VOLUME VI
ISBN No.: 978-81-953602-7-7
Computational Science

Estd. 1957

NALLAMUTHU GOUNDER MAHALINGAM COLLEGE

An Autonomous Institution, Affiliated to Bharathiar University, An ISO 9001:2015 Certified Institution,

Pollachi-642001

SUPPORTED BY

ABT Industries Ltd.,
Dairy Division

PROCEEDING

One day International Conference

EMERGING TRENDS IN SCIENCE AND TECHNOLOGY (ETIST-2021)

27th October 2021

Jointly Organized by

Department of Biological Science, Physical Science and Computational Science

NALLAMUTHU GOUNDER MAHALINGAM COLLEGE

An Autonomous Institution, Affiliated to Bharathiar University

An ISO 9001:2015 Certified Institution, Pollachi-642001.

Estd. 1957

Proceeding of the

One day International Conference on

EMERGING TRENDS IN SCIENCE AND TECHNOLOGY (ETIST-2021)

27th October 2021

Jointly Organized by

Department of Biological Science, Physical Science and Computational Science

ISBN No: 978-81-953602-7-7

Nallamuthu Gounder Mahalingam College

An Autonomous Institution, Affiliated to Bharathiar University

An ISO 9001:2015 Certified Institution, 90 Palghat Road, Pollachi-642001.

www.ngmc.org

ABOUT THE INSTITUTION

A nations's growth is in proportion to education and intelligence spread among the masses. Having this idealistic vision, two great philanthropists late. S.P. Nallamuthu Gounder and Late. Arutchelver Padmabhushan Dr.N.Mahalingam formed an organization called Pollachi Kalvi Kazhagam, which started NGM College in 1957, to impart holistic education with an objective to cater to the higher educational needs of those who wish to aspire for excellence in knowledge and values. The College has achieved greater academic distinctions with the introduction of autonomous system from the academic year 1987-88. The college has been Re-Accredited by NAAC and it is ISO 9001 : 2015 Certified Institution. The total student strength is around 6000. Having celebrated its Diamond Jubilee in 2017, the college has blossomed into a premier Post-Graduate and Research Institution, offering 26 UG, 12 PG, 13 M.Phil and 10 Ph.D Programmes, apart from Diploma and Certificate Courses. The college has been ranked within Top 100 (72nd Rank) in India by NIRF 2021.

ABOUT CONFERENCE

The International conference on “Emerging Trends in Science and Technology (ETIST-2021)” is being jointly organized by Departments of Biological Science, Physical Science and Computational Science - Nallamuthu Gounder Mahalingam College, Pollachi along with ISTE, CSI, IETE, IEE & RIYASA LABS on 27th OCT 2021. The Conference will provide common platform for faculties, research scholars, industrialists to exchange and discus the innovative ideas and will promote to work in interdisciplinary mode.

EDITORIAL BOARD

LIST OF ARTICLES

International Conference on Emerging Trends in Science and Technology (ETIST 2021)
Jointly Organized by Department of Biological Science, Physical Science and Computational Science
Nallamuthu Gounder Mahalingam College, Affiliated to Bharathiar University, Tamilnadu, India.
Published by NGMC - November 2021

Emphasizing the Factors Affecting Students' Academic Performance In Higher Education Using Feature Selection Techniques

Dr.S.Hemalatha

ABSTRACT: The aim of this proposed work is to rank the factors influencing Students' Academic Performance using various Feature Selection Algorithms. It is essential, in education to categorize the students who are facing problems in studies so that preventive actions can be taken to improve the students' performance. It is useful in identifying slow learners who are likely to accomplish poorly in studies. Here, the Feature Selection Algorithms like Correlation Chi Square, Extra Trees Classifier, Random Forest Classifier, Information Gain, Mutual Information,Univariate feature selection, etc. are implemented and the results are recorded.Feature Selection is a dynamic and vibrant research area in the field of Machine Learning and Data Mining. The key purpose of these Feature Selection algorithms is to select the most predictive elements from the chosen dataset for analysis and pay no attention to the rest of the attribute, which is non-predictive. It concludes that non-predictive features are not influencing the actual result, but the complexity of the analysis results is reduced.The accuracy and effectiveness of the student's performance prediction model can be improved with the help of these feature selection algorithms. The performance of different classifiers isoutsourcedby performing the experiments and the highest percentage of prediction accuracy is documented.

Keywords: Feature Selection, Prediction, Correlation, Efficiency, Accuracy.

1. INTRODUCTION

The development of the country is emotionally involved to the excellence of its Education System. Education is one of the most needed constituents in generating notable members of society. The data kept in academic institution databases plays noteworthy role for the improvement of educational process by exploring the hidden information [9].Many authors used feature selection (FS) algorithms in combination with classification algorithms to compare the prediction accuracy of varying student dataset. Feature Selection algorithms are used in in pre-processing step of data. It supports to select the appropriate subset of features to construct a model for predictions in Machine Learning. However, Feature Selection algorithms are utilized to improve the predictive accuracy and lower the computational complexity [10], [5], [7].The feature selection problem has the goal of finding the smallest subset of all features that are the most relevant for a machine learning task. Feature selection is regarded as either a single objective or multi-objective combinatorial optimization problem. There are anumber of objectives to consider when selecting a feature selection algorithm depending on the application at hand, namely Simplicity, Stability, Number of Reduced Features, Classification Accuracy, Storage and Computational Requirements. The application of machine learning approaches to

Dr.S.Hemalatha, Associate Professor, Department of Computer Applications,
Nallamuthu Gounder Mahalingam College, Pollachi, India. E-mail: tohemaragu@gmail.com

educational data is a recent trend in research. The activities of this research include feature engineering to create the students dataset, data collecting, data preprocessing, creating and evaluating machine learning models, finding the best model and analyzing the results.

Machine learning algorithms are techniques for calculating approximately the target function (f) for predicting the output variable (Y) with the given input variables (X).Linear regression algorithms fit a straight line, or another function that is linear in its parameters such as a polynomial, to numeric data, typically by performing matrix inversions to decrease the squared error between the line and the data. Nonlinear regression algorithms, which fit curves that are not linear in their parameters to data, are a little more difficult, because, unlike linear regression problems, they can't be solved with a deterministic method. Machine learning algorithms are even less clear-cut than nonlinear regression, partly because machine learning dispenses with the constraint of fitting to a specific mathematical function, such as a polynomial. There are two major categories of problems that are often solved by machine learning: regression and classification. Regression is for numeric data and classification is for non-numeric data.

2. LITERATURE SURVEY

HavanAgrawal et al [3]proposed a model is proposed to predict the performance of students in an academic organization. The algorithm employed is a machine learning technique called Neural Networks. Further, the importance of several different attributes, or "features" is considered, in order to determine which of these are correlated with student performance. *Muhammad Imran et al [8]proposed that i*n previous studies, it has been observed that many researchers have intension on the selection of appropriate algorithm for just classification and ignores the solutions of the problems which comes during data mining phases such as data high dimensionality, class imbalance and classification error etc. Such types of problems reduced the accuracy of the model.According toIsabel Fialho et al [4] the main challenge of the XXI century in the educational field is not only to assure that students acquire a wide range of skills which are permanently updated, but also to guarantee the quality of those learning experiences. Eyman Alyahyan et al., [2] aims to provide a step-by-step set of guidelines for educators willing to apply data mining techniques to predict student success. For this, the literature has been reviewed, and the state-of-the-art has been compiled into a systematic process, where possible decisions and parameters are comprehensively covered and explained along with arguments.Mahad Khalif Dhaqane et al., [6]study examines the role of satisfaction on students' academic performance and investigates the relationship between satisfaction of students and academic performance and explores other factors that contribute academic performance.

3. METHODOLOGY

The main objective of the proposed work is

- To study the factors affecting academic performance of the students.
- To implement the various feature selection algorithm in order to attain the selected features for finding out the key indicators that influences students' academic performance.
- To implement various classifier algorithms in order find the efficiency and the performance of it over the selected features.
- Finally, finding a model to predict the factors that affect the performance of the student.

The data collection is the preliminary step of the research work. The methodology is applied to a genuine data containing information about the students of various Department of Computer science streams during the academic years ranging from 2019 to 2022 through questionnaires. The questionnaires that are built on Google Forms are used to conduct a survey on students from Nallamuthu Gounder Mahalingam College, GVG College and STC College, Coimbatore district, Tamilnadu. Around Thirty MCQ type of psychometric response scale in which responders specify their level of agreement to a statement typically in five likert scale points: (1) Always (2) Often (3) Sometimes (4) Rarely (**5**) Never are designed. A total of 230 questionnaires were completed after combining the CSV files from Google Forms. The research sample (230 answers) represents an acceptable sample and the data collected are further processed by Machine Learning tools and Statistical tools.Implementation of classifiers is done Using Jupiter Notebook and Python Programming to find the factors influencing academic factors in higher education.

The Insights of the basic features like Gender, Area of Residence(AoR), Family Income/Month(FIPM) and Parent Educational Qualification(PE) with the Target Variable is show in the below graphical representation Figure 1.

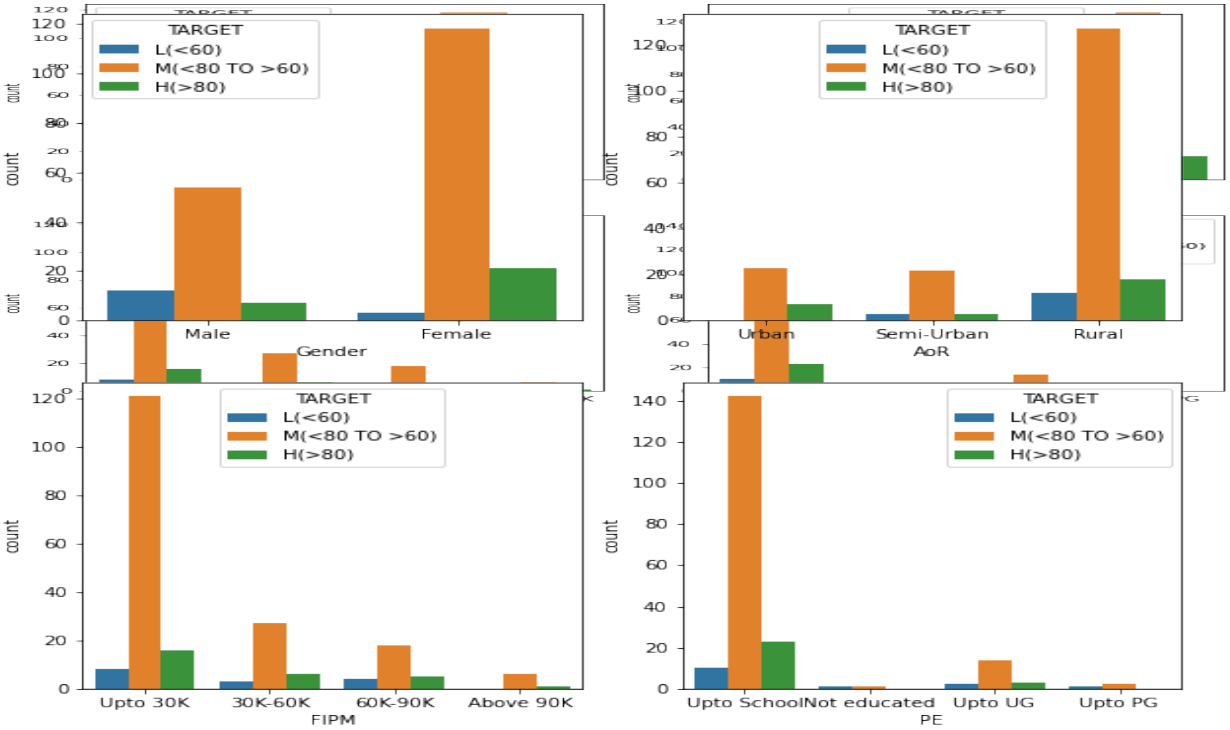

Figure 1.Insights of the Basic Features (Demographic)

The Label encoding is done for preprocessing the data in which each text label value is replaced with a number. Here in the proposed work the target variable which has been categorized as H, L and M is encoded as 0, 1 and 2 respectively.

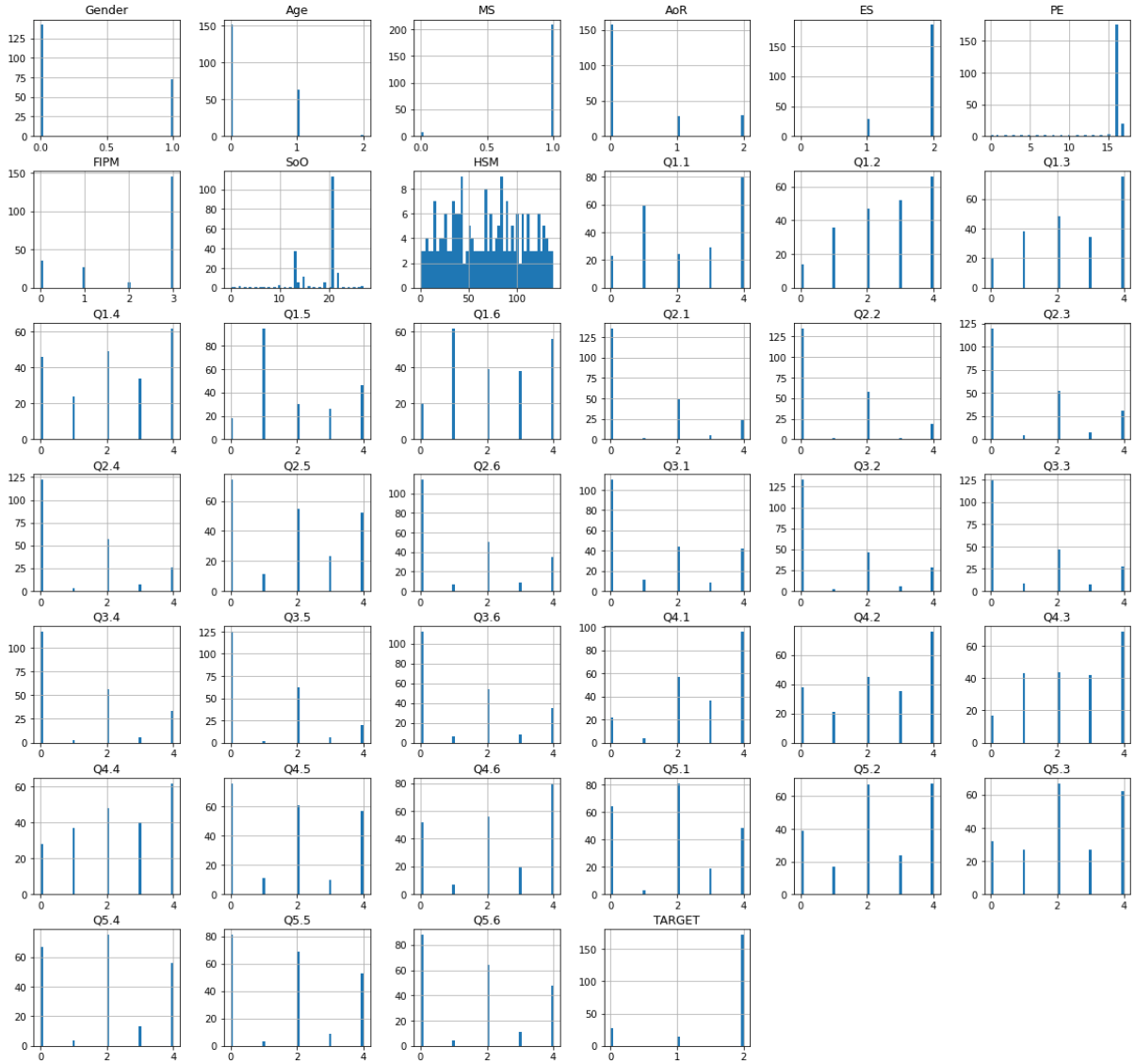

Figure 2.Insights of Label Encoding-Categorical Data (Text Format to Numeric Format)

3.1 FEATURE SELECTION

Feature selection and Data cleaning are considered to be the first and foremost steps for designing model. Feature Selection is the process where the features are selected automatically or manually which is contributed most to the prediction variable or output in which it is concerned. Feature selection is also known as variable selection or attributes selection; this can be used interchangeably for the purpose of this work. Feature selection is the process of selecting a subset of the relevant features for use in model construction (Brownlee, 2016) [1].

There are several Feature Selection Techniques such as Correlation, Chi Square, Extra Trees Classifier, Random Forest Classifier, Information Gain, Mutual Information, Univariate feature selection, Genetic Algorithms, etc. The different Feature are selected using Selection Techniques are recorded as below.

Table 1. Selected Features Using Feature Selection Techniques

	SFS (LR)	SFS (KNN)	SBS (LR)	SBS (KNN)	Chi-Square	ETC	RFC	IG	MI	Uni variate	GA	Rank
Gender		✓			✓	✓		✓	✓	✓	✓	7
Age	✓	✓	✓		✓			✓	✓	✓		7
ES	✓	✓	✓			✓	✓	✓	✓	✓		8
HSM	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓	10
Q1.2	✓	✓	✓			✓	✓		✓	✓	✓	8
Q1.3	✓	✓	✓	✓		✓	✓	✓	✓		✓	9
Q2.1	✓	✓	✓		✓			✓	✓	✓		7
Q2.5	✓		✓		✓		✓	✓	✓	✓	✓	8
Q3.5	✓	✓	✓	✓	✓	✓			✓	✓		8
Q4.1	✓	✓	✓	✓		✓	✓		✓	✓		8
Q4.3	✓	✓	✓		✓	✓	✓			✓	✓	8
Q4.4	✓	✓	✓	✓	✓		✓	✓		✓	✓	9
Q5.3	✓	✓	✓		✓	✓		✓		✓		7

The 13 features that are influencing student's academic performance from the above mentioned algorithms are listed in Table 10. The factors are Gender, Age,ES, HSM, Q1.2, Q1.3, Q2.1, Q2.5, Q3.5, Q4.1, Q4.3, Q4.4, Q5.3.

Table 2. Number of features selected using feature selection technique according to the factors influencing to the academic performance of the college students

	SFS (LR)	SFS (KNN)	SBS (LR)	SBS (KNN)	Chi-Square	ETC	RF C	IG	MI	Uni variate	GA	Total
Demographic Profile	6	6	6	3	4	4	3	5	6	4	5	**52**
Exam Related Factors	3	4	3	5	1	5	4	2	4	1	5	**37**
Department / Staff Related Factors	3	3	3	3	6	0	1	4	2	6	4	**35**
Social Media Related Factors	3	2	3	3	4	5	1	3	4	4	1	**33**
Extra-Curricular Activities	3	3	3	3	4	2	3	3	1	4	4	**33**
Personal Factor	2	2	2	2	1	3	1	3	3	1	2	**22**

From the above table 2, it is exemplified that the highest influencing factor is the Demographic Profile in which it secures as the total count of 52 evaluating with all the 11 techniques as mentioned above. Then the second highly influencing factor is Exam Related Factors. The next factors which is getting influenced is Department / Staff Related Factors, Social Media Related Factors, Extra-Curricular Activities and the least factor through which the students are getting influenced is the Personal Factor.

3.2 RANKING OF FACTORS INFLUENCING STUDENTS' ACADEMIC PERFORMANCE

According to the count as prescribed in the above table 2, the different factors are ranked and listed in the table 3. It is prevailed that the Demographic profile such as Gender, Age, Marital Status, Area of Residence, Educational Status, Parents Educational Qualification, Family Income per Month, State of Origin and Higher Secondary Mark greatly influences the students' Academic Performance and it is ranked number 1.

Table 3. Six Different Factors and its Ranking

FACTORS	RANKING
Demographic Profile	**1**
Exam Related Factors	**2**
Department / Staff Related Factors	**3**
Social Media Related Factors	**4**
Extra-Curricular Activities	**4**
Personal Factor	**5**

The Exam Related Factors has ranked number 2, but these factors give positive influences over the academic performance of the students. Since Department/Staff related factors are ranked number 3, no factors are highly influencing the students' academic performance. Social Media Related and Extra-Curricular Activities Related factors are ranked in position 4 stating that student's engagement to ECA and social media rarely affects students Grade Point Average (GPA). Finally, it is concluded that personal factor is the least factor which doesn't influence student performance to the great extent.

4. RESULTS AND DISCUSSIONS

It is worth discussing these interesting facts revealed for Predicting students' academic performance. It is widely useful to help the educators and learners improving their learning and teaching process. This is an important finding in the understanding of the dependent and independent factors which influence the performance of the students in the higher Education.

IMPLEMENTATION OF CLASSIFIERS

Implementation of classifiers is done Using Jupiter Notebook and Python Programming to find the factors influencing academic factors in higher education. The implications of accuracy using all features and the selected features are processed and the results are recorded using machine learning algorithms such as SVM, Decision Tree, Logistic Regression and Random Forest Algorithm. The below table depicts the confusion matrix and accuracy for all the features in the dataset. The features are 'Gender', 'Age', 'MS', 'AoR', 'ES', 'PE',

'FIPM', 'SoO', 'HSM', 'Q1.1', 'Q1.2', 'Q1.3', 'Q1.4', 'Q1.5', 'Q1.6', 'Q2.1', 'Q2.2', 'Q2.3', 'Q2.4', 'Q2.5', 'Q2.6', 'Q3.1', 'Q3.2', 'Q3.3', 'Q3.4', 'Q3.5', 'Q3.6', 'Q4.1', 'Q4.2', 'Q4.3', 'Q4.4', 'Q4.5', 'Q4.6', 'Q5.1', 'Q5.2', 'Q5.3', 'Q5.4', 'Q5.5', 'Q5.6'.

Table 4.Classifier Algorithm implemented for all features

Algorithm	Confusion Matrix	Accuracy
Random Forest	[0 0 11] [0 0 4] [0 0 50]	0.7692
Logistic Regression	[2 0 9] [0 0 4] [10 4 36]	0.5846
Support Vector Machine	[0 0 11] [0 0 4] [0 0 50]	0.7692
Decision Tree	[3 0 8] [1 0 3] [8 4 38]	0.6307

The above table 4 record the accuracy of Random Forest Algorithm which is 0.7692, Logistic Regression is 0.5846, SVM is 0.7992 and the Decision Tree is 0.6307. The result demonstrates that the performance of Random Forest and SVM is better than other classifiers.

Table 5. Results Using Different Classifiers for Selected Features

Algorithm	Confusion Matrix	Accuracy
Random Forest	[0 0 11] [0 0 4] [0 0 50]	0.7692
Logistic Regression	[1 0 10] [0 1 3] [0 1 49]	0.7846
Support Vector Machine	[0 0 11] [0 0 4] [0 0 50]	0.7692
Decision Tree	[2 0 9] [0 1 3] [7 3 40]	0.6615

The above table 5 exposes the Confusion Matrix and Accuracy for the selected features 'Gender', 'Age', 'ES', 'HSM', 'Q1.2', 'Q1.3', 'Q2.1', 'Q2.5', 'Q3.5', 'Q4.1', 'Q4.3', 'Q4.4', 'Q5.3' which are selected using the different feature selection techniques. As illustrated in Table 4.10, the performance of classifiers such as

Random Forest, Logistic Regression and Support Vector Machine are presented as 0.7692, 0.7846 and 0.7692 respectively.

Table 6. Comparison of All Features with Selected Features

	Selected Features		All Features	
Algorithm	**Confusion Matrix**	**Accuracy**	**Confusion Matrix**	**Accuracy**
Random Forest	[0 0 11] [0 0 4] [1 0 49]	0.7538	[0 0 11] [0 0 4] [0 0 50]	0.7692
Logistic Regression	[1 0 10] [0 1 3] [0 1 49]	**0.7846**	[2 0 9] [0 0 4] [10 4 36]	**0.5846**
Support Vector Machine	[0 0 11] [0 0 4] [0 0 50]	**0.7692**	[0 0 11] [0 0 4] [0 0 50]	**0.7692**
Decision Tree	[2 0 9] [0 1 3] [7 3 40]	0.6615	[3 0 8] [1 0 3] [8 4 38]	0.6307

From the table 6, it is concluded that the performance of Logistic Regression has been improved while selecting 13 features rather than selecting all features. Also, it is proved that the performance of SVM is same for all features and the selected features. The accuracy is same and recorded as 0.7692. Hence, these two algorithms can be considered in predicting the factors influencing students' academic performance.

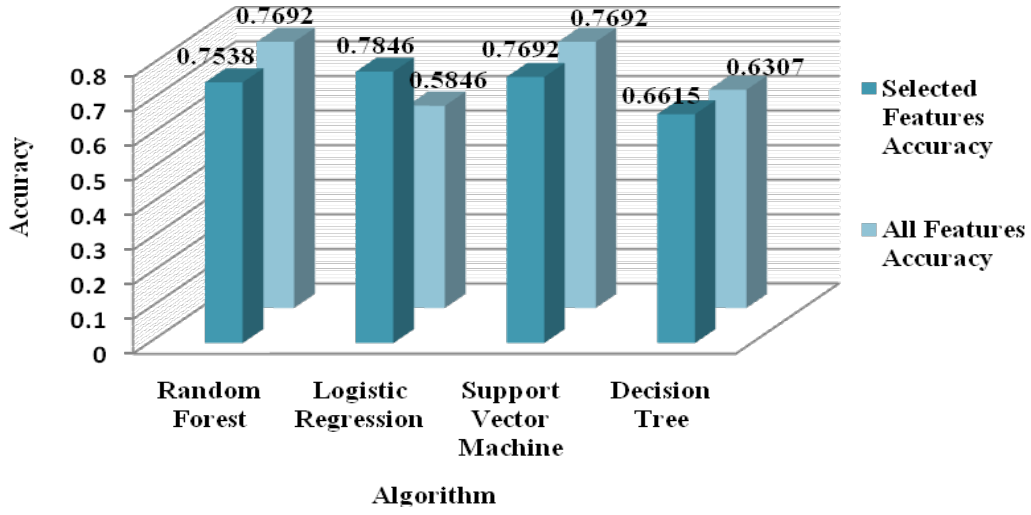

Figure 3.Performance of Different Classifiers

The above figure 3 shows the performance of different classifiers for all features and the selected features. The accuracy for all features in Logistic Regression is 0.5846 and for selected features is 0.7846. So it

is concluded that the LR is the best classifier when compared to the other algorithms. SVM can also be deployed since the accuracy for all the features and the selected features are same.

Table 7. Precision, Recall, F-Measure Using Different Classifiers

ALGORITHM	CLASS	PRECISION	RECALL	F1-SCORE
Random Forest	0 1 2	0.00 0.00 0.77	0.00 0.00 0.98	0.00 0.00 0.86
Logistic Regression	0 1 2	1.00 0.50 0.79	0.09 0.25 0.98	0.17 0.33 0.87
Support Vector Machine	0 1 2	0.00 0.00 0.77	0.00 0.00 1.00	0.00 0.00 0.87
Decision Tree	0 1 2	0.22 0.25 0.77	0.18 0.25 0.80	0.20 0.25 0.78

The results reported in the above table 7 are exhibiting the performance of different classifier algorithms for 13 selected features which influences the student academic performance. The result shows that Precision, Recall and F-Measure of Logistic Regression are better when compared to other classifier algorithms such as Random Forest, SVM and Decision Tree.

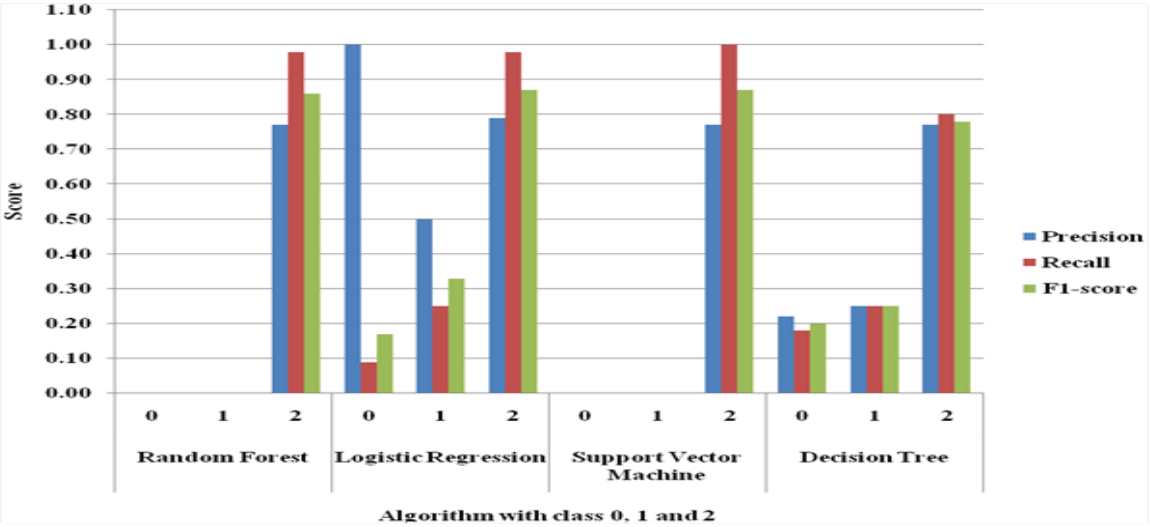

Figure 4.Performance of Algorithms with class 0, 1 and 2

The above figure 4 shows the performance of different classifiers with class 1 (Low), class 0 (High) and class 2 (Medium). Here Low, High and Medium are the grades that can be used for predicting the students' academic performance when the model is generated with the training data set.

5. CONCLUSION

Prediction of students' academic performance is mostly useful to facilitate the educators and learners to improve their teaching and learning method. This paper presents the study of various feature selection algorithms and analyzed their performance using student information datasets. The number of features is reduced and their performances in terms of accuracy are recorded in this work. The feature selection results are recorded through confusion matrix also. From the results and works, it is concluded that the feature selection technique on evaluating students' academic performance, Logistic Regression and SVM is better in keeping optimal and best attributes for prediction. In future, this study can be carried out for larger sample size in which the result might be improved than the existing one. Many other variables like teacher education, class environment, friends circle could be analyzed for further result.

REFERENCES

[1] Brownlee, J. "Machine Learning Algorithms," Machine Learning Mastery, Available: http://machinelearningmastery.com/. Diakses 15 May 2017.
[2] EymanAlyahyan et al.,"Predicting academic success in higher education: literature review and best practices", International Journal of Educational Technology in Higher Education, Vol.17, No.3, 2020.
[3] HavanAgrawal, HarshilMavani, "Student Performance Prediction using Machine Learning", International Journal of Engineering Research & Technology (IJERT) Vol. 4, No. 03, March 2015.
[4] Isabel Fialho et al., "Academic Performance in Public Higher Education Institutions: A Study on the Effects of Teacher Commitment, Teaching Methodologies and Evaluation Methodologies in Students attending Nursing and Management Courses".
[5] Koller.D and M. Sahami, "Toward optimal feature selection," Stanford InfoLab1996.
[6] MahadKhalifDhaqane et al., "Satisfaction of Students and Academic Performance in Benadir University", Journal of Education and Practice, Vol.7, No.24, 2016.
[7] Mitra.P,C. Murthy, and S. K. Pal, "Unsupervised feature selection using feature similarity," IEEE transactions on pattern analysis and machine intelligence, Vol. 24, pp: 301-312, 2002.
[8] Muhammad Imran, ShahzadLatif, "Student Academic Performance Prediction using Supervised Learning Techniques", International Journal of Emerging Technologies in Learning, March 2019.
[9] Osmanbegović.E, M. Suljić, and H. Agić, "Determining Dominant Factor For Students Performance Prediction By Using Data Mining Classification Algorithms," Tranzicija, Vol. 16, pp: 147-158, 2015.
[10] Ramaswami.M and R. Bhaskaran, "A Study On Feature Selection Techniques In Educational Data Mining," arXiv preprint arXiv:0912.3924, 2009.