Student's Academic Performance Prediction Using Machine Learning Approach

Vairachilai S[1*], Vamshidharreddy[1*]
[1]Department of Computer Science and Engineering, Faculty of Science and Technology (IcfaiTech), The ICFAI Foundation for Higher Education (IFHE), Hyderabad-501203.
Avvari Sai Saketh[2]
[2]Machine Learning Engineer, Nallakunta, Hyderabad-500044, Telangana, India.
Gnanajeyaraman R[3]
[3]Department of Computer Science and Engineering, SBM College of Engineering and Technology, Dindigul, Tamil Nadu 624005, India.

Abstract

Predicting academic performance is an important task for the students in university, college, and school, etc. The factors which affect the student's academic performance are class quizzes, assignments, lab exams, mid, and final exams. The student's academic performance should be informed to the class teacher in advance that will decrease the student's dropout and increase the performance. In this paper, machine learning classification algorithms such as decision tree, Support Vector Machine (SVM), and Naive Bayes are implemented to predict the student's academic performance. The performance of an algorithm has been evaluated based on confusion matrix, accuracy, precision, recall, and F1 score. The obtained result shows that the Naive Bayes classification algorithm performs better.

Keywords: *Machine Learning, Classification, Prediction, Decision Tree, Naive Bayes, Support Vector Machine.*

1. Introduction

Student's academic performance is a crucial part of an academic institution. This is considered as one of the important measures for many superior universities. Some researchers stated that the student's academic performance can be measured through learning assessment and co-curriculum activities. Though, the majority of researchers have mentioned that the student's past performances, achievements, and grades can play a vital role to predict the student's success rate. Predominantly, most of the higher-level institutions use grade as the main measure to assess student's performance. In addition, course structure, assignment marks, final exam scores, and extracurricular activities will affect the student's academic performance. The student's academic program can be well planned during their sophomore period of studies in an institution to analyze the performance of students. At present, machine learning algorithms are most popular to evaluate student's academic performance that has been extensively applied in the education sector. Mining the educational data used to predict the student's academic performance (Brijesh Kumar Baradwaj, Saurabh Pal, 2011). As a result, it would help the educators/faculty to improve the teaching approach in a constructive way. In addition, the teacher could observe student's achievements also. Nawal Ali Yassein et al. (2017) used classification and clustering techniques to predict student's academic performance for KSA (Kingdom of Saudi Arabia). In this research, the features which affect the student's academic performance is analyzed to predict the student's academic performance. The practical work and assignments given by the course instructor is the main factor for the student's success rate in the academic performance. They identified that student attendance in class is the most important factor than the final exam and the mid-exam grades. Md. Hedayetul Islam Shovon and Mahfuza Haque (2012) used the decision tree method to predict student's academic performance and data clustering method to predict General Point Average (GPA) that helps the instructor to improve the student's academic performance. Thaddeus Matundura Ogwoka et al. (2015) proposed a model to predict student's academic performance with high accuracy of 98.8439%. In this model, the data mining algorithm such as k-means and decision tree is used to predict the student's academic performance in advance. These mechanisms alleviate student dropout rates and improve student's academic performance. Prashant Sahai Saxena and M. C. Govil (2014) analyzed the relation between student's behavior and their success by using the clustering algorithm (O.J. Oyelade et.al., 2010). The variables used in this research are gender, parent's education, location, and parent's occupation, etc. The parent's occupation is a significant factor to predict the student's academic performance. Nguyen Thai-Nghe et.al. (2009), Carlos Márquez-Vera et.al. (2013) analyzed how to deal the imbalance class to improve the student's academic performance. Francisco Araquea et.al. (2009) analyzed the various factors which will affect the student's academic performance. The rest of the paper is organized as follows: Section 2, explains the description of the dataset. Section 3, explains the experiment results and performance analysis, Conclusion and future work are explained in Section 4.

2. Dataset Description

This dataset for the current study was collected from the website https://www.kaggle.com. It consists of 480 instances, 16 independent variables, and one dependent variable. The student's academic performance dataset is shown in Table 1. It consists of 305 male and 175 female students. During the first semester 245 student's records are collected and 235 records are collected during the second semester. This dataset includes the school attendance feature like how many students are absent for more than seven days. This also includes parent participation in the educational process. The sample dataset values are shown in Table 2. The students are classified into three numerical intervals based on their total grade/mark, interval values from 0 to 69 represents the low-level, interval values from 70 to 89 represents the middle-level, and interval values from 90 to100 represent high-level. In Table 2, the following shortcuts are used such as G–(gender), N–(NationalITy), PB-(PlaceofBirth), SID-(StageID), GID-(GradeID), SID-(SectionID), T-(Topic), SE-(Semester), R-(Relation), RH-(Raisedhands), VR-(VisITedResources), AV-(AnnouncementsView), D-(Discussion), MS-(Middle School), Lwl(Lower level), PAS-(ParentAnsweringSurvey), PS-(ParentschoolSatisfaction), SAD-(StudentAbsenceDays). As part of preprocessing, label encoding is applied to the dataset that converts the labels into numeric form. The sample student's academic performance dataset values after encoding is shown in Table 3. The independent variables histogram is shown in Figure 1 and box plot is shown in Figure 2.

Table 1. Students' Academic Performance Dataset Description

Independent Variables	Description	Type
Gender	Students Gender	Categorical(M/F)
NationalITy	Resident of a country	Categorical
PlaceofBirth	Country of origin	Categorical
StageID	Educational level student belongs	Categorical
GradeID	Grade student belongs	Categorical
SectionID	Classroom students belong	Categorical
Topic	Course Topic	Categorical
Semester	Semester year school	Categorical
Relation	Parent responsible for student	Categorical
Raisedhands	Number of times the student raises his/her hand on classroom	Continuous
VisITedResources	Number of times the student visits a course content	Continuous
AnnouncementsView	Number of times the student checks the new announcements	Continuous

Discussion	Number of times the student participates on discussion groups	Continuous
ParentAnsweringSurvey	Parents answered the surveys which are provided from school or not	Categorical
ParentschoolSatisfaction	Degree of parent satisfaction from school	Categorical
StudentAbsenceDays	The number of absence days for each student	Categorical

Table 2. Sample Student's Academic Performance Dataset Values

G	N	PB	SID	GID	SID	T	SE	R	RH	VR	AV	D	PAS	PS	SAD	Class
M	KW	KuwaIT	Lwl	G-04	A	IT	F	Father	15	16	2	20	Yes	Good	Under-7	M
M	KW	KuwaIT	Lwl	G-04	A	IT	F	Father	20	20	3	25	Yes	Good	Under-7	M
M	KW	KuwaIT	Lwl	G-04	A	IT	F	Father	10	7	0	30	No	Bad	Above-7	L
M	KW	KuwaIT	Lwl	G-04	A	IT	F	Father	30	25	5	35	No	Bad	Above-7	L
M	KW	KuwaIT	Lwl	G-04	A	IT	F	Father	40	50	12	50	No	Bad	Above-7	M
F	Jordan	Jordan	MS	G-08	A	Chemistry	S	Father	5	4	5	8	No	Bad	Above-7	L
F	Jordan	Jordan	MS	G-08	A	Geology	F	Father	50	77	14	28	No	Bad	Under-7	M
F	Jordan	Jordan	MS	G-08	A	Geology	S	Father	55	74	25	29	No	Bad	Under-7	M
F	Jordan	Jordan	MS	G-08	A	History	F	Father	30	17	14	57	No	Bad	Above-7	L
F	Jordan	Jordan	MS	G-08	A	History	S	Father	35	14	23	62	No	Bad	Above-7	L

Table 3. Sample Student's Academic Performance Dataset Values after Encoding

G	N	PB	SID	GID	SID	T	SE	R	RH	VR	AV	D	PAS	PS	SAD	Class
1	4	4	2	1	0	7	0	0	15	16	2	20	1	1	1	2
1	4	4	2	1	0	7	0	0	20	20	3	25	1	1	1	2
1	4	4	2	1	0	7	0	0	10	7	0	30	0	0	0	1

1	4	4	2	1	0	7	0	0	30	25	5	35	0	0	0	1
1	4	4	2	1	0	7	0	0	40	50	12	50	0	0	0	2
0	3	3	1	5	0	2	1	0	5	4	5	8	0	0	0	1
0	3	3	1	5	0	5	0	0	50	77	14	28	0	0	1	2
0	3	3	1	5	0	5	1	0	55	74	25	29	0	0	1	2
0	3	3	1	5	0	6	0	0	30	17	14	57	0	0	0	1
0	3	3	1	5	0	6	1	0	35	14	23	62	0	0	0	1

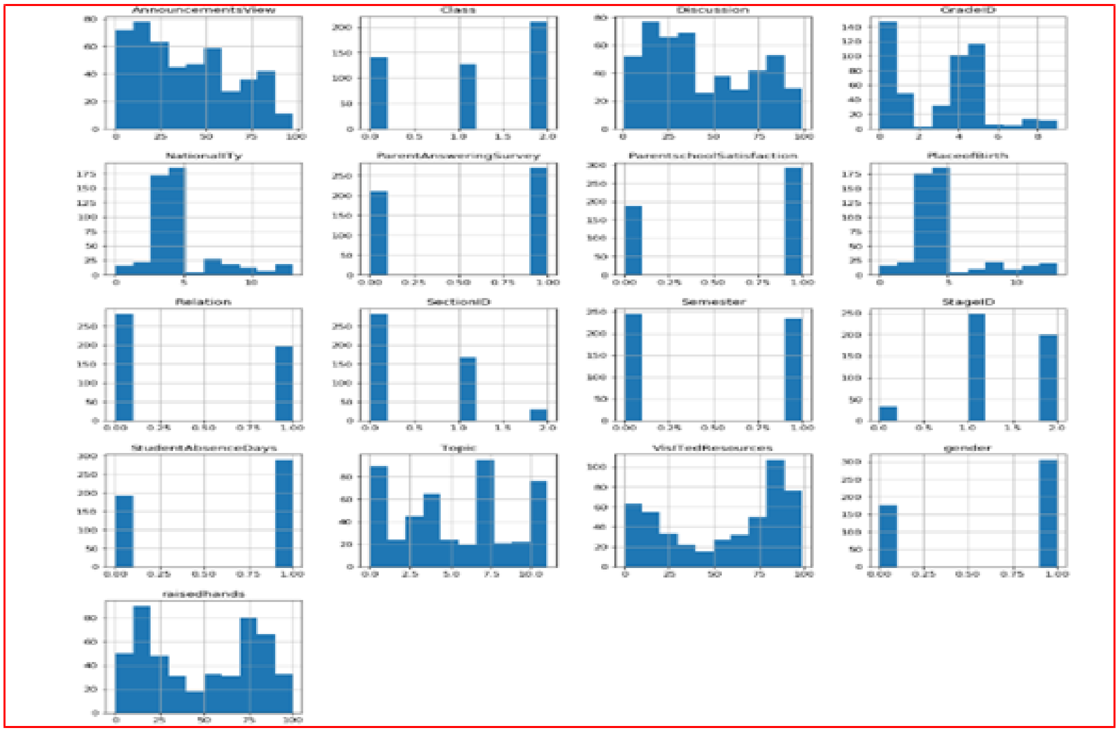

Figure 1. Histogram Distribution of Attributes

3. Experiment Results and Performance Analysis

The performance of a classification algorithm is analyzed based on confusion matrix which is used to describe the performance of a classification model on a set of test data for which the true values are known. The confusion matric values for decision tree, SVM, and Naïve Bayes are shown in Table 4. Machine learning algorithms performance metrics are calculated from the confusion matrix. The confusion matrix values are shown in Table 4. The performance metric values such as accuracy, precision, recall and f1 score are shown in Table 5. Decision tree provides 71% accuracy, Support Vector Machine provides 38% accuracy, and Naïve Bayes gives 77% accuracy. The Naïve Bayes algorithm gives better result. The comparison chart for the performance matric is shown in Figure 3.

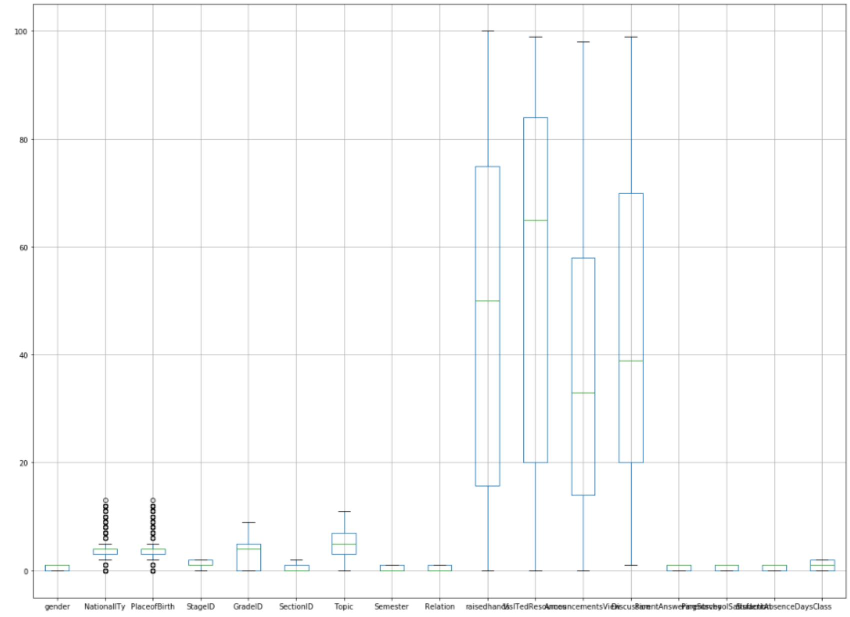

Figure 2. Box plot Distribution view of attributes

4. Conclusion and Future Work

Predicting student's academic performance is exceptionally useful to help the instructors and learners to improve their learning and teaching process schematically. This paper analyzed the student's academic performance with various machine learning algorithms. The classification algorithms are used frequently in educational data mining. In this paper, Decision tree, SVM and Navie Bayes algorithms are used to predict student's academic performance. The Naïve Bayes algorithm provide better performance for predicting the student's academic performance. In conclusion, student's academic dataset analysis on predicting student's academic performance has motivated us to carry out further research to be applied in our domain. It will help the educational system to track the student's academic performance in a structured way.

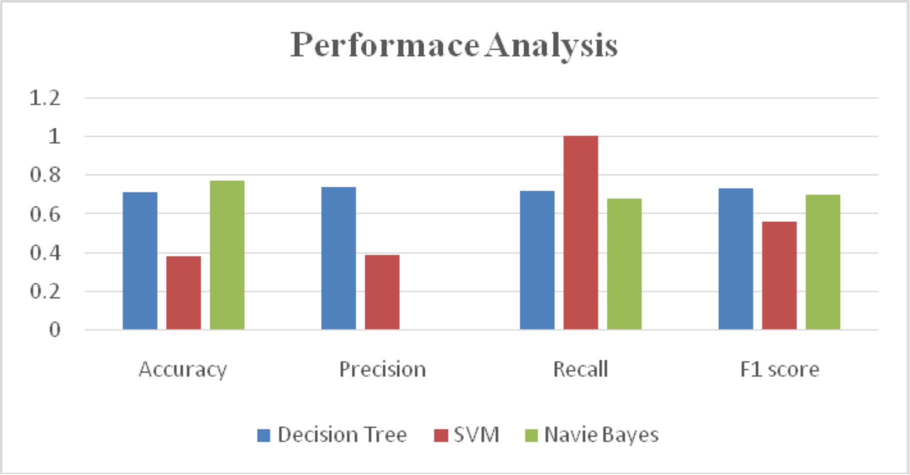

Figure 3. Comparison chart for performance Metric Values

Table 4. Confusion Matrix Values for the algorithms

Decision Tree			
Actual Value	Predicted Value		
	Low Level	Middle Level	High Level
Low Level	31	0	21
Middle Level	2	30	4
High Level	10	4	42
SVM			
Actual Value	Predicted Value		
	Low Level	Middle Level	High Level
Low Level	0	0	52
Middle Level	0	0	36
High Level	0	0	56
Navie Bayes			
Actual Value	Predicted Value		
	Low Level	Middle Level	High Level
Low Level	45	1	6
Middle Level	0	28	8
High Level	13	5	38

Table 5. Performance Metric Values for the algorithms

Algorithm	Accuracy	Precision	Recall	F1 score
Decision Tree	0.71	0.74	0.72	0.73
SVM	0.38	0.39	1.0	0.56
Navie Bayes	**0.77**	**0.73**	**0.68**	**0.70**

References

1. Nawal Ali Yassein, Rasha Gaffer M Helali and Somia B Mohomad , “Predicting Student Academic Performance in KSA using Data Mining Techniques”, Journal of Information Technology & Software Engineering., Vol.7, No. 5, (2017).
2. Md. Hedayetul Islam Shovon and Mahfuza Haque, “An Approach of Improving Student’s Academic Performance by using K-means clustering algorithm and Decision tree”, International Journal of Advanced Computer Science and Applications, Vol.3, No. 8, (2012).
3. Thaddeus Matundura Ogwoka, Wilson Cheruiyot, and George Okeyo, “A Model for Predicting Students’ Academic Performance using a Hybrid of K-means and Decision tree Algorithms”, International Journal of Computer Applications Technology and Research, Vol. 4, No. 9, (2015), pp.693 – 697.
4. Prashant Sahai Saxena and M. C. Govil, “Prediction of Student’s Academic Performance using Clustering”, National Conference on Cloud Computing & Big Data., (2015).
5. Nguyen Thai-Nghe, Andre Busche, and Lars Schmidt-Thieme, “Improving Academic Performance Prediction by Dealing with Class Imbalance”, International Conference on Intelligent Systems Design and Applications, IEEE Xplore., (2009)

6. Brijesh Kumar Baradwaj, and Saurabh Pal, “Mining educational data to analyze students' performance. International Journal of Advanced Computer Science and Applications Vol. 2, No. 6. (2011).
7. Carlos Márquez-Vera, Alberto Cano, Cristóbal Romero & Sebastián Ventura, “Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data”, Applied Intelligence, Vol.38, (2013), pp.315-330.
8. Francisco Araquea, Concepción Roldán, and Alberto Salgueroa, “Factors influencing university drop out rates”,Computers &Education,Vol. 53, No. 3, (2009), pp. 563-574.
9. O.J. Oyelade, O.O. Oladipupoand I.C. Obagbuwa, “Application of K-means clustering algorithm for prediction of student’s academic performance”, International Journal of Computer Science and Information Security, Vol.7, No.1, (2010), pp-292-295.