iJRASET
International Journal For Research in Applied Science and Engineering Technology



# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 13 **Issue:** V **Month of publication:** May 2025

**DOI:** https://doi.org/10.22214/ijraset.2025.71502

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: 7.538*
*Volume 13 Issue V May 2025- Available at www.ijraset.com*

# Educational Data Classification using Different Classifiers for Real Time Student Applications

Dr. Abhilasha Dangi
*Assistant Professor, Department of Computer Science Engineering, Sangam University, Bhilwara, Rajasthan*

***Abstract:*** ***The application of data mining techniques to educational datasets is gaining increasing attention due to the growing availability of student-related information. However, organizing and interpreting this data effectively poses a significant challenge because of its high dimensional and complexity. This study explores the use of the Linear Support Vector Classifier (Linear - SVC), SVM, and naive bias known for its computational efficiency and robustness, in categorizing educational data. The model's output can reveal actionable insights into student performance, offering valuable support for real-time academic assessments. Additionally, it holds potential for informing future strategies related to student admissions and selection processes in higher education institutions.***

***Index Terms: Education data Classification, Linear - Svc Classifie, naive bayes, SVM, Data Mining.***

## I. INTRODUCTION

The continuous accumulation of student data over time presents both opportunities and challenges in the education sector. Efficiently organizing this growing volume of information to meet analytical and decision-making needs is vital for enhancing educational outcomes. However, manually analyzing and classifying such datasets is both resource-intensive and impractical at scale. To address this, researchers have increasingly turned to automated classification methods, which have gained traction in the field of educational data mining. Techniques such as Support Vector Machines (SVM), Naive Bayes Decision Trees, Neural Networks, and Linear Support Vector Classifiers (Linear-SVC) have been widely explored for this purpose. Among them, Linear-SVC has emerged as a preferred method in text classification tasks, valued for its simplicity and speed in both model training and prediction [4]. Although it may not always achieve the highest accuracy compared to more complex models, its effectiveness in diverse applications has been well demonstrated. The model's ability to evaluate each feature independently contributes to its computational efficiency.

This paper investigates the use of the Linear-SVC algorithm for classifying student-related data and benchmarks its classification performance against other widely used machine learning models like SVM.

The section II specifies the background methodology of the Linear SVC followed the problem description and analysis of section III. The papers concludes with the proposed result and analysis on data set collected.

## II. LITERATURE SURVEY

Educational data mining (EDM) is a field that exploits machine- literacy, statistical and data- mining algorithms over the different types of educational data [1]. Its main ideal is to analyses data in order to resolve educational exploration issues. EDM is concerned with developing styles to explore the unique types of data in educational settings and, using these styles, to more understand scholars and the settings in which they learn. This data helps to understand that data uprooted could be used to Find Meaningful Pattern for the scholars on the real time problem script operation to be covered at council position. Also the model can be used for the unborn planning of pupil selection criteria at council position [2].

The ideal of anticipation is to estimate the unknown value of a variable that describes the students. anticipation of performance, knowledge, score, or mark is done. This value can be numerical/ nonstop value (regression task) or categorical/ separate value (bracket task) [5]. regression analysis finds the relationship between a dependent variable and one or further independent variables. In bracket individual particulars are placed into groups grounded on quantitative information regarding one or further characteristics essential in the particulars and grounded on a training set of preliminary labeled particulars.Data data mining tools for educational exploration issues are prominently developed and used in numerous countries. In country like India, demand for educational data mining has been increased from last many times, because of increase in educational database time to time and need of discovery of knowledge from that database to take important opinions and remedial results for unborn purpose. numerous marketable data mining software packages are available with colorful situations of complication and cost.

©IJRASET: All Rights are Reserved 

One of the popular data mining tools is linear SVC [3]. Since this is open source tool and supports numerous data mining algorithms, this has been used for pre-processing and bracket of pupil data. Python has been used to classify the scholars using Linear SVC algorithms.

## III. SELECTIVE LINEAR SVC CLASSIFIER AND SVM

Support vector machines represent a category of algorithms known for their effectiveness in multi-class classification tasks. Similar to other machine learning techniques, they assign new instances to pre-labeled classes. Support Vector Machine (SVM)
is a powerful machine learning method that's mainly used for classifying data into different categories [1] . It works by finding the best possible dividing line or in higher dimensions, a plane—that separates the data into groups. What makes SVM special is that it tries to create this dividing line with the largest possible gap (or margin) between the groups, which helps make more accurate predictions on new, unseen data.

When the data isn't neatly separable in a straight line, SVM can use something called a kernel function to transform the data into a different space where separation becomes easier. Think of it as reshaping the space so the computer can draw a better boundary.

Some key ideas behind SVM:

*1)* Support vectors are the data points that are closest to the boundary—it uses these to decide where the line goes.
*2)* The margin is the space between the groups; the bigger it is, the better the model tends to perform.
*3)* The kernel trick is like a mathematical shortcut that helps SVM deal with complex, curved boundaries without making the problem too hard to solve [2].
*4)* SVM is especially good at handling tricky problems like sorting text, recognizing images, or analyzing medical data. It's popular because it's both accurate and efficient, especially when the number of features (or variables) is high [4].

The underlying mechanisms for each classification algorithm differ, yet strategies for improving performance may be analogous across similar types. This type of algorithm offers various kernel functions, including RBF (Radial Basis Function), linear, polynomial, Gaussian, and sigmoid kernels. The linear kernel is most effective for linear problems, while alternate kernels are utilized for non-linear scenarios.

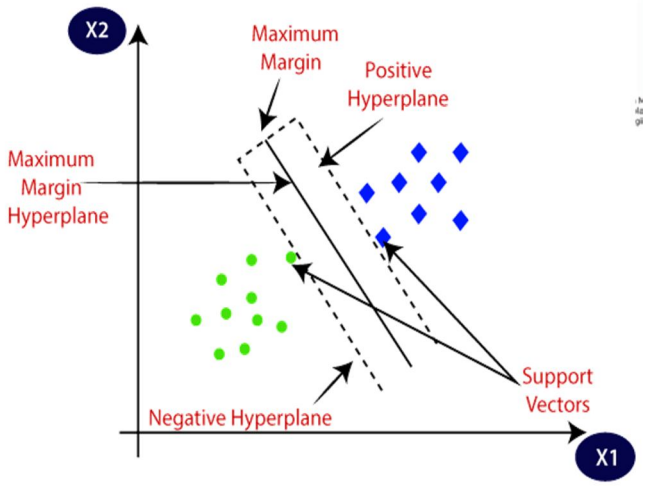



Fig. 1. Hyperplane generation

## IV. PROBLEM DEFINITION

The University of Rajasthan's courses are the primary means of promoting the state of Rajasthan's educational system. There are about 10,000 students registered on the portal for the university with its established colleges. Due to the SC, ST, and OBC reservation quota, there is fierce competition among the state universities under UOR for the cutoff list each year. When sub-castes are established for the reservation quota classes, the issue occurs. Therefore, the different data mining problem is to classify the sub-caste into the given set of quota classes. The Centre of Converging Technologies (CCT), an independent organization supported by AICTE/UGC and located on the University of Jaipur campus, made available 152 data points for the data set. The entries also included the the applicant's personal information for the dual B.Tech/M.Tech programs at CCT. 42 categorical/nominal attributes, such as father/mother invasion, social position, sub-caste, 10th–12th percentage, and form of media of education, were included in data set (fig. [2]).

©IJRASET: All Rights are Reserved 

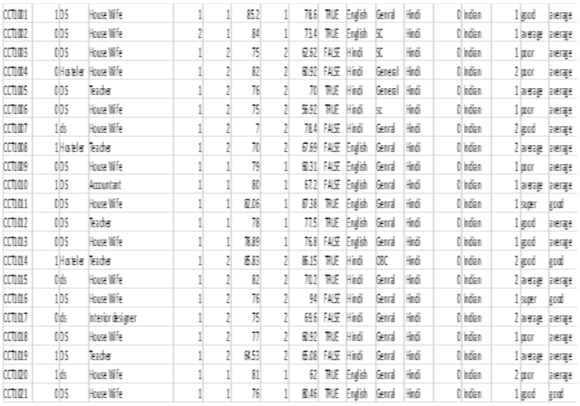


Fig. 2. Snapshot of the CCT Sample data-set

Although the data can be categorized using any of the categorical attributes, the caste and sub-caste combination is one of the primary attributes used to place the specified sub-caste in the appropriate reservation quota class. In order to demonstrate the effectiveness of the Indian government's reservation quota policies for promoting the education of the backward classes, the percentage is also another classifier. To test the data, the data set is first normalized using eight nominal and one categorical attributes selected from the designated set of forty-two attributes fig. 3.

| column1 | column2 | column3 | column4 | column5 | column6 | column7 | column8 | column12 |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | average |
| 0 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | average |
| 0 | 1 | 2 | 2 | 1 | 2 | 2 | 2 | average |
| 0 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | average |
| 0 | 1 | 2 | 2 | 1 | 2 | 1 | 2 | average |
| 0 | 1 | 2 | 2 | 1 | 2 | 1 | 2 | average |
| 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | average |
| 1 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | average |
| 0 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | average |
| 1 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | average |
| 0 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | good |
| 0 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | average |
| 0 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | average |
| 1 | 1 | 2 | 2 | 2 | 2 | 1 | 2 | good |
| 0 | 1 | 2 | 2 | 2 | 2 | 1 | 2 | average |
| 1 | 1 | 2 | 2 | 1 | 2 | 2 | 2 | good |
| 0 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | average |
| 0 | 1 | 2 | 2 | 1 | 2 | 1 | 2 | average |
| 1 | 1 | 2 | 2 | 1 | 2 | 2 | 2 | average |
| 1 | 1 | 1 | 1 | 2 | 2 | 1 | 1 | average |
| 0 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | good |
| 1 | 1 | 1 | 1 | 2 | 2 | 1 | 1 | average |
| 1 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | average |

Fig. 3. Snapshot of the Normalized data-set

Using the normalized data were classified set, the using the 10th–12th percentile, with the reservation category as the classifier criterion. Depending on the classifier definition, the categorical output can be categorized as either "good" or "average." Selective Linear SVC classifiers with a 66% split percentage are used for training cases and 33% for testing. Python is used to run the classifier, and the dataset is supplied in CSV format.

## V. SIMULATION & INFERENCE

The basic concept behind the SVM technique is to make a plane known as the hyper-plane that would best separate the data-point as classes if conceivable. The selected features for T1 are T2, T3 & Stime. These features were used as training attributes for the predictive analysis of T1. For the insight of the model, the Accuracy, Precision, Recall & F1 scores are calculated. The results are summarized in Table below.

Table 1. Accuracy, Precision, Recall & F1 scores for T1 for Liner-SVC

| Folds | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| 1 | 0.80384 | 0.80362 | 0.80384 | 0.79761 |
| 2 | 0.80384 | 0.80786 | 0.80384 | 0.79458 |
| 3 | 0.85769 | 0.85439 | 0.85769 | 0.85142 |
| 4 | 0.79615 | 0.82156 | 0.79615 | 0.80341 |
| 5 | 0.81744 | 0.82897 | 0.81744 | 0.81002 |

The research work deduces that for the given dataset out of five-folds, the best Accuracy, Precision, Recall & F1 scores were achieved on 3rd fold similar to Decision Tree algorithm, but with different scores, which are 0.85769, 0.85439, 0.85769 & 0.85142 respectively. The graph given below gives the variation of Accuracy, Precision, Recall & F1 scores with the number of folds.

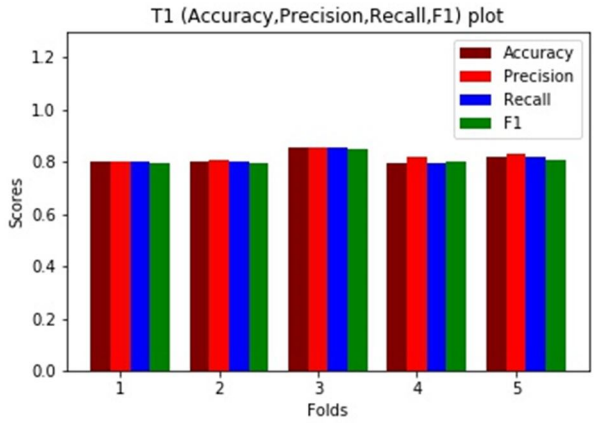



Fig. 4. Variation of Accuracy, Precision, Recall & F1 scores with a number of folds for T1.

The best machine learning technique for the issue is linear SVC. T2, T3, and Stime are the features chosen for T1. The predictive assessment of T1 has made use of these features as training attribute values. Accuracy, precision, recall, and F1 ratings are computed for the model's vision. Table 1 below provides a summary of the effects:

Table 2. Accuracy, Precision, Recall & F1 scores for T1for SVM

| Folds | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| 1 | 0.73384 | 0.57611 | 0.70384 | 0.63429 |
| 2 | 0.88846 | 0.75716 | 0.78846 | 0.76810 |
| 3 | 0.74076 | 0.68146 | 0.78076 | 0.70993 |
| 4 | 0.84153 | 0.81923 | 0.79153 | 0.82142 |
| 5 | 0.80193 | 0.78298 | 0.82193 | 0.78641 |

According to the table, the fourth fold of these five folds—0.84153, 0.81923, 0.79153, and 0.82142, respectively—has achieved the first-rate Accuracy, Precision, Recall, and F1 scores for the data-set in question. The graph below shows how Accuracy, Precision, Recall, and F1 scores vary over a wide range of folds.

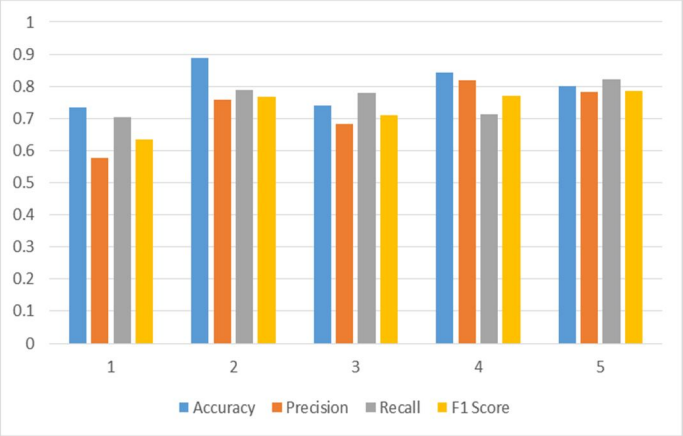



Fig. 5. Variation of Accuracy, Precision, Recall & F1 scores with a number of folds for T1.

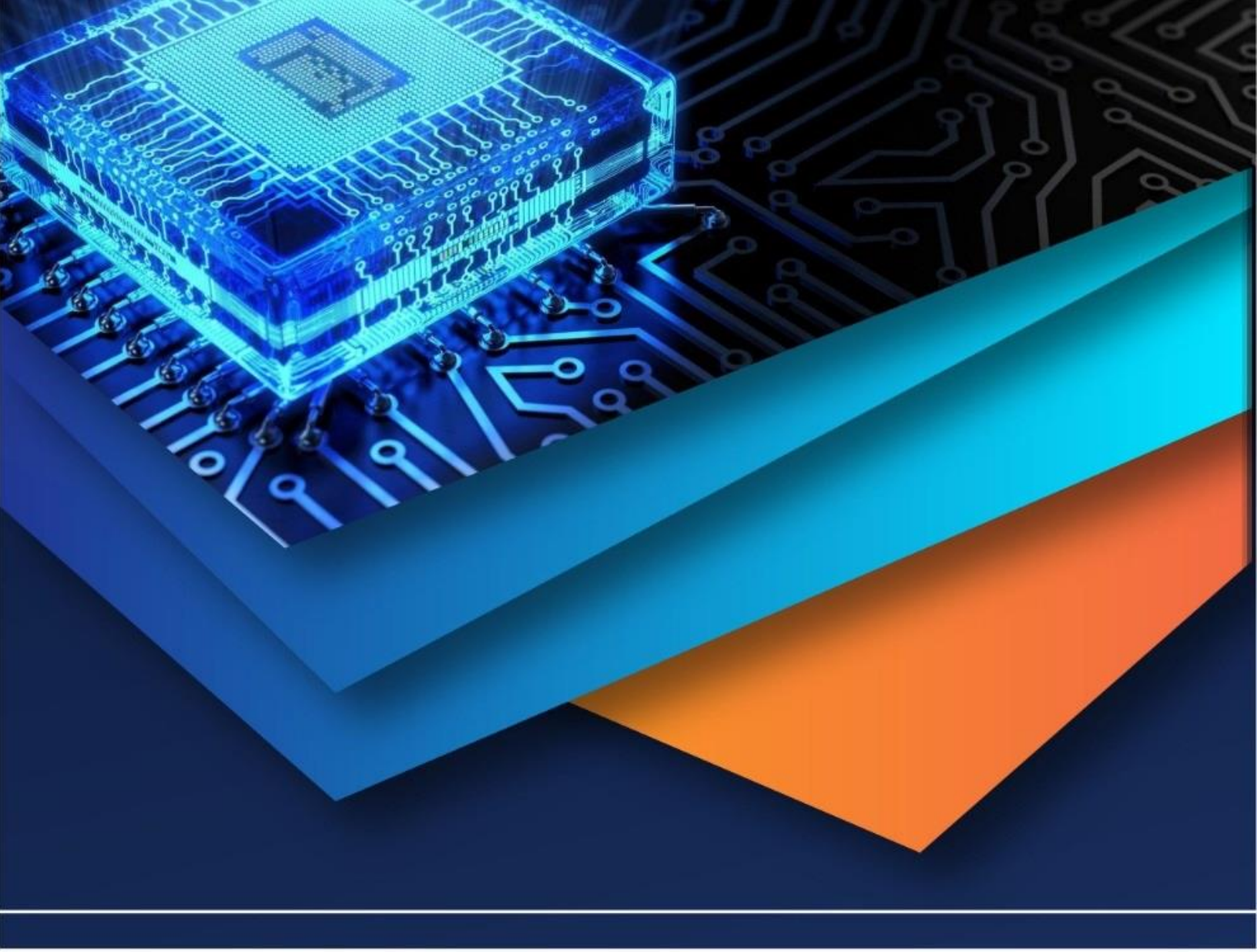
## VI. CONCLUSION

The analysis of the results suggests that the linear SVC algorithm be seen as the the classification medium for the small data set. However, by employing different classifiers, such as Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), etc., in less time, the percentage classified can still be increased. In the future, it is suggested that the same process be used, with performance comparisons with other classifiers for larger data sets planned. It is also possible to use mass estimation for similarity measures when processing data in parallel.

## REFERENCES

[1] Dangi, S. Srivastava, Multi-Class Sentiment Analysis Comparison Using Support Vector Machine (SVM) and BAGGING Technique-An Ensemble Method, In July 2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE) pp. 1-6
[2] A. Dangi, S. Srivastava,, An application of student data to forecast education results of student by using classification techniques, In July 2020 Journal of Critical Reviews (JCR), SCOPUS pp. 3339-3343
[3] G. A. Verma and M. Kumari, "Prediction of Students' Performance Using Machine Learning Techniques," in Proc. 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2020, pp. 689–693.
[4] R. Ahmad, S. A. Malik, and M. Hussain, "Machine Learning Techniques for Predicting Academic Performance: A Case Study of a Pakistani University," in Proc. 2021 4th International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), Sukkur, Pakistan, 2021, pp. 1–6
[5] Ghosh and S. Dey, "Predicting Students' Academic Performance Using Linear SVC and SMOTE for Imbalanced Dataset," in Proc. 2021 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2021, pp. 1–6.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)