Forecasting Students' Academic Performance using Support Vector Machine

N. Gowri Sreelakshmi[1], J. Sai Mukesh[2], G. Srinivasa Rao[3], M. Sai Youzen[4], M. Karthik[5] , K. Nithin[6]
Department of CSE, GIET- A, Rajamahendravaram
Srilakshmi@giet.ac.in[1],
{saimukeshjoga6[2], gamapasanisrinivas[3], saiyouzen2001[4], karthikmylapalli21[5], nithinkolli951[6] } @gmail.com

Abstract
Student success forecasting is a goal of higher education institutions and a significant area of study. Both students and educational institutions are genuinely concerned with predicting students' academic achievement. Academic advancement and personality traits related to learning activities are only a few variables that might affect pupils' academic performance. Common traits of students like age, gender, how much they study, how often they miss classes, the size of their families, and so on… are the training dataset's primary constituents for the supervised machine learning algorithm to determine whether the student will be successful on their final exam or not. This study investigated the effectiveness of various supervised algorithms in machine learning. Previous studies on this dataset relied heavily on the K-Nearest Neighbor and Logistic Regression algorithms, with mixed results. Surprisingly, the Support Vector Machine algorithm is rarely used for predictions, even though it's a popular and reliable technique. To make sure we had a fair comparison, we decided to use Support Vector Machines to predict the students' grades and see how they compared to the other methods.
Keywords:- Students' Academic Performance Prediction, Machine Learning Techniques, K-Nearest Neighbors(KNN) classifier, Logistic Regression(LR) classifier, Support Vector Machine (SVM) classifier, Quality Metrics.

1. Introduction

Estimating student performance is critical for educators because it allows them to gather feedback early on and take quick action or put in protections if needed to help the student do better. This forecast is manageable if the cause of the problem is identified. If it's from doing after-school activities, having family problems, or health issues, all that stuff can really mess up a student's grades. We can investigate such scenarios with the use of a dataset for student performance. We already said that by looking at the records from previous students, we can create a forecasting model that will help students do better in their tests. We'll try out different classifiers such as KNN or SVM and compare them how they stack up with each other. A variety of elements, such as familial issues or alcohol use, can impact a student's exam performance. By employing our machine learning model

we can modify the lectures and curriculum, and lecturers and institutional managers which can support students' learning plans by drawing on the prediction of learning outcomes. The results indicated that the SVM classifier was the most reliable and effective among the three classifiers for predicting students' final grades when the supervised machine learning approaches we have discussed were examined. The machine learning algorithms for comparing and predicting whether a student would pass the final exam/ not are supervised which machines are being trained by using "well labelled" data.

2. Literature Survey:

Earlier models made use of ML algorithms including Random Forest, Naive Bayes Decision Tree, Linear Regression, and Bagging [1][2][3][4]. The previous system uses the machine learning algorithm, which is a decision tree and random forest, to evaluate student performance, and the decision tree is used to take a group of data values from the dataset [1]. The machine learning algorithm is linear regression, which was used to produce the evaluation of student performance but provided less accuracy because of underfitting and overfitting issues [2]. The outcomes of the Empirical study proved that the Support Vector Machine had a slight edge over the K-Nearest Neighbor and Logistic Regression algorithms. [3]. Regression algorithms are slightly better than Neural networks [4]. Boosting student retention rates, streamlining enrolment processes, keeping alumni connected, better-targeted marketing, and all-around improvement of the educational institute's efficacy all benefited from the students' performance projection [5].

3. Proposed Methodology:

The proposed system incorporates algorithms such as K-Nearest Neighbor (KNN) classifier, Support Vector Machine (SVM) classifier, and Logistic Regression classifier that are used to categorize features that aid students' academic performance as well as the prediction of final exam marks, and it compares the above three algorithms based on quality indicators such as receiver operating characteristic (ROC) curve, F1 score, Root Mean Square Error (RMSE) and confusion matrix. The working of proposed model is represented in **Fig 5**

3.1 Machine Learning (ML) Algorithms:

ML is described as the subfield of AI that uses specific programs and data sets to help computers gain knowledge from the data they process and become more proficient over time. Supervised, Unsupervised, and Reinforcement are the 3 types of Machine Learning models. And now in this system, we are going to use Supervised learning models which make use of labeled datasets to train the machine.

3.2 Logistic Regression (LR):

This is a ML classifier which is employed to examine association between different variables and identify patterns. It employs supervised learning to make guesses

about what the outcome of a certain variable will be based on a set of independent factors. A discrete or binary value between 0 and 1 is generated using the sigmoid function in Logistic Regression.

3.3 K-Nearest Neighbor (KNN):

KNN which can be also known as a non-parametric algorithm used to estimate the correlation between the new case of data and existing data. Whenever the new input of data enters the model, it will predict the value and put it into the category which is familiar to the categories by using the Euclidean distance formula.

3.4 Support Vector Machine (SVM):

SVM is a famed model for doing supervised learning, which is employed to solve regression issues, classification, and outlier detection. Support Vector Machine is used to resolve linear and non-linear problems. The concept behind SVM is really straightforward - it makes a hyperplane that divides the data into different categories. In high-dimensional spaces, SVM is more effective. It transforms non-linearly separable data from lower dimensions to higher dimensions to facilitate linear classification by using a kernel.

3.5 Feature Selection:

Feature selection is a method that is used to handle large metric sets, to identify which metrics contribute to the estimation of students' grades in final exam. By using this feature selection, redundant and non-independent attributes are removed from the dataset.

3.6 Performance Evaluation:

Here we will use four quality metrics to measure the performance of 3 classifiers These quality metrics are:

- Confusion Matrix (**Fig 1**)
- F1 Score (**Fig 2**)
- ROC (Receiver Operating Characteristic) Curve & Score (**Fig 3**)
- Root Mean Square Error(**Fig 4**)

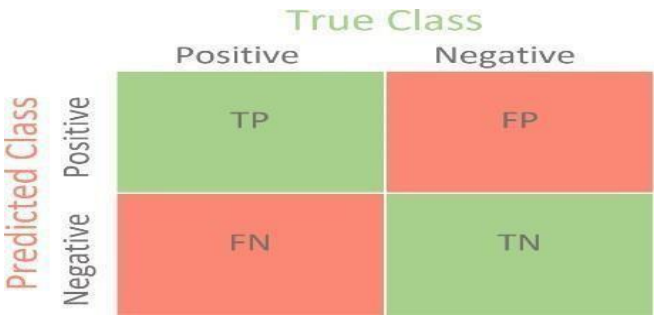

Fig- 1: Confusion Matrix

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{\text{TP}}{\text{TP} + \frac{1}{2}(\text{FP} + \text{FN})}$$

Fig- 2: F1 Score

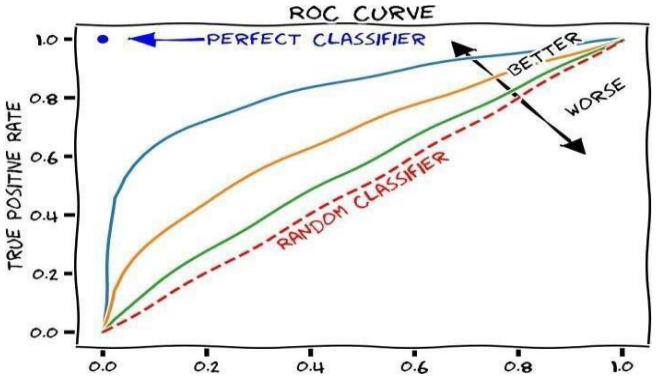

Fig-3: ROC Curve

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

Fig- 4: Root Mean Square Error

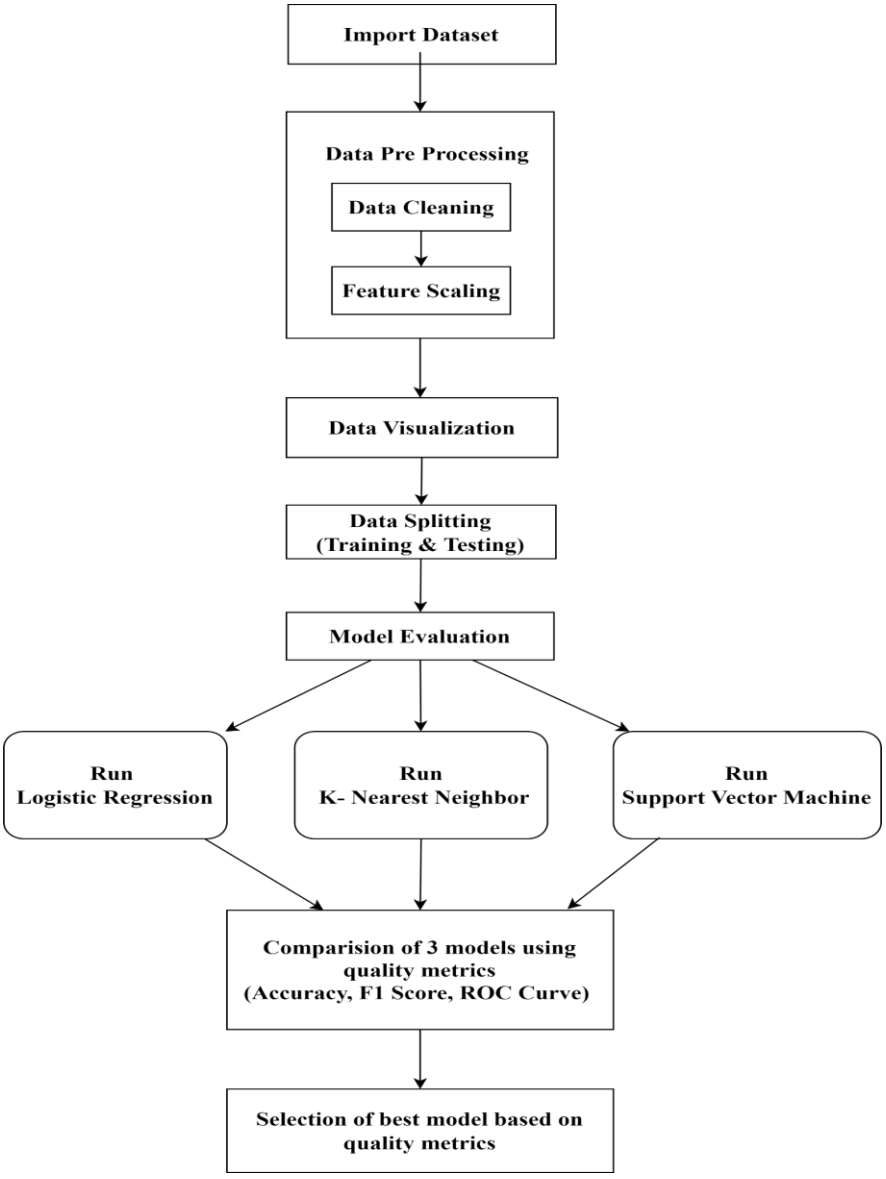

Fig-5: Working of the model

4. Result

The result of this paper depicts the findings of a model evaluation. It found that the SVM (Support Vector Machine) model had the lowest RMSE score (**Fig 7**) and highest ROC score (**Fig 6**) among the three models, suggesting that it was the most accurate in forecasting the students' academic performance. The outcomes of the model can be used to recognize trends and data patterns, which can inform future educational policy and practice. To communicate the results, it is important to create visualizations, tables, and reports that highlight the key findings and the implications for education.

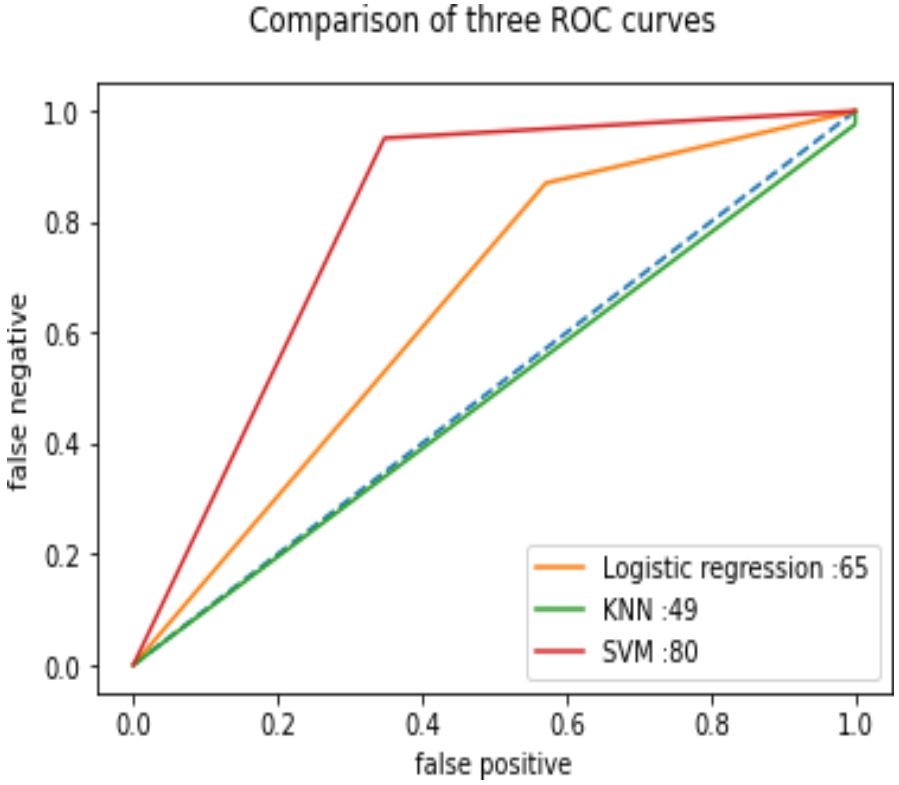

Fig- 6: Comparision of ROC Curves

metric	Logistic regression	KNN	SVM
Mean Square Error	0.535	0.558	0.395

Fig- 7: Comparison of Mean Square Error

5. Conclusion

Figuring out which classification algorithm works best and determining which factors have the greatest impact on students' academic standing so they can get a clear picture of what it takes to succeed in school and stay away from trouble. In conclusion, the results of the comparison between the SVM, logistic regression, and KNN models for forecasting students' academic performance showed that the SVM model outperformed the other two models in terms of multiple quality metrics, including F1 score, accuracy, and ROC. The SVM model had the highest F1 score, accuracy, and ROC AUC value (**Fig 8**), indicating that it was the most effective model in correctly classifying students as pass or fail. The results of this comparison emphasize the significance of considering multiple quality metrics when evaluating model's performance. The SVM model's superior performance in all four metrics suggests that it is the best choice for this particular task of forecasting students' academic performance.

In summary, the use of SVM models in this area has the potential to improve the accuracy of predictions and provide valuable insights into the aspects that effects students' exam score. By making the use of these ML techniques, it is possible to make informed decisions and take targeted actions to improve students' outcomes.

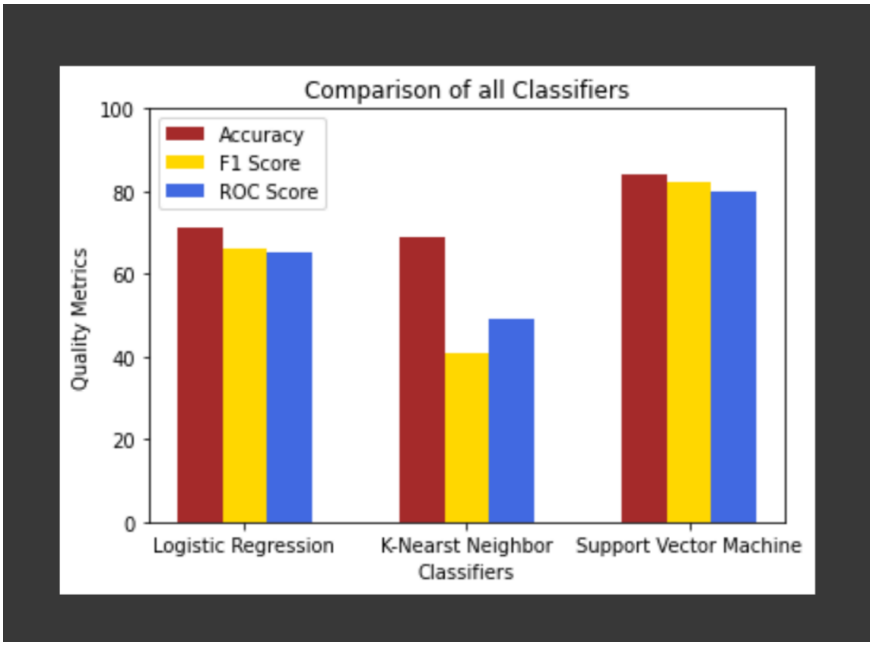

Fig- 8: Comparison of all Classifiers

6. References

1. "W. Tomohisa, M. Dawodi, N. Ahmadi, and M. Mohammadi (2019). A comparison of supervised learning algorithms for predicting student performance. The International Conference on Artificial Intelligence in Information and Communication (ICAIIC) will be held in 2019. (pp. 124-127). IEEE, doi:10.1109/ICAIIC.2019.8669085.
2. M. Ayan, M. Garca (2008). academic performance of university students predicted using linear and logistic models. 11(1), 275-288 The Spanish Journal of Psychology. https://doi.org/10.1017/S1138741600004315
3. Iqbal, Z., Qadir, J., and Qayyum, A., along with Latif, S. (2019). An empirical investigation of early student grade prediction. The 2019 Second International Conference on Advances in Computational Sciences (ICACS) will be held at (pp. 1-7). The IEEE. https://doi.org/10.23919/ICACS.2019.8689136
4. N. M. Rusli, Z. Ibrahim, and R. M. Janor (2008). Comparison of neuro-fuzzy, logistic regression, and artificial neural networks for predicting students' academic success. International Information Technology Symposium in 2008 (pp. 1-6). IEEE, doi:10.1109/ITSIM.2008.4631535.
5. A systematic review of the literature on machine learning techniques for predicting student performance. Science and education 2021; 11(9):552 by Albreiki B, Zaki N, and Alashwal H. https://doi.org/10.3390/educsci11090552
6. "Student academic performance prediction model employing fuzzy genetic algorithm and decision tree." by Simi Indurable, Hamsa, Hashmi, and Joyful J. Kizhakkethottam Procedia Technology 25, 326-332 (2016)
7. V. Vijayalakshmi and K. Venkatachalapathy "Comparison of predicting student performance using machine learning algorithms,". International Journal of Intelligent Systems and Applications, Vol. 11, No. 12 (2019), p. 34.
8. "Predicting students' academic performance using education data mining," by Borkar, Suchita, and K. Rajeswari. 2.7 (2013): 273-279 in International journal of computer science and mobile computing.
9. Suthaharan, Shan, and Shan Suthaharan "Support vector machine, in Machine learning models and methods for big data classification: reasoning with examples for successful learning", 2016, pp. 207–235.
10. "Predicting student performance using personalised analytics," Asmaa Elbadrawy et al. Computer 49.4 (2016): 61-69.
11. Osman Yildiz et al "Rules Optimization Based Fuzzy Model for Predicting Distance Education Students' Grades," 4.1 (2014): 59-62 in International Journal of Information and Educational Technology.

12. "Comparative Study Of Machine Learning Knn, Svm, And Decision Tree Algorithm To Predict Student's Performance", by S. Wiyono and T. Abidin, Int. J. Res. Granthaalayah, vol. 7, no. 1, pp. 190-196, Jan. 2019.
13. PREPRINT (Version 1) of Student Performance Prediction Using Machine Learning Techniques, 23 March 2022 by Ahmed Omar and Tarek Abd El-Hafez. Research Square [https://doi.org/10.21203/rs.3.rs-1455610/v1].