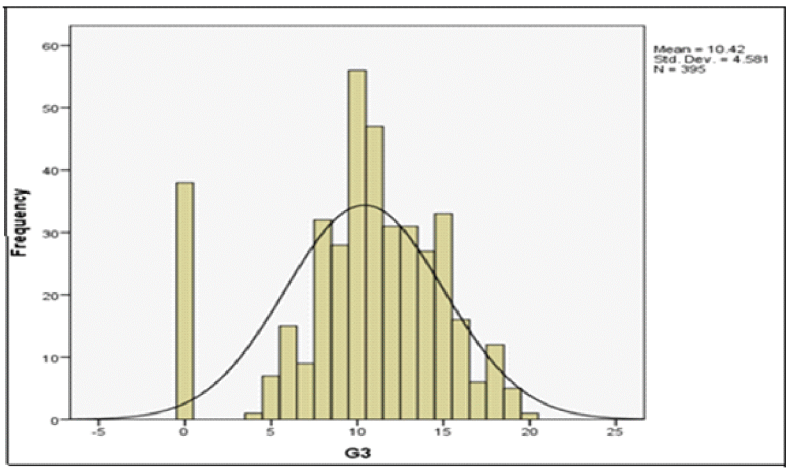


Fig 7: MLE of attribute G3 (final grade)

3.4 SVM

Support Vector Machine is a supervised technique which is being used for classification as well as for regression related challenges. Basically it is used for linearly separable binary sets. The goal is to design a hyper-plane that classifies all training vectors into 2 classes. We have many hyper-planes, but the best choice will be the hyper-plane that leaves the maximum margin from both classes, where support vectors are the coordinates (fig 8). By applying SVM with linear kernel 96.61% accuracy is obtained with confusion matrix as depicted in fig 10.

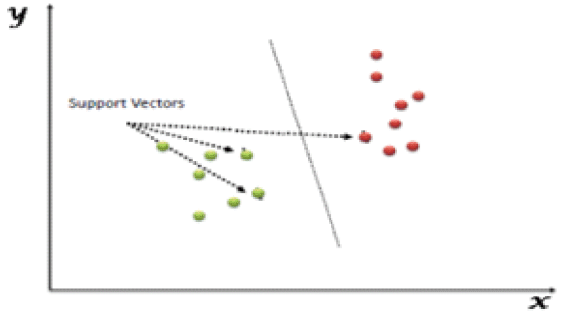


Fig 8: SVM hyper-plane and support vectors of example [7]

```
Support Vector Machines with Linear Kernel

277 samples
 33 predictor
  2 classes: '0', '1'

Pre-processing: centered (33), scaled (33)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 249, 249, 250, 250, 249, 249, ...
Resampling results:

  Accuracy   Kappa
  0.9278219  0.82007

Tuning parameter 'C' was held constant at a value of 1
```

Fig 9: Accuracy and Kappa value using SVM model

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 47  0
         1  4 67

               Accuracy : 0.9661
                 95% CI : (0.9155, 0.9907)
    No Information Rate : 0.5678
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.9303
 Mcnemar's Test P-Value : 0.1336

            Sensitivity : 0.9216
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.9437
             Prevalence : 0.4322
         Detection Rate : 0.3983
   Detection Prevalence : 0.3983
      Balanced Accuracy : 0.9608

       'Positive' Class : 0
```

Fig 10: Confusion matrix and accuracy of data using SVM

3.5 Decision Tree (J48)

Decision tree algorithm transforms raw data to rule based decision making trees.Decision tree J48 is the implementation of algorithm ID3 (Iterative Dichotomiser 3) developed by the WEKA project team. [7]

$$\text{Entrophy}(S) = \sum - p(I) \cdot \log_2 p(I)$$

$$\text{Gain}(S, A) = \text{Entrophy}(S) - \sum [\, p(S|A) \cdot \text{Entrophy}(S|A)]$$

Table 4: Results of J48

a	b	c	d	e	f
66	21	0	0	0	5
15	145	4	0	0	1
0	13	37	1	9	0
0	0	0	15	3	0
0	2	10	2	8	0
10	2	0	0	0	26

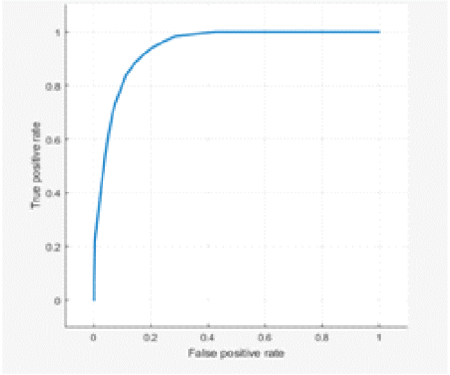
Correctly classified instances	**297**	**75.1899%**
Incorrectly classified instances	98	24.8101%

3.6 ROC

ROC curves are frequently used to show in a graphical way the connection/trade-off between sensitivity(TP) and specificity (TN).

Accuracy=TP+TN/Total

Error rate=FP+FN/Total



	+	–
+	True Positive (TP)	False Positive (FP)
–	False Negative (FN)	True Negative (TN)

Fig 11: Results of ROC curve

3.7. Naïve Bayes

Bayesian are statistical classifiers based on Baye'stheorem.P(H|X) is the posterior probability of H conditioned on X and P(H) is prior probability.

$$P(H|X)=P(X|H)P(H)/P(X)$$

Table 5: Results of Naïve Bayes

a	b	c	d	e	f	
63	21	0	0	0	8	a= E
18	132	14	0	0	1	B=D
0	6	44	1	9	0	c= C
0	0	3	14	1	0	d=A
0	1	11	0	10	0	e=B
0	5	0	0	0	30	f= F

Correctly classified instances	**293**	**74.1772%**
Incorrectly classified instances	102	25.8228%

4. Results

The performance analysis of G3 attribute applying various model is depicted in table(6). Among all the models SVM has shown the accuracy 96.61% for predicting the final grade.

Table 6: Accuracy results of different models based on attribute G3 (final grade)

S. No.	Model	Accuracy (G3)
1	Logistic regression	88.88
2	SVM	96.61
3	Decision Tree	75.18%
4	ROC	94.06
5	Naïve Bayes	74.14
6	Clustering	62

5. Conclusion and Future Work

Various parameters of student performance dataset are evaluated in this work. Considering correlations, positive relation is seen between G1, G2 and G3 attributes. For G3 (final grade) MLE for the mean is 10.42.Among different classifiers, SVM classifier used for the data classification has the highest accuracy with 96.61%. Some facts are also evaluated such as 71% G1 (A)= G3 (A), 50% G1 (B)= G3 (B), 75% G2 (A)= G3 (A) and 73% G2 (B)= G3 (B). All students who had opted for higher education had A or B final grades.Considering the internet usage parameter, 92 % students get A grade and 86% students get B grades. Considering the family size, 75% students whose family size is greater than 3 obtains A grade.The students whose guardian is mother, 68% students obtain A grade and 80% obtain B grade.Future work includes many more additional experiments on analyzing the data from other fields. Research based on hybrid approach of feature selection techniques and their impact on classification.Integration of other sources of information, different combination of feature selection methods and fuzzy set theory to improve the model development. Analyzing the results with unsupervised technique.

REFERENCES

1. H. Wenglinsky, "Teacher Classroom Practices and Student Performance/ : How Schools Can Make a Difference," Stat. Res. Div. Princeton,Princeton, NJ 08541, no. September, 2001.
2. I. Ianus, "Prediction of F reshmen Academic Performance", Research Showcase @ CMU, 2001.
3. D. Nguyen, A. Smith, P. Rose, "Author Age Prediction from Text using Linear Regression", (2010).
4. S. Sunthornjittanon, "Linear Regression Analysis on Net Income of an Agrochemical Company in Thailand", Portland State University PDXScholar, University Honors Theses (2015).
5. G. Narasinga Rao, Ch.SreenuBabu, "A Study on the Academic Performance of the Students by Applying Multiple Linear Regression Analysis using the method of Least Squares", IJETCAS 15-164; 2015.
6. Sandra W. Pyke& Peter M. Sheridan, "Logistic Regression Analysis of Graduate Student Retention", The Canadian Journal of Higher Education ,vol-23-2, 1993.
7. https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/,2/4/2018.
8. EdinOsmanbegoviæ, Mirza Suljiæ, "Data Mining Approach For Predicting Student Performance", Economic Review – Journal of Economics and Business, Vol. X, Issue 1, May 2012.
9. Michael L. Anderson, Justin Gallagher, Elizabeth Ramirez Ritchie, "School Lunch Quality and Academic Performance", NBER Working Paper No. 23218 ,2017.
10. Gina A. N. Chowa, Rainier D. Masa, Jenna Tucker, "Parental Involvement's Effects on Academic Performance", University of North Carolina at Chapel Hill, CSD Working Papers No. 13-15, 2013.
11. Prof. S. P. Singh, Savita Malik, Priya Singh, "Factors affecting Academic Performance of students Paripex-Indian Journal of Research", Vol:5, Issue:4, April 2016.
12. Ganyaupfu, E. M.,"Factors Influencing Academic Achievement in Quantitative Courses among Business Students of Private Higher Education Institutions". Journal of Education and Practice, 4(15), 57-65, 2013.
13. Grace A. Fayombo, "Emotional Intelligence and Gender as Predictors of Academic Achievement among Some University Students in Barbados", International Journal of Higher Education, doi:10.5430/ijhe.v1n1p102 ,Vol. 1, No. 1; May 2012.
14. Isabel Fialho , José Saragoça , Hugo Rebelo , Marília Cid , Manuela Oliveira , Jorge Bonito, AdelindaCandeias , VitorTrindade, "Academic achievement in public higher education quality – A study on the effects of teachers' commitment, teaching and evaluation methodologies in Nursing and Management degrees students.", Research on teacher education and training (pp. 277-290), 2012.
15. *MahadKhalifDhaqane, Nor AbdulleAfrah,* "Satisfaction of Students and Academic Performance in Benadir University", Journal of Education and Practice, Vol.7, No.24, 2016.
16. EkberTomul, GökhanPolat, "The Effects of Socioeconomic Characteristics of Students on Their Academic Achievement in Higher Education", American Journal of Educational Research 1(10):449-455, DOI: 10.12691/education-1-10-7, 2013.
17. P. Kavipriya, "A Review on Predicting Students' Academic Performance Earlier, Using Data Mining Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 12, December 2016.
18. DorinaKabakchieva, "Student Performance Prediction by Using Data Mining Classification Algorithms", International Journal of Computer Science and Management Research, Vol. 1 Issue 4 November 2012.

❑❑❑