iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** III **Month of publication:** March 2020

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue III Mar 2020- Available at www.ijraset.com

Prediction on Performance of Slow Learners using Machine Learning

Ajay Kumar. C[1], Ajith. M. S[2], Nirmal Kumar. R[3], Geetha. R[4]
[1, 2, 3]*Student, Department of CSE, S.A. Engineering College, Thiruverkadu, Chennai, Tamil Nadu, India.*
[4]*Professor, Department of CSE, S.A. Engineering College, Thiruverkadu, Chennai, Tamil Nadu, India.*

Abstract: ***Maintaining of immense measure of data has always been a great concern. With expansion in awareness towards educational data, the amount of data in the educational institutes is additionally expanded. To deal with increasing growth of data leads to a new approach machine learning. Predicting student performance before the final examination can help management, faculty and as well as students to make timely decisions and avoid failing of students. With additional to this using sentimental analysis, which all the insights that will affect the students' performance can gained and focused on these insights to improve their performance on their next term. various machine learning algorithm techniques are used to build predictive models are XGboost, K-Nearest Neighbors (KNN), Support Vector Machine (SVM). Machine learning algorithm performance improved by precision, recall and F-1 Score. Algorithms were compared one another in terms of indicator values such as accuracy, to determine which algorithm gives best results.***
Keywords: K-Nearest Neighbors (KNN), Performance prediction, Support Vector Machine (SVM), XG Boost.

I. INTRODUCTION

Machine learning is a method of identifying the patterns in data and utilizing them automatically to make predictions. Machine learning can automatically learn from experience. The computer analyses a large amount of data, and discovers patterns and rules covered in the data. These patterns and rules are quite mathematical in nature, and they can be easily defined and processed by a computer. The computer can then utilizes those rules to meaningfully characterize new data. The creation of rules from data is an automatic process, and it is something that persistently improves with newly presented data. In this paper we predicting the student performance before the university exam. Prediction is done by using internal exam results. And also finding the student interest in course by taking Sentimental Analysis.

II. RELATED WORKS

In this paper [1] the authors did their project to predict the future carrier options and possibilities for students to get violent in future. Data collected from Educational Data Mining (EDM). The aim is to give the carrier options based on their interest, skills, links, hobbies etc. ID3 algorithm used for learning. The learned rules represented in the form of decision tree. K-Fold cross validation used for training & testing accuracy measure and low misclassification rate from the confusion matrix Numerous Decision Tree algorithm were used they are C4.5, CART, CHAFD, MARS. ID3 only handle categorical where as C4.5 handle both. ID3, C4.5, and CHAI were compared for accuracy. The advantage is along with the performance it also predicts the behavioural patterns of the students. The disadvantage is the algorithm (ID3) used here only used for categorical data. The classes were not explained clearly.

In this paper [2] the authors have used records of mathematical graduates' students from the academic year 2008 to 2014 as their dataset from the mathematics department in a college of science. Built a model to predict performance of students in a programming course based on their grades in English and Mathematics Subjects. Additional java code needed to convert and combine the data. They predict likelihood of success in a course before enrolling that course by taking two English and mathematics courses. They used Association rule algorithms. User manually fills the support and confidence threshold before doing prediction. It gives 62.75 accuracy of four subjects and 67.33 accuracy for only English subject. Advantage is, it predicts 9 out of 17 with 52.94%. Disadvantage is the dataset contains many irrelevant courses and multiple unnecessary details, and also additional java code is needed for translation and combining java code.

In this paper [3] the authors have taken dataset from the IT department at King Stand University. The dataset contains 100 students record. Each record has six parameters such as ID, GPA, HSC Percentage, GAT Score, EAT Score, and Courses taken by student. Out of 100 75 were used for train and remaining 25 were used for test data. The academic performance is predicted based on course difficulty, and EAT score. Classification model is based on the second year is more accurate. It gives 80 percentage of prediction accuracy.

In this paper [4] the authors did their prediction by gathering data from text-based self-evaluation comments written by students. They predicted dropout and failures using data based on student social behaviours such as student to student, student to teacher relationship. Developed a Dynamic diagnostic and self-regulated (DDS) to support decision making. Categorical approach is used to recognize the student's emotional states. Each comment is rated from 0 to 9 and categorized into three different categories positive, neutral, and negative. Support Vector Machine and Convolution Neural Network algorithms are preferred. It improves the early stage prediction accuracy. Early stage accuracy is improved. Disadvantage is it includes various methods which makes it complex for prediction.

In this paper [5] the authors focuses on Predicting student performance and also focuses on how the prediction algorithm can be used to identify the most importance attributes in student data. The dataset is created from Educational Datamining (EDM). Each record contains the following attributes GPA, Internal assessment, Internal assignment classified as-assignment mark, quizzes, lab work, class test and attendance, Student demographic (gender,age,family background) and external assessment. Decision tree, neural network, Navies Bayes-Nearest neighbour, Support vector machine algorithms were used against the dataset for prediction. At finally it produces an overall accuracy of all algorithms which are used for predicting the student performance. Disadvantage is the accuracy is not standard, It depends only on attributes, so can't be used to decide which algorithm is best.

III.METHODOLOGY

A. Data Description

The dataset which is used in this study was obtained from the CSE Department at S.A. Engineering college. The dataset contains 15 attributes and 151 student data. Internal exam results of three subjects are used for analysing the student academic performance. It also contains other behaviours of students such as travel time, Internet Usage, Late to Class. Each record has the following information.

1) Register Number.
2) Name.
3) Age.
4) Section.
5) Travel Time.
6) Internet usage.
7) Absent Days.
8) Late to Class.
9) Response.
10) Subject Score (IA1, IA2, IA3).
11) SA(Sentimental Analysis).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	RegNumbe	Name	Age	Section	TravelTim	InternetUsage	AbsentDays	LatetoClass	Response	IA1-CS400	IA1-CS4002	IA1-CS4003	IA1-Avg	IA2-CS4001	IA2-CS4002	IA2-CS4003	IA2-Avg	IA3-CS4001	IA3-CS400(	IA3-CS400	IA3-Avg	SA
2	CS1001	Aishwarya K S	21	C	3	2	8	Frequently	Fair	63	40	60	54	63	56	64	61	42	55	76	58	Performance is bad due to frequently missing and distraction
3	CS1002	Ajay S	21	A	2	2	9	Rare	Bad	51	73	57	60	52	54	45	50	32	50	53	45	Performance is good but need to improve learning habit to achieve higher grade
4	CS1003	Ajay Kumar C	21	A	3	3	5	Sometimes	Fair	65	47	42	51	57	72	40	56	53	55	69	59	Performance is good but need to improve learning habit to achieve higher grade
5	CS1004	Ajith.M.S	20	C	1	1	5	Rare	Fair	75	58	59	64	58	50	44	51	47	44	75	55	Performance is good but need to improve learning habit to achieve higher grade
6	CS1005	Akash S	21	C	1	3	6	Sometimes	Fair	57	57	57	57	52	50	62	55	32	42	75	50	Performance is bad due to frequently missing and distraction
7	CS1006	Akash J	20	C	3	2	8	Rare	Fair	61	50	61	57	72	61	46	60	61	52	61	58	Performance is bad due to frequently missing and distraction
8	CS1007	Akshya D	21	C	1	1	7	Sometimes	Bad	64	45	58	56	68	42	33	48	63	56	53	57	Performance is neutral need to put extra effort to achieve higher grade
9	CS1008	Amirtha P	21	B	1	2	6	Rare	Good	47	67	41	52	58	58	35	50	61	48	41	50	Performance is bad due to frequently missing and distraction
10	CS1009	Anbu S	20	C	1	1	8	Frequently	Good	31	67	42	47	46	61	42	50	43	56	77	59	Performance is good but need to improve learning habit to achieve higher grade
11	CS1010	Ashok K	21	B	1	1	10	Frequently	Bad	54	72	56	61	72	46	53	57	50	51	76	59	Performance is good but need to improve learning habit to achieve higher grade
12	CS1011	Ashok Kumar P	20	C	3	2	9	Sometimes	Bad	70	56	48	58	73	48	63	61	45	59	51	52	Performance is good but need to improve learning habit to achieve higher grade
13	CS1012	Ashu C	20	B	3	1	4	Frequently	Fair	42	75	52	56	66	46	53	55	64	48	62	58	Performance is neutral need to put extra effort to achieve higher grade
14	CS1013	Baasha M	21	B	1	2	4	Frequently	Fair	57	78	40	58	49	51	56	52	56	59	36	50	Performance is neutral need to put extra effort to achieve higher grade
15	CS1014	Balaji S	21	A	1	1	8	Rare	Bad	60	63	61	61	61	64	59	61	34	49	78	54	Performance is neutral need to put extra effort to achieve higher grade
16	CS1015	Balaji S	20	A	3	1	7	Sometimes	Good	54	79	51	61	64	67	54	62	44	55	46	48	Performance is good but need to improve learning habit to achieve higher grade
17	CS1016	Banumathi J	20	C	1	1	10	Sometimes	Good	32	64	68	55	48	59	61	56	56	59	64	60	Performance is good but need to improve learning habit to achieve higher grade
18	CS1017	Bhanu Prakash G	20	C	3	2	5	Frequently	Fair	72	56	40	56	49	54	61	55	30	43	45	39	Performance is good but need to improve learning habit to achieve higher grade
19	CS1018	Bhuvanesh K	21	A	1	1	7	Frequently	Fair	54	57	55	55	48	65	53	55	31	45	46	41	Performance is neutral need to put extra effort to achieve higher grade
20	CS1019	Bhuvaneshwari L	20	C	1	1	5	Frequently	Good	75	62	41	59	48	54	61	54	42	46	60	49	Performance is bad due to frequently missing and distraction
21	CS1020	Bhaasha J	21	A	1	3	10	Sometimes	Bad	69	58	67	65	57	61	62	60	36	57	48	47	Performance is good but need to improve learning habit to achieve higher grade
22	CS1021	Bujima K	21	A	1	2	6	Frequently	Fair	70	55	59	61	45	65	64	58	45	55	35	45	Performance is good but need to improve learning habit to achieve higher grade
23	CS1022	Candy V	21	A	2	3	8	Rare	Fair	75	42	67	61	61	60	64	62	41	60	64	55	Performance is good but need to improve learning habit to achieve higher grade
24	CS1023	Cassilas K	21	A	1	3	6	Rare	Bad	77	73	51	67	53	39	46	46	37	48	62	49	Performance is neutral need to put extra effort to achieve higher grade
25	CS1024	Chandar B	21	C	1	2	7	Frequently	Bad	51	42	51	48	65	59	51	58	47	45	71	54	Performance is good but need to improve learning habit to achieve higher grade
26	CS1025	Chelsea M	21	C	2	2	4	Frequently	Fair	73	49	68	63	51	66	61	59	50	45	45	47	Performance is neutral need to put extra effort to achieve higher grade
27	CS1026	Cipher N	20	B	3	3	7	Frequently	Bad	60	65	59	61	47	58	62	56	53	56	79	63	Performance is neutral need to put extra effort to achieve higher grade
28	CS1027	Cummins J	20	A	2	2	7	Sometimes	Fair	73	61	55	63	58	56	61	58	58	60	51	56	Performance is neutral need to put extra effort to achieve higher grade
29	CS1028	Daniel J	21	A	1	3	5	Rare	Good	76	68	56	67	49	42	45	45	59	50	39	49	Performance is neutral need to put extra effort to achieve higher grade
30	CS1029	Deepak L	20	A	2	2	8	Rare	Fair	71	80	55	69	59	69	51	60	44	45	56	48	Performance is neutral need to put extra effort to achieve higher grade
31	CS1030	Dembele O	21	A	3	3	6	Rare	Bad	51	52	58	54	29	39	54	41	40	57	63	53	Performance is good but need to improve learning habit to achieve higher grade
32	CS1031	Demitri K	20	A	1	1	7	Sometimes	Fair	56	44	57	52	50	58	47	52	53	53	58	55	Performance is neutral need to put extra effort to achieve higher grade
33	CS1032	Dhanasekhar V K	20	C	1	2	6	Frequently	Fair	42	77	54	58	39	61	56	52	54	54	59	56	Performance is good but need to improve learning habit to achieve higher grade
34	CS1033	Dhanilo P	21	B	1	2	5	Sometimes	Fair	43	64	69	59	26	57	50	44	42	51	47	47	Performance is good but need to improve learning habit to achieve higher grade
35	CS1034	Dhoni M S	20	B	1	1	6	Frequently	Fair	30	59	67	52	61	54	52	56	46	50	64	53	Performance is neutral need to put extra effort to achieve higher grade
36	CS1035	Eclessia R	20	C	1	2	5	Frequently	Fair	37	64	63	55	30	52	38	40	41	51	30	41	Performance is neutral need to put extra effort to achieve higher grade
37	CS1036	Ellakiya R	21	A	1	1	6	Rare	Bad	72	57	69	66	39	44	42	42	36	50	45	44	Performance is good but need to improve learning habit to achieve higher grade
38	CS1037	Ellakiya S	21	B	2	2	7	Rare	Fair	30	45	53	43	27	54	57	46	44	45	50	46	Performance is neutral need to put extra effort to achieve higher grade
39	CS1038	Emma	20	C	3	2	7	Frequently	Bad	74	84	53	70	35	53	53	47	62	51	44	52	Performance is good but need to improve learning habit to achieve higher grade
40	CS1039	Emeniem R	21	B	2	3	4	Rare	Bad	68	76	68	71	66	62	65	64	64	54	36	51	Performance is good but need to improve learning habit to achieve higher grade
41	CS1040	Emmanuel	20	B	1	1	4	Sometimes	Bad	79	30	60	56	58	47	49	51	57	49	45	50	Performance is bad due to frequently missing and distraction
42	CS1041	Femina S	20	A	1	1	6	Rare	Fair	71	32	69	57	50	50	60	53	44	50	60	51	Performance is bad due to frequently missing and distraction

Fig. 1.Sample of Original Data Set

B. Data Pre-Processing

Data Pre-processing is the process of converting data into understandable format. Raw data cannot be used directly for prediction. Only useful data is extracted for prediction.

Replacing missing values: The missing values in data occurs due to many reasons,such as during data collection. To handle missing data mean imputation method is used. It works by replacing the missing values by the mean/median values of that attribute in the class. In our case all data were filled.

Feature Selection: It is a process of reducing the inputs into relevant inputs for processing an analysis. To make predictive model some individual relevant inputs alone necessary that can be achieved through Feature Selection.

Feature Creation: Process of making new feature from existing data to obtain a machine learning model is known as Feature Creation.

C. Data Transformation

Data transformation is the process of reducing the number of values for a given continuous attribute, by dividing the attribute into a range of intervals. A sample of transformed data is shown in Figure 3. Certain rules are made to identify how well the students have performed in each test.

Students who scored marks between 50 to 60 corresponds to D, marks between 61 to 70 corresponds to C, marks between 71 to 80 corresponds to B, marks between 81 to 90 corresponds to A, marks greater than 90 corresponds to S and marks less than 50 corresponds to E.

Students who scored marks between 50 to 100 corresponds to pass, marks below 50 corresponds to fail.

IA3-CS4002	IA3-CS4003	IA3-Avg	pass_mark1	pass_mark2	pass_mark3	total_score	percentage	s
55	72	17	Pass	Pass	Pass	173	57.666667	
59	42	8	Pass	Pass	Fail	169	56.333333	
53	64	19	Pass	Pass	Pass	166	55.333333	
53	65	15	Pass	Pass	Pass	170	56.666667	
48	77	23	Pass	Pass	Pass	174	58.000000	

Fig. 2.Sample of transformed Data Set

D. Sentimental Analysis

Sentimental analysis is used to identify the behaviour of each student and gain insights which affecting their result in order to improve their education quality. Here we are using the report card comments to read the behaviour and education performance of the students. Sentimental Analysis will help group students and classify them based on remarks. That will help to analyze as group and leads what actions to be performed so that they will progress next semester.

	Name	SA
0	Aishwarya K S	positive
1	Ajay S	positive
2	Ajay Kumar C	negative
3	Ajith.M.S	positive
4	Akash S	negative
...	...	...
145	Vimal Kumar T	negative
146	Vimala G	negative
147	Wasim M	negative
148	Zack L	positive
149	Zamba O	negative

Fig. 3.Sample of Sentimental Analysis

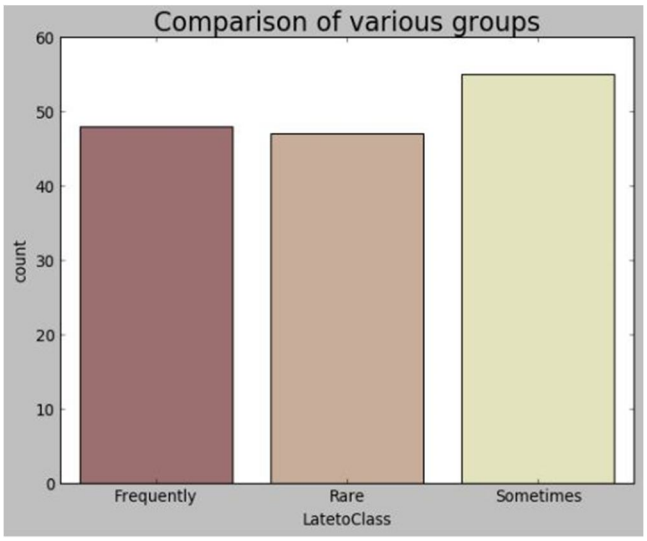


Fig. 4.Comparison of student behaviour

E. K-Nearest Neighbours

K-Nearest Neighbours(KNN) algorithm is simplest and yet strongest supervised learning model widely used for classification as well as for regression. KNN works by classifying the data points by separating into several classes inorder to predict the new data points by similarity measure(Euclidean distance) between the data points. In KNN algorithm, 'K' refers to number of neighbours for classification. Inorder to avoid the overfitting and underfitting of the model right choice of 'K' value need to be choose. In most of the cases taking 'K' = {square-root of (total number of data 'n')} gives good result. If the value 'K' considered as right else we make it odd either by adding or subtracting 1 from 'K' value.The accuracy of the K-Nearest Neighbours(KNN) algorithm is 82%.

F. Support Vector Machine

Support Vector Machine(SVM) are supervised learning model that allow analyse the data for classification analysis. SVM allows to classify the data by linearly separable such that it creates hyperplane to classify the among the data and based on best hyperplane(distance to the nearest data) the classification is performed. SVM works as robust, less affected by noisy data and low prone overfitting issues. SVM gives the accuracy of 80%.

G. XGBoost

It means extreme gradient boosting. XGBoost is a advanced version of the gradient boosting method. The main aim of this algorithm is to increase speed and to increase the efficiency.It is ensemble method that allows to correct the errors that made by already existing models until no further improvements can be achieved. New predictive models are created by using the residuals obtained from previously used models and then added together to perform the final prediction. It is called gradient boosting because it uses a gradient descent algorithm to minimize the loss when adding new models subsequently are added and hence provides good predictive model. XGBoost gives the more accurate results of 90%.

H. Performance Evaluation

To evaluate the performance obtained predictive models, three measures were used,such as accuracy, precision, and recall. Accuracy measures percentage of correctly classified records in the dataset. Precision measures the ratio of the true positives to all actual positives. Recall measures the ratios of the positives to all predicted positives.These measures are calculated as follows:

Accuracy = T P + T N /P + N

Precision = T P/T P + F P

Recall = T P /T P + F N

Where T(True) is the record collection labelled as Good. N(Negative) is the record collection labelled as Weak. TP(True Positive) is the number of records that were correctly classified as Good. TN(True Negative) is the number of records those correctly classified as Weak. FN(False Negative) is the number of records those misclassified as Weak. FP(FalsePositive) is the number of records those misclassified as Good.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue III Mar 2020- Available at www.ijraset.com

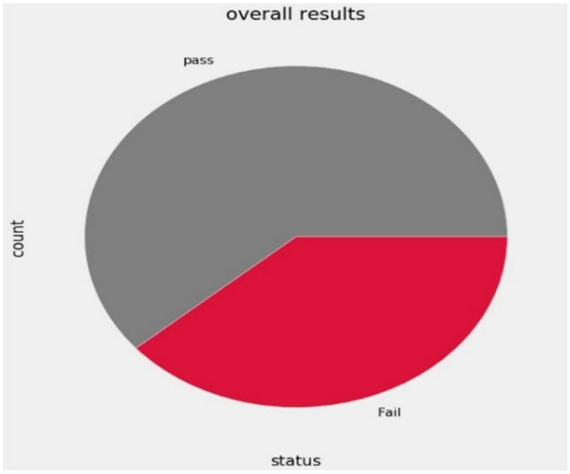


Fig. 5.Overall student results

	RegNumber	Name	status
0	CS1001	Aishwarya K S	0
1	CS1002	Ajay S	0
2	CS1003	Ajay Kumar C	1
3	CS1004	Ajith.M.S	0
4	CS1005	Akash S	0
...	...	...	...
145	CS1146	Vimal Kumar T	1
146	CS1147	Vimala G	1
147	CS1148	Wasim M	0
148	CS1149	Zack L	1
149	CS1150	Zamba O	1

Fig.6 Result of predicted student Performance

TABLE I
Accuracy Comparison

Algorithm	Accuracy
Support Vector Machine	0.80
K-Nearest Neighbor	0.82
XGBoost	0.90

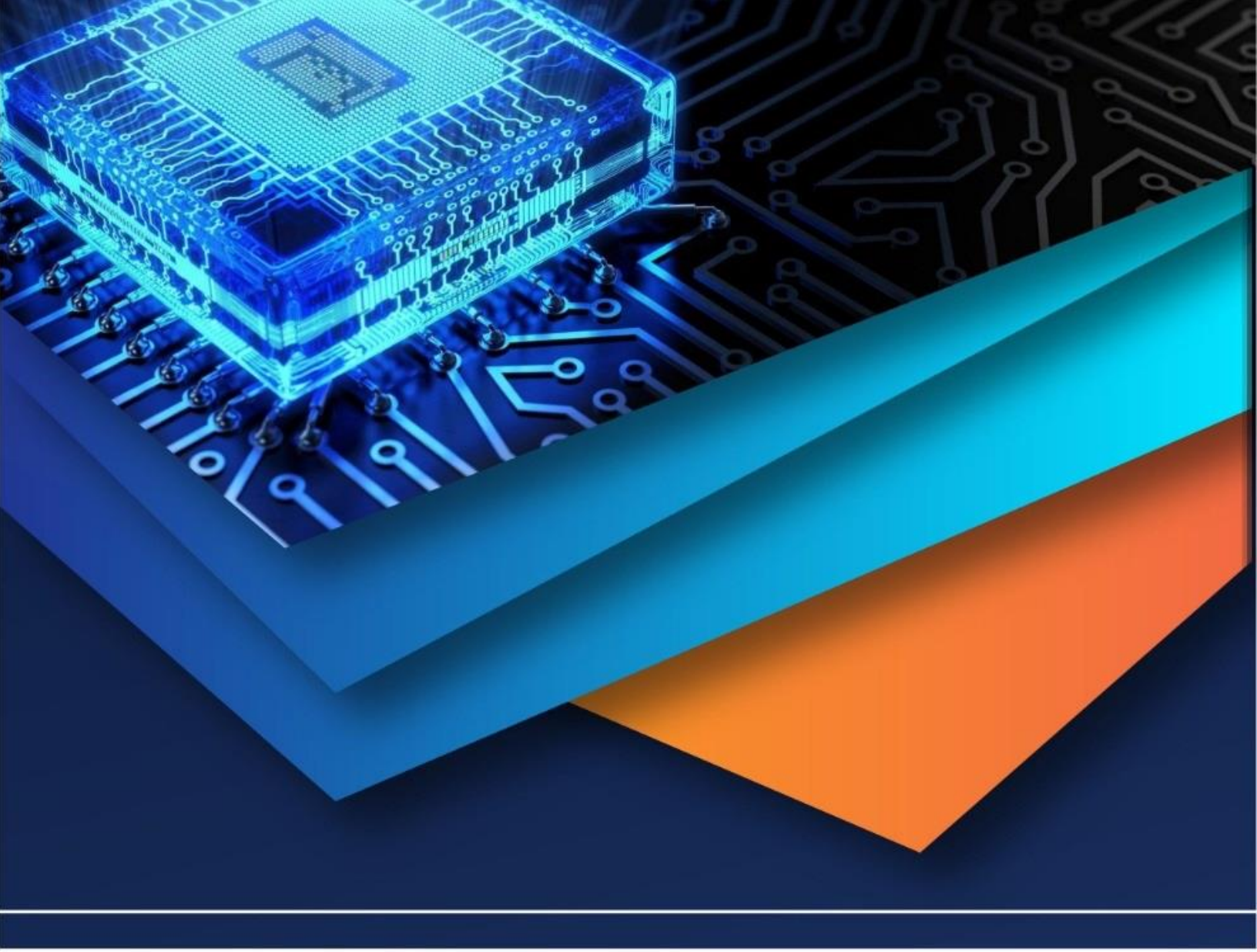
By comparing Support Vector Machine, K-Nearest Neighbor, XGBoost algorithms for evaluating accuracy, XGBoost algorithm gives the best predictive result such as 90%.

IV. CONCLUSION

The purpose is to accurately identify students those who are at risk before they take a final (semester) exam. In (fig.6) the status indicates student result as follow that 0(fail) and 1(pass). These students might fail, drop or perform worse than expected. We can notify the student's instructor to take the appropriate steps to assist that particular student before take final (semester) exam. The students' performance evaluation is done by based on academic and personal data collected from college's progress report. The dataset was used to perform prediction using KNN, SVM, XGBoost classification algorithms and accuracy is compared. Based on the accuracy comparison table (Fig. 7) one may conclude that the XGBoost Classification method was the most suitable algorithm for the dataset to predict and gives best result. The dataset may be extended to collect some of other insights that will effect student performance for improvement of student performance. Based on the prediction one may define what kinds results may expected for every students who shares the same characteristics. We can further improve this performance by using more characteristics in the prediction techniques and can implement in Chat Bot that gives automatic recommendation after predicting.

REFERENCES

[1] Elakia, Gayathri, Aarthi, and Naren, "Application of Data Mining in Educational Database for Predicting Behavioural Patterns of the Students.", IJCSIT Vol.5 No.3, June 2014.
[2] Ghada Badr, Afnan Algobail, Hanadi Almutairi, Manal Almutery, "Predicting Student Performance in University Courses: A Case Study and Tool in KSU Mathematics Department", SDMA Vol.82 80-89, March 2016.
[3] Yasmeen Altujjar, Wejdan Altamimi, Isra Al-Turaiki, Muna Al-Razgan, "Predicting Critical Courses Affecting Students Performance: A Case Study", SDMA Vol.82 65-71, March 2016.
[4] L.C. Yu, C.W. Lee, H.I. Pan, C.Y. Chou, P.Y. Chao, Z.H. Chen, S.F. Tseng, C.L. Chan, and K.R. Lai, "Improving Early Prediction of Academic Failure using Sentiment Analysis on Self-Evaluated Comments", JCAL Vol.34 No.4, March 2018.
[5] Amirah Mohamed Shahiri, Wahidad Husain, Nur aini Abdul and Rashid, "A Review on Predicting Students Performance Using Datamining Techniques", ISICO Vol.72, 2015.
[6] Ajman Abu Saa, "Educational Datamining & Students Performance Prediction", IJACSA Vol.7 No.5, March 2016.
[7] Kolluru Venkata Nagendra, k.Sreenivas, P.Radhika, "Student Performance Prediction Using Different Classification Algorithms", IJCESR Vol.5 No.4, 2018.
[8] Mack Sweeney, Huzefa Rangwala, Jaime Lester, Aditya Johri, "Next- Term Student Performance Prediction: A Recommender System Approach", JEDM Vol.8 No.1, September 2016.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY