

A Comprehensive Review of Student Performance Prediction using Machine Learning Techniques

Shruti Mehta¹ and Dr. Mukta Agarwal²

¹⁻²Indus University, Ahmedabad, Gujarat, India

Email: shrutimehta.21.rs@indusuni.ac.in, muktaagarwal.dcs@indusuni.ac.in

Abstract—Education plays a crucial role in the lives of individuals, shaping their purpose and adding value to their existence. This paper presented student performance prediction models by applying KNN, SVM and Naive Bayes on educational dataset of school taken from Kaggle. The main objective of this paper is to provide early estimation of the students' performance so that teachers can plan their pedagogy to improve the student learning. Most of the prior research conducted on the same dataset involved comparing KNN and SVM and they have employed a restricted set of attributes from this dataset. We have introduced an additional machine learning technique to extend the comparative analysis of results and we have utilized all available attributes. The experimental results show that SVM is performing better than KNN and Naive Bayes.

Index Terms— Educational Dataset, Support Vector Machine, K-Nearest Neighbour, Naive Bayes.

I. INTRODUCTION

Predicting student performance is a crucial concern for enhancing the educational process. Several factors, including the father's occupation, gender, and previous years' average marks, may influence the student's academic performance. [4] Early prediction of student performance can contribute significantly to the improvement of the educational process. In this research paper, we employed K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Naive Bayes algorithms to predict the categorization of students as Below Average, Average, or Above Average. This classification is intended to facilitate the planning of special assignments and pedagogical strategies to enhance their learning experience.

II. LITERATURE REVIEW

[1] They have conducted a comparison of the outcomes yielded by two algorithms, revealing high correlation coefficients of 0.96 and 0.95, respectively. They have applied Support Vector Machine (SVM) regression and Instance-Based Learning with k-Nearest Neighbors (IBK) classifiers to the student dataset for the prediction of student grades within the range of 0 to 20. It is noteworthy that their research primarily focused on predicting grades with three categories (Bad, Good, Excellent), and the reported accuracies ranged from 52% to 82.3%. From their results, it can be inferred that both SVM and KNN regression are suitable for addressing this type of problem. [5]

[2] Through a comparative analysis of three evaluation parameters (accuracy, recall, and precision) for both

KNN and Naive Bayes, the Naive Bayes algorithm demonstrated the highest accuracy at 93.17%. They have concentrated on K-Nearest Neighbors (KNN) and Naive Bayes as primary classification algorithms to devise a student performance prediction model. This significant accuracy suggests a robust relationship among the features influencing student performance, thereby enhancing the model's predictive capabilities for students' future performance.

[3] They have used various supervised learning techniques, including Decision Trees (DT), Random Forests (RF), Extra Trees Classifier (ETC), and Gradient Boosting Classifier (GBC), were applied to the dataset and evaluated using a 10-fold cross-validation method. Data was sourced from the publicly available Kaggle repository, specifically curated for research purposes. This study employed a combination of machine learning and statistical methods to explore the field of study selected by students. Subsequently, the data were preprocessed to a format suitable for machine learning models. The results indicated that Random Forests (RF) and Gradient Boosting Classifier (GBC) exhibited accuracies of 0.75% and 0.61%, respectively, in predicting students' chosen study fields.

[4] Logistic regression was employed for the prediction of dropout among first-year students.

[5] Applied different algorithms on different courses. SVM is selected as the most suitable algorithm for the computer department, yielding an F-measure of 0.73. The QDA algorithm has been identified as the optimal machine learning algorithm for the Urban department, achieving the highest F-measure of 0.91. RF is designated as the best machine learning algorithm for the Civil department, achieving the highest F-measure of 0.79. Similarly, the RF algorithm has been chosen as the most effective machine learning algorithm for the Mechanical department, exhibiting the highest F-measure of 0.78. For the Architectural department, KNN is the selected algorithm with an F-measure of 0.89, and for the Electrical department, KNN is again chosen with an F-measure of 0.77. LR is the algorithm of choice for the Mining and Petroleum department, attaining an F-measure of 0.91. Lastly, the average F-measure across all departments is calculated to be 82.57%

[6] The current findings indicate that the Grade Point Average (GPA) holds significant importance in determining a student's final performance. Upon comparing algorithms in terms of accuracy, it becomes evident that K-nearest neighbor surpasses Decision Tree in accuracy. Most students fall into the Pass category, with only 19% achieving excellence and 31% falling into the good category. Notably, only 6% of students are projected to fail based on their GPA, and this figure can be reduced to zero through proactive measures. Predicting results in advance allows the opportunity to enhance the performance of a percentage of students, either moving them from good to excellent or from excellent to good.

III. MATERIAL AND METHODOLOGY

A. Data Collection

The dataset was obtained from the Kaggle dataset repository, comprising 28 attributes and 395 rows. The average of G1 and G2 was computed, and the outcomes were categorized into three groups: Below average (≤ 12), Average (> 12 and ≤ 16), and Above average (> 16 and ≤ 20). This categorization was performed with the intention of tailoring pedagogical strategies to students' comprehension levels. Assignments can be assigned separately to enhance their learning experiences.

B. Data Conversion

The dataset has been transformed into a numeric format, facilitating the ease of calculating statistical values and correlations. The average of Grade 1 and Grade 2 has been computed, and the results have been categorized into Below Average, Average, and Above Average.

The review of existing literature reveals that researchers have conducted comparisons between KNN and SVM, as well as between SVM and Naive Bayes. Consequently, our decision is to perform a comprehensive comparison of all three algorithms on a similar database. Notably, previous studies utilized a limited set of features from the provided database and reported performance solely in terms of Pass and Fail outcomes.

C. Techniques

K-Nearest Neighbour

[2], [6] This algorithm computes the nearest neighbors based on a specified k-value, indicating the number of nearest neighbors to be considered. Consequently, this determines the class assignment for the sample data point. The fundamental concept of K-nearest neighbors (KNN) lies in categorizing an unknown data point based on the class of its nearest neighbors, whose classes are already known.

TABLE I. ATTRIBUTE CONVERSION

SNo	Attribute	Description	Conversion
1	Sex	Student' gender	M=0, F=1
2	Address	Urban or Rural	U=0, R=1
3	Famsize	Family size	GT3=0, LE3=2
4	Pstatus	Parent Living apart or together	A=0,T=1
5	Medu	Mother's education	none=0,primary education(4th grade)=1, 5th to 9th grade=2, secondary education=3,higher education=4
6	Fedu	Father's education	none=0,primary education(4th grade)=1, 5th to 9th grade=2, secondary education=3,higher education=4
7	Mjob	Mother Education	services=0, health=1,teacher=2,at_home=3,other=4
8	Fjob	Father Education	services=0, health=1,teacher=2,at_home=3,other=4
9	reason	Reason to choose the school	course=0,home=1,reputation=2,other=3
10	guardian	Mother, Father or Other	mother=0,father=1,other=3
11	traveltime	Travelling time to school	<15 min=1,15 to 30 min =2,30 min to 1 hour=3 and >1 hour=4
12	studytime	Time spends by the students on studies	<2 hours=1,2 to 5 hours=2, 5 to 10 hours=3,>10 hours=4
13	failure	No of past class failures	0,1,2,3,4
14	schoolsup	Extra Educational support	0,1
15	famsup	Extra family support	0,1
16	paid	Extra paid classes	0,1
17	activities	Extra-Curricular activities	0,1
18	nursery	Attended pre school	0,1
19	higher	Attended higher education	0,1
20	internet	Access of internet at home	0,1
21	romantic	Romantic or Practical	0,1
22	famrel	Family Relation	1,2,3,4,5
23	freetime	Free time after school	1,2,3,4,5
24	goout	Going out with friends	1,2,3,4,5
25	Dalc	Workday alcohol consumption	1,2,3,4,5
26	Walc	Weekend alcohol consumption	1,2,3,4,5
27	Health	Current health status	1,2,3,4,5
28	Absence	No. Of school absence	0 to 93
29	Avg(G1,G2)	Average of Grade 1 and Grade 2	0 to 20
30	Res	Performance of student	Below Average=0, Average=1, Above average=2

KNN stands out as one of the simplest machine learning algorithms, finding widespread applications across various domains. Its versatility is evident in applications such as pattern recognition, internet marketing, image database analysis, and clustering.

This straightforward definition encapsulates the essence of KNN. The classification report and confusion matrix of the KNN machine learning technique after applying on the student database taken from Kaggle are as follows:

TABLE II. CLASSIFICATION REPORT FOR KNN MACHINE LEARNING MODEL

	precision	recall	F1-score	support
0	0.93	0.98	0.96	65
1	0.86	0.83	0.85	30
2	1.00	0.25	0.40	4
accuracy			0.91	99
mac avg	0.93	0.69	0.73	99
weight avg	0.91	0.91	0.90	99

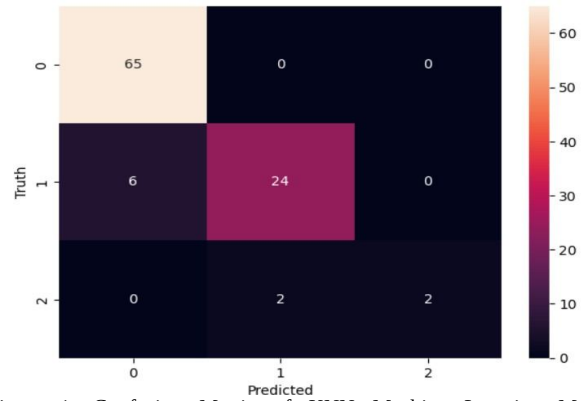


Figure 1. Confusion Matrix of KNN Machine Learning Model

The highest correlation is achieved when K-NN value is increased to 6 with a correlation value of 0.87. Furthermore, we notice 6KNN achieves the maximum level of accuracy with 91%. Parameter values achieved by different KNN values are given in table III.

TABLE III. EVALUATION MATRIX FOR KNN MACHINE LEARNING MODEL

KNN	Accuracy	Correlation Coefficient	Mean Square error	Root Mean Square Error	Relative Absolute Error	Relative Root Mean Square Error
3	89%	0.82	0.11	0.33	0.22	0.51
5	90%	0.85	0.09	0.30	0.18	0.46
6	91%	0.87	0.08	0.28	0.16	0.48

Naïve Bayes

Naive Bayes is recognized for its user-friendly nature, suitability for analysing and predicting high-dimensional data and its compatibility with large datasets. So, we have employed the Naive Bayes (NB) on a similar database in comparison with KNN and SVM. Operating on the principles of the Bayesian theorem, this algorithm computes the probability for each class based on multiple independent input variables.

There are three types of Naive Bayes available i.e. Gaussian Naive Bayes (used for normal distribution), Multinomial Naive Bayes (suitable for discrete data) and Bernoulli Naive Bayes (used for document classification). We have utilized Gaussian Naive Bayes technique. Naive Bayes has demonstrated robust implementation across diverse scenarios, including spam classification, weather forecasting systems, and sentiment analysis. The correlation coefficient is 0.80 and accuracy achieved by the Naive Bayes is 77%.

The classification report of the Naïve Bayes machine learning technique is presented in Table 4. The confusion matrix is presented in Fig 2.

TABLE IV. CLASSIFICATION REPORT FOR NAÏVE BAYES MACHINE LEARNING MODEL

	precision	recall	F1-score	support
0	0.94	0.91	0.92	65
1	0.68	0.43	0.53	30
2	0.24	1.00	0.38	4
accuracy			0.77	99
mac avg	0.62	0.78	0.61	99
weight avg	0.83	0.77	0.78	99

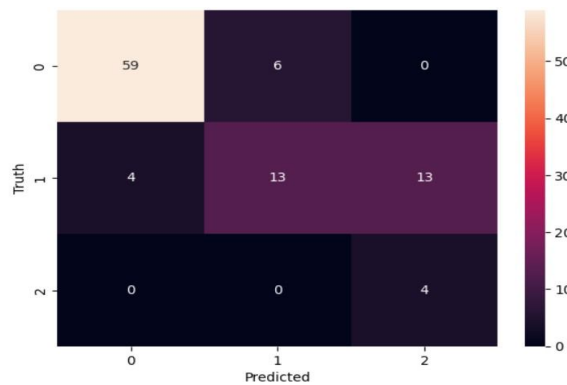


Figure 2. Confusion Matrix of Naïve Bayes Machine Learning Model

The highest correlation is achieved when partition of training and testing data is kept 75:25. Furthermore, we notice Naïve Bayes achieves accuracy of 77%. Correlation Coefficient for different partition values shown in Table 5.

TABLE V. EVALUATION MATRIX FOR NAÏVE BAYES MACHINE LEARNING MODEL

	Accuracy	Correlation Coefficient	Mean Square error	Root Mean Square Error	Relative Absolute Error
75:25	77%	0.80	0.23	0.48	0.46
60:40	72%	0.76	0.32	0.56	0.53
90:10	85%	0.79	0.15	0.38	0.53

Support Vector Machine

The core objective of the SVM algorithm is to create an optimal line or decision boundary capable of segregating an n-dimensional space into distinct classes. This delineation enables the accurate categorization of new data points in future instances. The optimal decision boundary is commonly known as a hyperplane.

The Support Vector Machine (SVM) stands out as one of the most widely utilized algorithms in Supervised Learning, demonstrating proficiency in addressing both Classification and Regression challenges. However, its primary application lies in tackling Classification problems within the realm of Machine Learning.

SVM identifies crucial points or vectors essential for establishing the hyperplane. These pivotal instances are termed support vectors, thereby leading to the nomenclature "Support Vector Machine."

We have optimized the parameter for SVM which includes gamma, kernel type, cost, degree and loss. Accuracy gets decreased by increasing the value of gamma to 1 and improvement occurs only when kernel type is linear. The correlation coefficient is 0.91 and accuracy achieved by the SVN is 94%.

The classification report and confusion matrix of the SVM machine learning technique after applying on the student database taken from Kaggle is as follows:

TABLE VI. CLASSIFICATION REPORT FOR SVM MACHINE LEARNING MODEL

	precision	recall	F1-score	support
0	0.97	0.97	0.97	65
1	0.90	0.93	0.92	30
2	1.00	0.75	0.86	4
accuracy			0.95	99
mac avg	0.96	0.88	0.91	99
weight avg	0.95	0.95	0.95	99

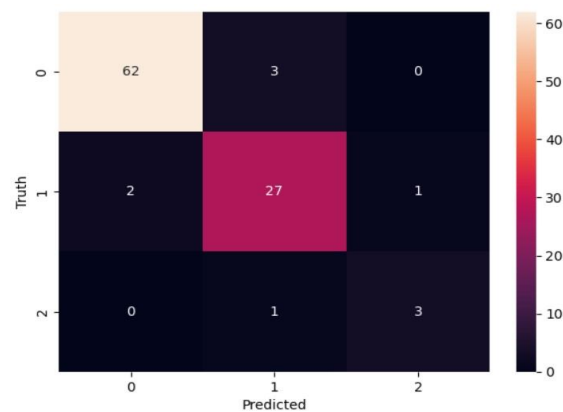


Figure. 3. Confusion Matrix of SVM Machine Learning Model

Correlation Coefficient for different partition values for SVM is shown in Table 7.

TABLE VII. EVALUATION MATRIX FOR SVM MACHINE LEARNING MODEL

	Accuracy	Correlation Coefficient	Mean Square error	Root Mean Square Error	Relative Absolute Error
75:25	95%	0.91	0.05	0.22	0.10
60:40	90%	0.86	0.10	0.31	0.45
90:10	92%	0.79	0.07	0.27	0.54

IV. RESULT AND DISCUSSION

During our experiment, it became evident that SVM demonstrates superior performance compared to KNN and Naive Bayes when utilizing 28 features, with the testing dataset comprising 75% and the training dataset making up 25%. Final comparison is demonstrated in Table VIII.

TABLE VIII. COMPARISON MATRIX FOR SVM, KNN AND NAÏVE BAYES MACHINE LEARNING MODEL

	Accuracy	Correlation Coefficient	Mean Square error
Accuracy	77%	92%	95%
Correlation Coefficient	0.80	0.87	0.91
Mean Square error	0.23	0.08	0.05
Root Mean Square Error	0.48	0.28	0.10
Relative Absolute Error	0.46	0.16	0.10

V. CONCLUSIONS

In the comparison of the three techniques, SVM exhibits superior performance. In future, the incorporation of additional techniques can be explored which has the potential to enhance result accuracy further. Additionally, fine-tuning the categorization of target values may play a crucial role in achieving the most optimal outcomes for improving student performance.

REFERENCES

- [1] Al-Shehri Huda, Student performance prediction using Support Vector Machine and K-Nearest Neighbor. 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), 2017. Accessed: Dec. 25, 2023. [Online]. Available: <https://ieeexplore.ieee.org/document/7946847>
- [2] I. A. A. Amra and A. Y. A. Maghari, "Students performance prediction using KNN and Naïve Bayesian," ICIT 2017 - 8th International Conference on Information Technology, Proceedings, pp. 909–913, Oct. 2017, doi: 10.1109/ICITECH.2017.8079967.
- [3] A. O. Alsayed et al., "Selection of the right undergraduate major by students using supervised learning techniques," Applied Sciences (Switzerland), vol. 11, no. 22, Nov. 2021, doi: 10.3390/app112210639.
- [4] M. Segura, J. Mello, and A. Hernández, "Machine Learning Prediction of University Student Dropout: Does Preference Play a Key Role?," Mathematics, vol. 10, no. 18, Sep. 2022, doi: 10.3390/math10183359.
- [5] M. Ezz and A. Elshenawy, "Adaptive recommendation system using machine learning algorithms for predicting student's best academic program," Educ Inf Technol (Dordr), vol. 25, no. 4, pp. 2733–2746, Jul. 2020, doi: 10.1007/S10639-019-10049-7/METRICS.
- [6] Srivastava Akhilesh Kumar, "PREDICTION OF STUDENTS PERFORMANCE USING KNN AND DECISION TREE- A MACHINE LEARNING APPROACH," Strad Research, vol. 7, no. 9, Sep. 2020, doi: 10.37896/sr7.9/018.