Journal of Engineering Sciences

Vol 11, Issue 7,July/2020

ISSN NO:0377-9254

Analyzing students' performance based on various attributes

Dr. G.R.Wani[1], Mr. Anand J. Maheshwari[2]

[1]*Associate professor, BASPONC College, Bhusawal, K.B.C.N.M.U., Jalgaon, gajuwani03@gmail.com*

[2]*Assistant professor, R.C.Patel Arts, Commerce, and Science College, Shirpur, K.B.C.N.M.U., Jalgaon, a.j.maheshwari39@gmail.com*

ABSTRACT-

Information mining methodologies are explored in many enterprises as a blend of techniques for analyzing huge amounts of task-related data and to extract expressive knowledge for decision-making strategy. This technique can also be applied in the education sector for improving student's grades in various tests. The aim of this study is to infer issues like gender of students, their ethnicity, parental education, lunch they take, and preparation for the test might have impact on test scores. We examined their results by training classification models. Also, assessed and compared performances using decision trees and Bayesian classifiers.

Keywords: data mining, classifier, decision tree, grade, Bayesian

1. Introduction:

Traditionally computing technology only focus on automating routine of functional (clerical) task, improving existing process efficiency and collecting data. Even though a huge amount of data have been collected, the extent to use these available data is not being significant. In general, data is used to give results of simple queries and daily traditional reports. To manipulate these data to obtain knowledge, conventional statistical techniques were used. However, these statistical techniques are inefficient for large amount of data.

With the advent of data mining, organizations can intelligently process already collected data (historical data) as well as current data, to obtain result (knowledge) that helps in management decision.

Earlier, educational institutes are collecting huge volumes of records related to learners and teachers but the extent to use these available documents is not being significant. In our proposed work we infer various attributes of a learner like its gender, ethnicity, and parental education, type of lunch, test-preparation etc. and elaborate whether these attributes really effect in grading them. For experiment, we used decision tree and Bayesian classification to classify the grades of students using the WEKA tool. WEKA is famous machine learning tool that has machine learning procedures and visualization tools for analysis and prediction system design.

This paper is organized as follows: Section-2 described related work. Section-3 presents dataset, data-preprocessing and research methodology used for experiment, Section-4 present results and Section-5 includes analysis and observations thereof; Section-6 concludes the paper.

2. Related Work:

C. Anuradha and T. Velumurgan [1] uses various classification techniques to predict the performance of students in end semester university examinations and developed a model of student performance predictors.

Md. Imdadul Hoque et al. [4] predict student performance and also helps to improve their performance by knowing their lacking. Experiment perform with J.48, REPTree and Hoeffing decision tree classification methods and comparative study shows that, J.48 algorithm achieves highest accuracy.

3. Experimental Details:

The main objective of this study is to find out whether the facilities, background and course preparation assists to the students are really affect their grades?

For this, in this section, we present all experimental details in the following subsections namely, dataset, data-preprocessing and research methodology used respectively.

3.1 Dataset:

A standard dataset student-performance.csv is used for experiment. This dataset is open source in nature and downloaded from Kaggle.com website. We add three columns namely total, result and grade to the dataset. The data set contains 1000 records with 11 attributes. All attributes are relevant for our study.

Table 1 shows description of all attributes and their conceivable values.

Table 1. Description of attributes and their conceivable values used in dataset.

Attribute	Description	Conceivable values
Gender	Students gender	Male / Female
Race/ethnicity	Social or cultural groups	Group-A, Group-B, Group-C, Group-D, Group-E
Parental-education	Education of parents	High-school, some-high-school, some-college, bachelor-degree, associate-degree, master-degree
Lunch	Type of lunch	Standard, free or reduced
Test-preparation-course	Course-preparation	Completed or not completed
Math-score	Score of mathematics	Numeric
Reading-score	Score of reading-test	Numeric
Writing-score	Score of writing-test	Numeric
Total	Total score obtained in three subjects	Numeric
Result	Result based on total score	Pass or Fail
Grade	Average score of three scores	O : >=85 A+::>=75 &&<85 A:>=70&& <75 B+: >=60&& <70 B : >=55 && <59 C: >=45 && <54 P: >=40 && <44 F: <40

3.2 Data-preprocessing:

The original data is available in excel sheets.CSV format. Data is filled with standard values and there are no missing values. For experiment, data-preprocessing is done with the following steps:

(i) We converted original dataset into .ARFF format.

(ii) We apply WEKA → filter → supervised → attribute → discretize → Total to be

The preprocessed data is used to train the instances in data set using WEKA implementation tool as shown in figure 2.

(iii) We categorized tuples in dataset by decision tree classifier j-48 and naïve Bayesian classifier in order to assort grades based on a variety of factors, the total score obtained by the student.

(iv) Result is based on total score obtained by the student. If total score is more than 120 then the student is marked as passed otherwise false.

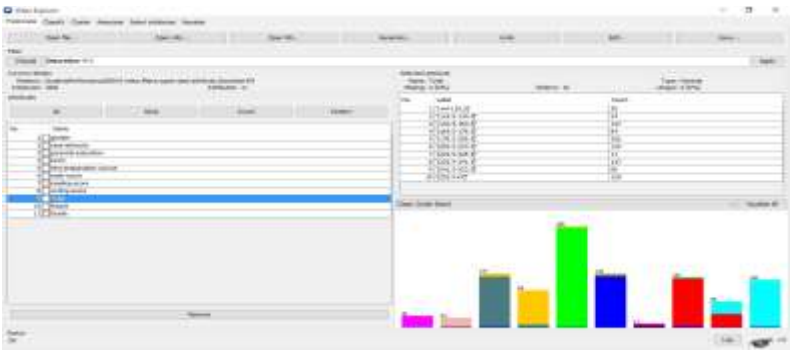

Fig. 2: Preprocessing of Data Set

3.3 Research Methodology:

To study is to find out whether the factors like gender, race/ethnicity, parental education, lunch, test preparation course are really affect their grades. The experiment is performed through two classification techniques: (i) decision tree classifier and (ii) naïve Bayesian classifier.

(i) Decision tree classifier: Decision tree induction is a top-down recursive algorithm, which uses an attribute selection measure to select the attribute tested for each non-leaf node in the tree. Tree pruning algorithm improves accuracy [4],[6]-[8].

(ii) Naïve Bayesian classifier: This is based on Bayes' theorem of posterior probability. It assume the effect of an attribute value on a given class is independent of the values of the other attributes [3], [6]-[8].

We used WEKA tool for classification. For classifying tuples, j48 decision tree classifier algorithm is chosen with test option of cross-validation with 10 folds. The output of this classifier generated with the number of object as two and seed is three. The visualization of attributes is shown in fig. 3 and the output of the Weka implementation is shown in fig. 4.

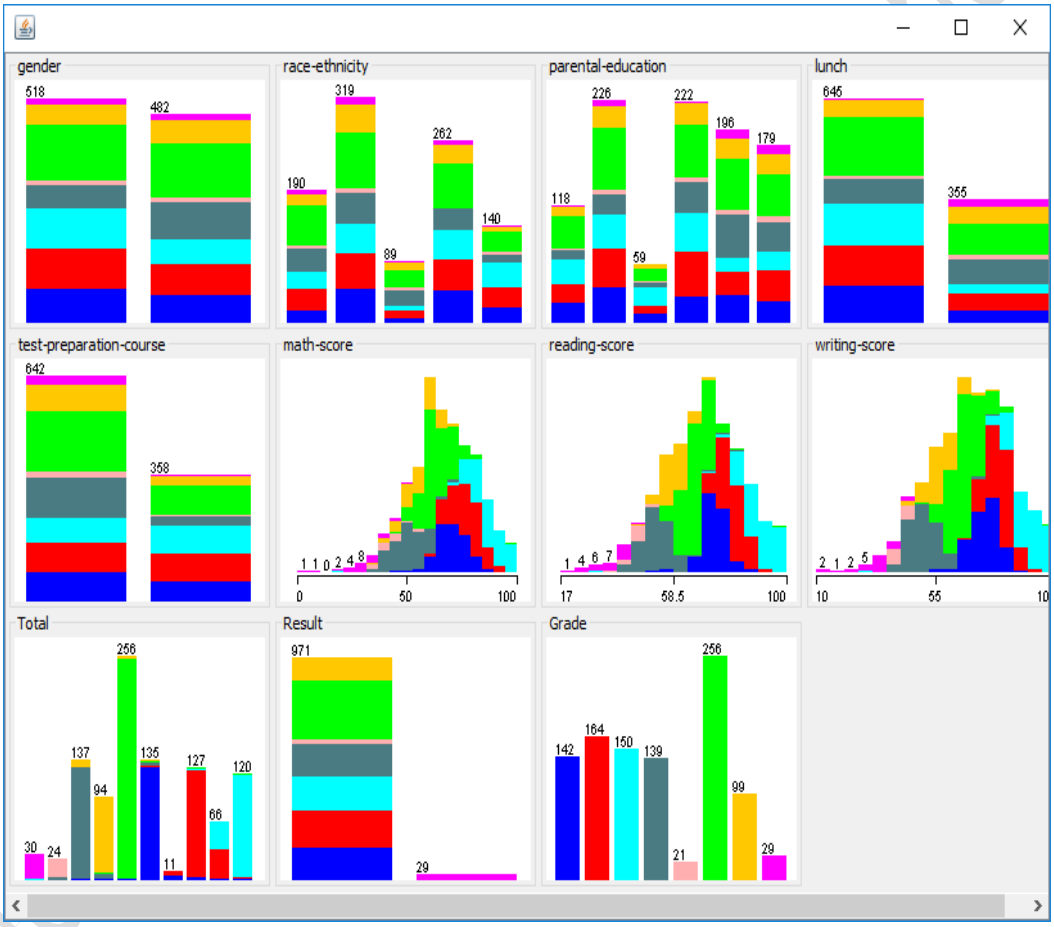

Fig. 3: Visualization of attributes

Fig, 4: Output of decision tree classifier j48

We also classify the tuples with Naïve Bayesian classifier with 10 folds of cross-validation; the generated output is shown in fig. 5.

Fig. 5: Output of Naïve Bayesian classifier.

4. Experimental Results:

This section present experimental results using both (i) J-48 classifier and (ii) Naïve Bayesian classifier.

J-48 classifier: It generates 25 nodes in the pruned tree with the size of the tree are 34. The pruned tree generated by this classifier is shown in fig 6. The total time required to build a model is 0.06 sec. For different outcome categories, true positive false positive and accurate precision results are shown in table 2.

Table 2. Experimental results of J48 classifier

Grade	TP-rate	FP-rate	Accurate Precision %
O	0.88	0.02	88%
A+	0.87	0.028	87%
A	0.923	0.014	92%
B+	0.98	0.007	98%
B	0.86	0.009	91%
C	0.914	0.012	91%
P	1.0	.0003	87.5%
F	1.0	0.001	96.7%

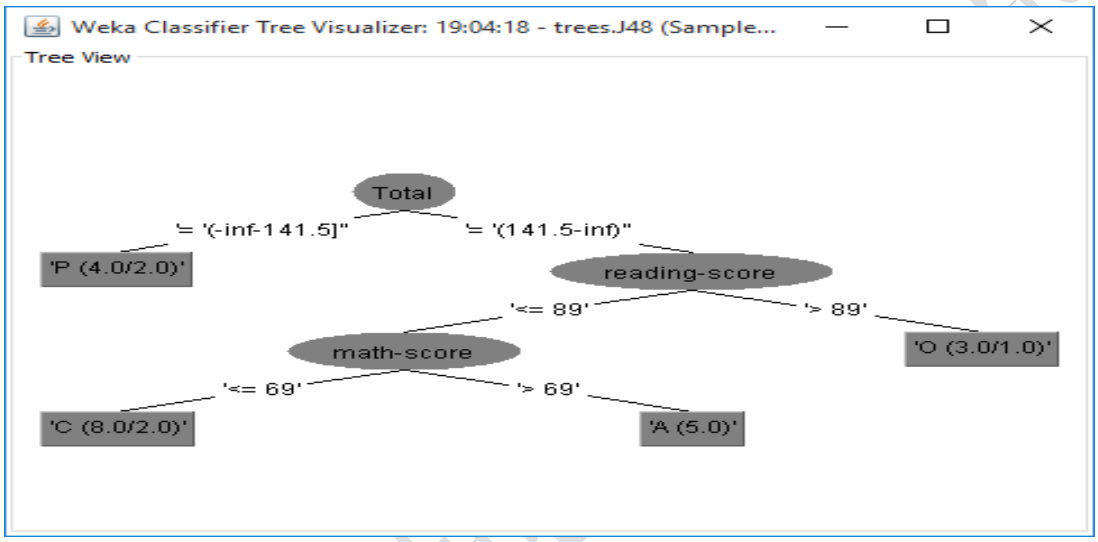

Fig. 6:J-48-pruned-tree

Naïve Bayesian classifier: This classifier classifies all the attributes based on attribute grade in the range of 0 to 1 with a distribution is as: 0.15 for O, 0.16 for A+, 0.14 for A, 0.05 for B+, 0.1 for B,0.14 for C, 0.02 for P and 0.03 for F. The time taken to build a model is 0.02 sec. For different outcome categories, true positive, false positive and precise precision results are shown in table 3.

Table 3. Experimental results of Naïve Bayesian classifier.

Grade	TP-rate	FP-rate	Precise Precision %
O	0.83	0.014	0.912
A+	0.90	0.039	0.819
A	0.915	0.012	0.929
B+	0.98	0.007	0.98
B	0.869	0.09	0.915
C	0.914	0.013	0.92
P	0.952	0.004	0.833
F	0.966	0.001	0.966

5. Analysis:

To proceed with our work initially we choose the decision tree algorithm and Bayesian classifier for our data set. The decision tree algorithm can handle both numeric and nominal attributes and provide a clear indication of which field is most probable for classification and to generate understandable rules. The accuracy of this classifier is also good, but we prefer the Bayesian classifier for our detailed analysis since decision tree classifier works on only one attribute at a time and if the attribute is numeric then splitting test is an inequality also it is subject to overfitting. These classifiers may construct models that use discrete variables with many values; hence the behavior of classifiers is not desirable. [2] Bayesian classifier works on more than one attribute at a time also it has a minimum error rate. This algorithm is robust to isolated noise points, irrelevant attributes, and can handle missing values by ignoring instance during probability estimate calculations.

In this section, from the experimental results, we analyzed the performance of student based on gender, ethnicity, parental education, type of lunch they take, and test preparation. Finally we compare both the classifier in terms of accuracy based on a dataset. The findings of experiments are as follows:

(i) Analysis of genders based on their grades: We found that, Female students performed well to achieve higher grades but male students are achieving more average grades than females as depict in Fig. 7.

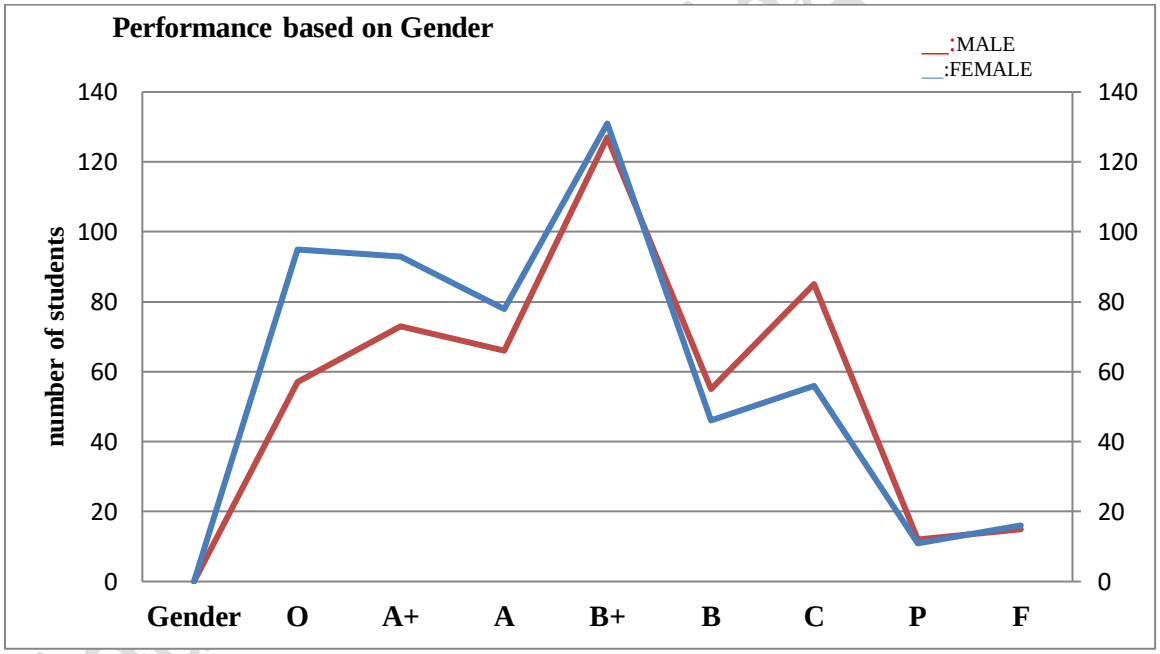

Fig. 7: Performance analysis based on gender

(ii) **Analysis on ethnicity:** we analyzed that whether ethnicity plays an important role in student performance. To do this analysis we combine higher-grade score as the sum of class obtained from O, A+, and A. The lower rank as the sum of B+, B, C, and P. Data for this analysis is taken by us from the output of naïve Bayesian classier as shown in fig 8.

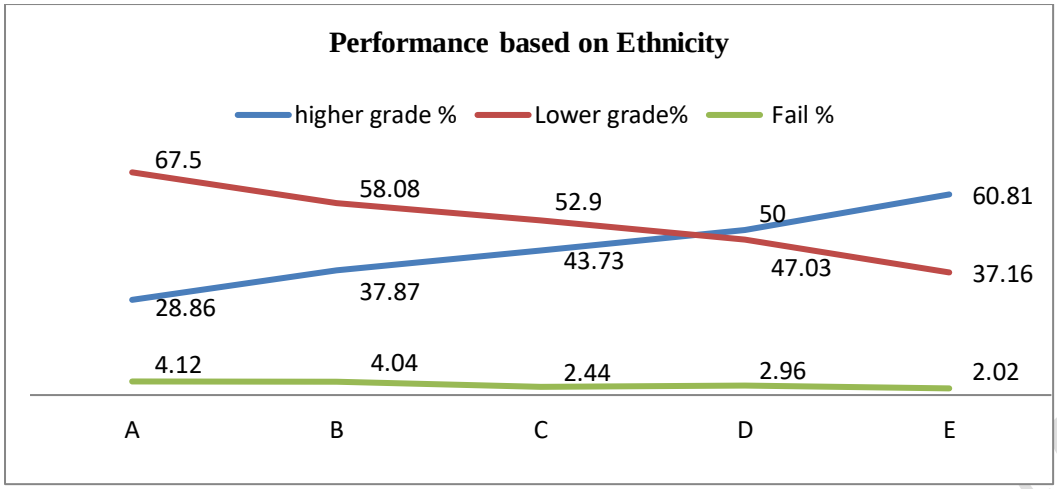

Fig. 8: Performance analysis based on ethnicity

Fig. 8 depicts that ethnicity affects the performance. Out of five groups, group D and E students rank more in higher level (O, A+ and A) and groups A, B and C students rank lower level(B+, B, C &P)

(iii) Analysis based on parental education:

Parental education can provide a significant impact on student performance. In fig. 9 students whose parents-education as bachelor-degree, associate-degree and master-degree can have higher grade percentage (O, A+ & A) than parents whose education some-school, some-high-school, and some-college level. They also have more failure rates.

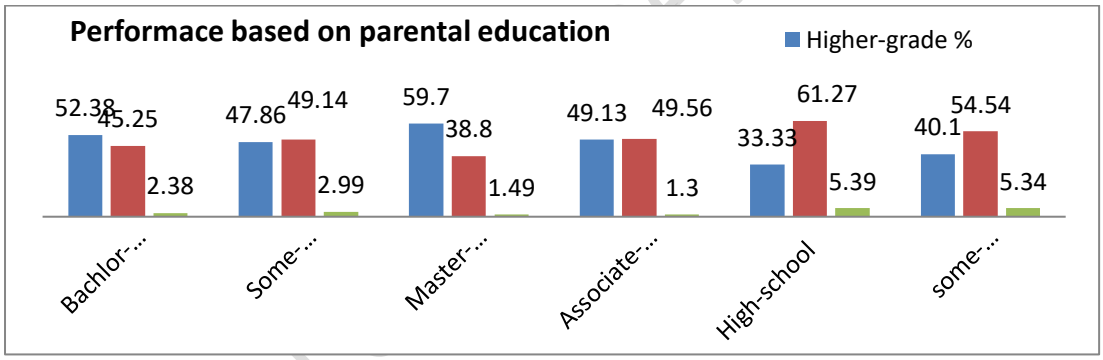

Fig. 9: Performance analysis based on parental education

(iv) **Analysis of student performance based on Lunch:** Students getting standard lunch secure more high-grade percentage (O, A+, & A) than students getting free or reduced lunch also their failure percentage is more as in fig. 11.

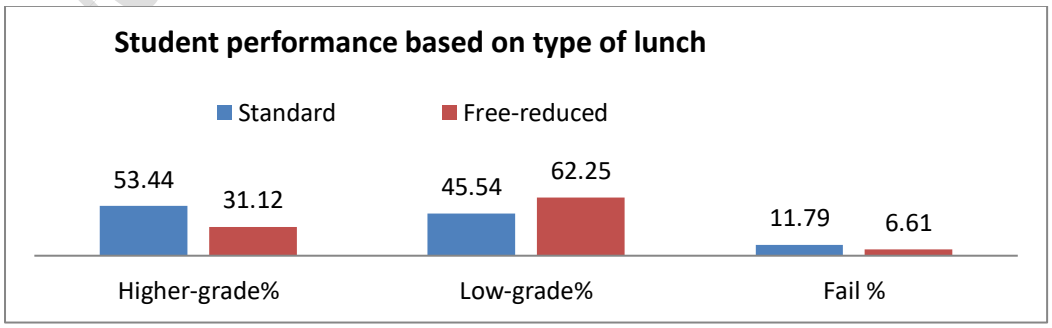

Figure 10: Performance analysis based on type of lunch

(V) Analysis of students' performance based on whether they have completed test-preparation-course or not. From fig. 11, we analyze that those students who completed the test preparation course securing higher grades than those who not to complete also the failure rate of them is also high.

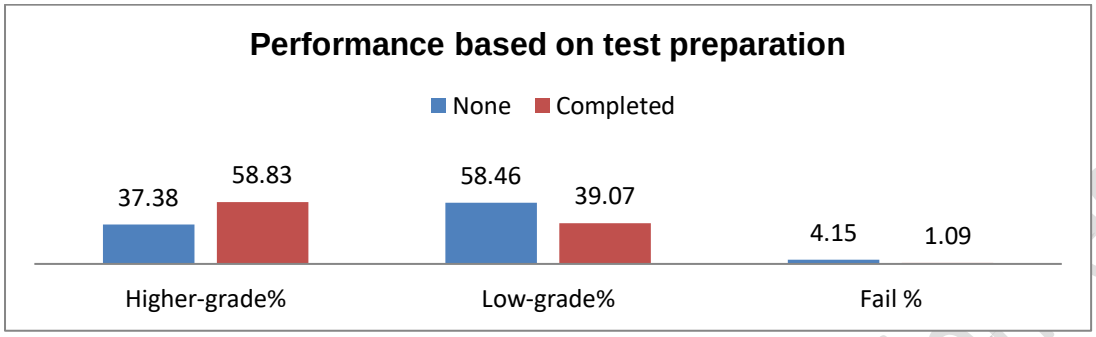

Figure 11: Performance analysis based on test preparation course

Following table 4 shows accuracy comparison of both classifiers based on dataset in a study which is observed as:

Table 4: Comparative analysis of J-48 and Naïve Bayesian classifiers.

Classifier	Correctly Classified Instances	Incorrectly Classified Instances	Kappa Statistics	Mean Absolute Error	ROC Area (Weighted Avg.)
J-48	92.1	7.9	0.90	0.0297	0.969
Naïve Bayesian	91.6	8.4	0.89	0.0246	0.982

From the above table, we say J48 is a slightly good classifier than a naïve Bayesian classier for our dataset.

6. Conclusion:

An exposure in educational-sector relies on outcomes produced by students, especially their performances in diverse exams at numerous stages of their entire education. Our aim is to categorize the student outcomes by investigating their background conditions. The analytic done by us reveals that the ethnicity, parental education, type of lunch they take and completion of test preparation course have a significant impact on grades that they secured. Experimental results show that a true positive rate for securing grades like O, A+, A, B+, B, C, P, etc. is more along with their accurate precision percentage. To do our evaluation we have used two classifiers J48 and Naïve Bayesian, the J48 classifier performs slightly well at computing accuracy but Naïve Bayesian classifier is best suited for our assessment. Naïve Bayesian algorithm deeds well in categorizing different aspects of our analysis. In conclusion, students from varied backgrounds will be highly benefited to achieve their performance improvements.

The future work comprises utilizing our work on a dataset having more attributes from students in diverse stages of their educational tasks with more accuracy. We also aim to extend the work with more exposure to other techniques like multilayer perceptron, k-means, etc.

References:

[1] C. Anuradha and T. Velumurgan, "A comparative analysis of the evaluation of classification algorithm in the prediction of students' performance", Indian Journal of Science and Technology, Vol 8(15), no IPL057, (2015)

[2] Archit Verma, "Study and Evaluation of Classification Algorithms in Data Mining", IRJET 5, no. 08,(2018)

[3] N. Chandra Sekhar Reddy, K. Sai Prasad and A Maunika, "Classification algorithms on Data Mining: A study", Indian journal of computational intelligence research 13, no. 08, (2017)

[4] Md. Imdadul Hoque, Abdul Kalam Azad, Mohamad Abu Hurayra Tuhin, Zayed Us Salehin, "University students result analysis and prediction system by decision tree algorithm", Advances in science, technology, and engineering systems Journal 5, no. 3,(2020) :115-122

[5] Evaristus Didik Madyaatmadia, Mediana Aryuni," Comparative Study of Data Mining Model for Credit Card Application scoring in Bank", Journal of Theoretical and Applied Information Technology (https://studyres.com/) 59, no.2(01 2014)

BOOK

[6] Camber, Jiawei Han and Micheline, "Data Mining Concepts and Techniques", Edited by Elsevier Vol. 2nd, Morgan Kaufmann, 2006

[7]Arun K. Pujari, "Data Mining Techniques", Vol.3rd, University Press (India) Limited, 2015.

[8]Vikram Pudi and P. Radha Krishna, Vol 5th "Data Mining", Oxford university press, 2012