CHAPTER 6

PREDICTION OF STUDENT PERFORMANCE

6.1 Introduction

This chapter discusses the design of student performance prediction models using different classifier models. We evaluate their performance, based on predictive accuracy and other measures. We also investigate the performance of the classifiers based on class labels.

The purpose of this study is to examine the student performance prediction models that can be used to classify students based on their results. Different ML techniques are applied to the student datasets to predict students' performance. The results are evaluated by measuring the accuracy of all the models and compared with each other. The six classifier models will also be used to build the student performance prediction models, under three different cases, by varying class variable values. In the first case, the target variable is divided into two labels: *fail* and *pass*. In the second case, the target variable is classified into three labels: *fail, good* and *very good*; and in the third case, five class labels: *fail, poor, satisfactory, very good* and *excellent* are used. Results of six algorithms will be evaluated in order to predict the students' performance based on the collected data.

Data was obtained from different universities and institutions in Jaipur, and 1980 student records were used for the experimental work. This chapter discusses the results and analysis of the four datasets evaluated from the confusion matrix, discussed in chapter 4. The metrics used for this study are the same, i.e., accuracy, precision, sensitivity, specificity, f-measure and ROC curve.

6.2 Categorization of Data

The data was collected from different sources. The original data includes the attributes Gender, Father and mother education, occupation, Family income, Board, 10^{th} and 12^{th} marks and Result in Program. Three categories of data were considered, as given below:

1. Categorize the students with two class labels, represented by 1 and 2: "fail" students who obtained less than 35% of marks, "pass" for passed students who got minimum 35% marks.
2. Categorize them into three class labels, represented by 1, 2 and 3: "fail" representing failed students who secured below 40%, "good" representing

passed students ranges from 40% to 70% and “very good” for passed students with 70% and above.

3. Categorize into five class labels, represented by the numbers 1 to 5: “Excellent” represent students who got 85% and above, “very good” for students who obtained marks ranges from 70% to 85%, “satisfactory” for students who secured between 55% to 70% , “poor” for students who secured between 40 and 50% and “fail” for uncovered cases.

The class values distribution and the number of classes in each category are shown in Figure 6.1. All the experiments were done by using Jupyter Notebook that facilitates all classifier techniques.

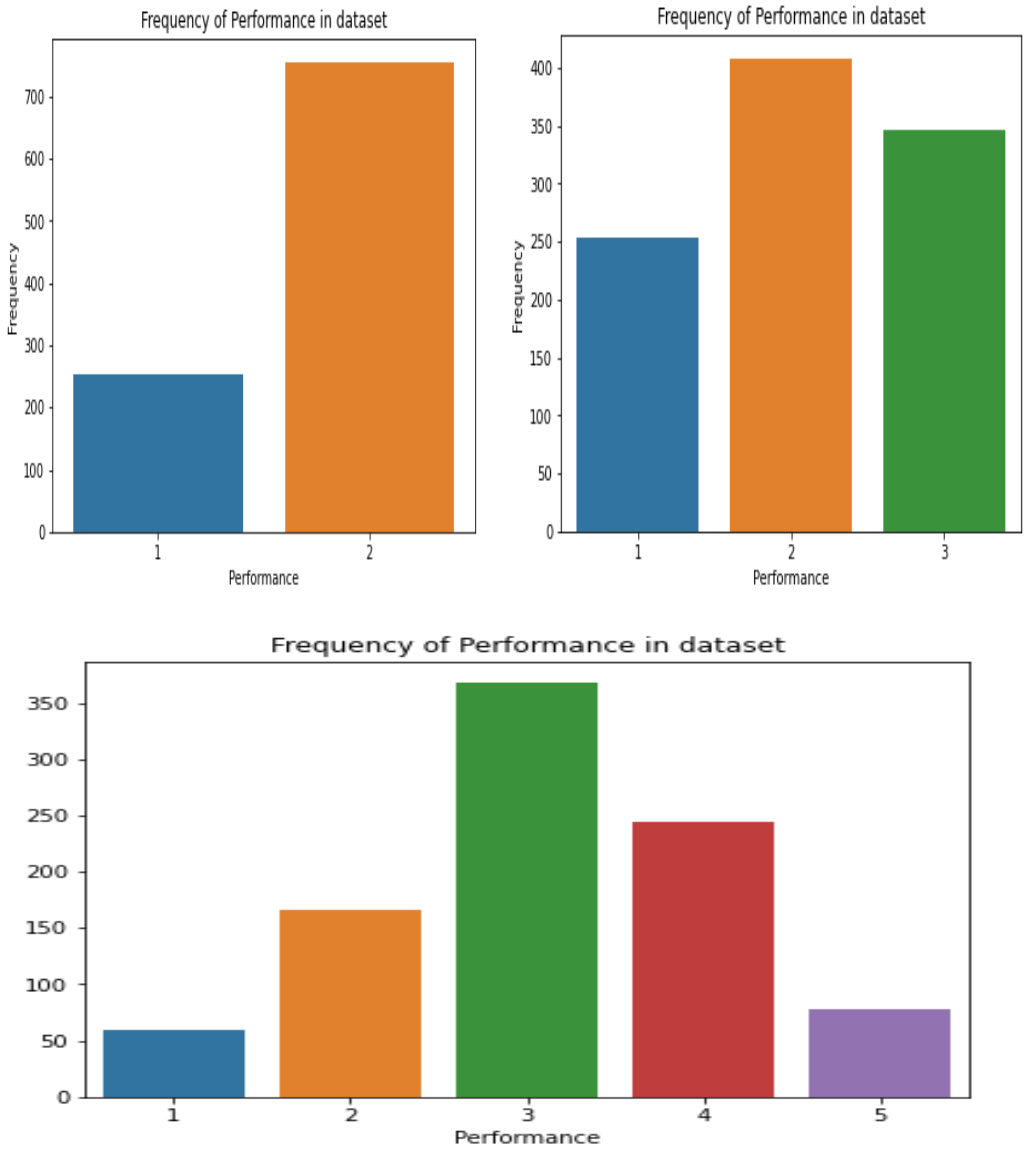

Figure 6.1 Representation of of Result Categories

This dataset was used for analysis of student performance in technical programs. It includes 1850 students' records.

To choose the ML techniques for predicting the student potential and performance, data had to be transformed in a form suitable to ML techniques, because some algorithms work with numerical attributes and some algorithms prefer nominal attributes to perform prediction. The performance or result was transformed as given in Table 6.1 below:

Total Percentage	Classification
75 & above	Excellent
65-75	Very Good
50-65	Satisfactory
40-50	Poor
Below 40	Fail

Table 6.1 Result Categories

6.3 Finding Best Classifier model to analyze student performance

In this Section, we present the prediction of student performance, based on different class labels. Value of predictive accuracy is highly dependent on the different class base labels. The performance of the classifier models is dependent on the list of attributes used in the student dataset.

Individual classifier models discussed in Chapter 3 are used in this study. The performance of the classifier models can be compared in terms of accuracy and cross validation score. This study has recommended the ideal classifier for student performance prediction models.

Students' Dataset was pre-processed and later fed to Decision Tree, Random Forest, Support Vector Machine, Logistics Regression, Ada Boosting, and Stochastic Gradient Boosting. Training and testing were performed in ten different folds resulting in an accurate Model. The results obtained from the built model were measured in terms of Accuracy. We discuss these models next.

Logistics Regression

As mentioned earlier, Logistic Regression (LR) is a simple but powerful algorithm for new and available data and assigns new data into the category which is most similar to the available categories. A certain function between the target categorical variable and independent variables is calculated by estimating the probabilities using the logistic function. It gives good accuracy in predicting the performance.

LR was used for evaluating the performance of students based on class labels. This method was applied on the three class labels and results showed that accuracy for the two labels case is higher than for the three-label and five-label cases, as shown in Table 6.2. The difference in accuracy between the two-label case and the five-label case is approximately 5%.

Class Labels	**Accuracy**	**Cross validation score**
Two	0.99	0.9504
Three	0.983	0.9661
Five	0.9475	0.8943

Table 6.2 LR result evaluation for analysis of student performance

Support Vector Machine

SVM is also a frequently implemented classification algorithm and is a Supervised Learning method. This algorithm has been used in many studies, for predicting student performance. It is most suitable for small datasets and is faster than other methods.

The SVM model performance results are shown in Table 6.3. Here the difference between the label classes is approximately 6%. The accuracy of the two-label case was better than LR but the accuracy of the five-label case was better in LR.

Class Label	Accuracy	Cross validation score
Two	0.995	0.9831
Three	0.9801	0.9669
Five	0.9348	0.9009

Table 6.3 SVM result evaluation for analysis of student performance

AdaBoost

AdaBoost or Adaptive boosting method is an ensemble classifier made strong by combining multiple classifiers to increase accuracy. Multiple sequential models are built, each correcting the errors from the previous models. Many papers in the literature have used AdaBoost algorithm to predict student performance with different attributes.

AdaBoost was applied on the student datasets for evaluating the student performance. The results are shown in Table 6.4. The difference in accuracy between the two-label and five-label cases was approximately 15% which is much greater than both LR and SVM classifiers.

Class Label	Accuracy	Cross validation score
Two	0.9872	0.9834
Three	0.9801	0.9735
Five	0.8399	0.7986

Table 6.4 AdaBoost result evaluation for analysis of student performance

Decision Tree

Decision tree (DT) is one of the most usable techniques for prediction. Most of the studies have used this method because it can be used on both small or large datasets. It gives good accuracy for prediction. It uses IF-THEN rules for prediction of student performance. DT was applied on the dataset and the results evaluated for analysis of student performance.

The summary of DT model performance is shown in Table 6.5. The difference between accuracy of the three cases is approximately 5%. This model was better than AdaBoost but underperformed as compared to LR and SVM.

Class Label	Accuracy	Cross validation score
Two	0.9872	0.9800
Three	0.9801	0.9735
Five	0.932011	0.90759

Table 6.5 DT result evaluation for analysis of student performance

Random Forest

It is an Ensemble method that has been used in education data mining. In this method, a large dataset is divided into a number of subsets and the model is applied on each subset independently. It is a combination of trees. Random Forest was applied on the student data for analysis of student performance.

The RF model performance details are shown in Table 6.6. The difference in accuracy between the two-label and five-label was 3% and it performed better than the other algorithms.

Class Label	Accuracy	Cross validation score
Two	0.9985	0.9834
Three	0.9886	0.9834
Five	0.9645	0.891

Table 6.6 RF result evaluation for analysis of student performance

Stochastic Gradient Boosting

This is also an Ensemble method which uses boosting. Here each succeeding model attempts to correct the errors of the previous one. Boosting is often applied to similar or the same algorithms and uses majority voting for its decision strategy. Ensemble methods give better predictive performance compared to a single model.

A summary of GB model performance details is shown in Table 6.7. Approximately 8% difference is found in the accuracy between two- and five- class label cases.

Class Label	Accuracy	Cross validation score
Two	0.9929	0.9801
Three	0.973	0.9471
Five	0.81303	0.75577

Table 6.7 GB result evaluation for analysis of student performance

6.4 Overall Evaluation of student performance prediction models

In this section we present a comparison of the performance of the six classifier models discussed above. We do it for each of three classes separately.

6.4.1 Performance for the two-labels class

The overall results of classifiers performance for this class are given in Table 6.8. we note that the GB classifier method achieved predictive accuracy of 99%, while DT achieved lowest accuracy of 98%. The accuracy of the class thus ranges from 98 to 99%. Figure 6.2 shows the same comparison as a Bar chart.

ML Models	Accuracy	Cross-Validation Score
Decision Tree	0.9872	0.9800
Random Forest	0.9985	0.9834
Support Vector Machine	0.9950	0.9831
Logistics Regression	0.9900	0.9504
Ada Boosting	0.9872	0.9834
Stochastic Gradient Boosting	0.9929	0.9801

Table 6.8 Performance Analysis for Students dataset with two- labels

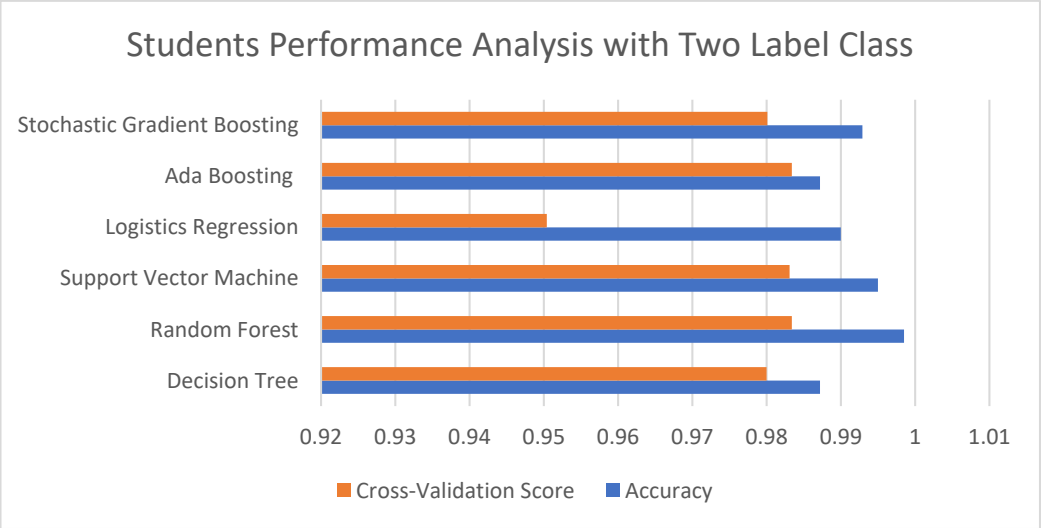

Figure 6.2 Bar chart of Performance for Students dataset with two-label class

6.4.2 Performance for the three-label class

The classifier accuracy for this class is shown in Table 6.9. It is seen to have marginally come down to the range from 94% to 97%.

Two class accuracy is found to be always greater than the multi-label classes. The predictive accuracy of the classifiers decreases on increasing the label values. The highest accuracy of 98.86% was achieved by Random Forest. Figure 6.3 shows the performance analysis of students with three-label classes.

ML Models	Accuracy	Cross-Validation Score
Decision Tree	0.9801	0.9735
Random Forest	0.9886	0.9834
Support Vector Machine	0.9801	0.9669
Logistics Regression	0.9830	0.9661
Ada Boosting	0.9801	0.9735
Stochastic Gradient Boosting	0.9730	0.9471

Table 6.9 Performance Analysis of Students dataset with three-labels

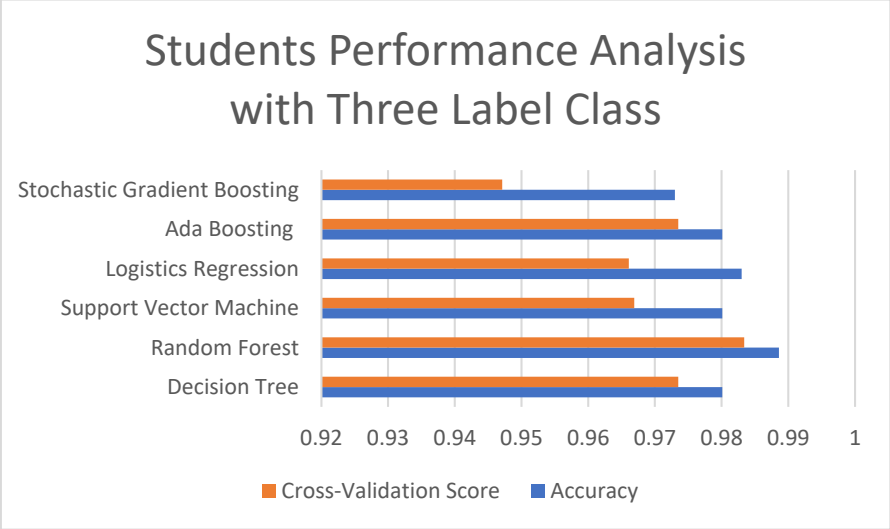

Figure 6.3 Bar chart of Performance for three-label class

6.4.3 Performance for the five-label class

The performance of the classifiers is shown in Table 6.10. Random Forest gives highest accuracy while GB gives lowest accuracy. Figure 6.4 shows the comparison between six classifiers for analysis of students' performance in bar chart for this class.

ML Models	Accuracy	Cross-Validation Score
Decision Tree	0.932011	0.90759
Random Forest	0.964589	0.89108
Support Vector Machine	0.934844	0.90099
Logistics Regression	0.947592	0.8943
Ada Boosting	0.83994	0.79867
Stochastic Gradient Boosting	0.813031	0.75577

Table 6.10 Performance Analysis of Students dataset with five-labels

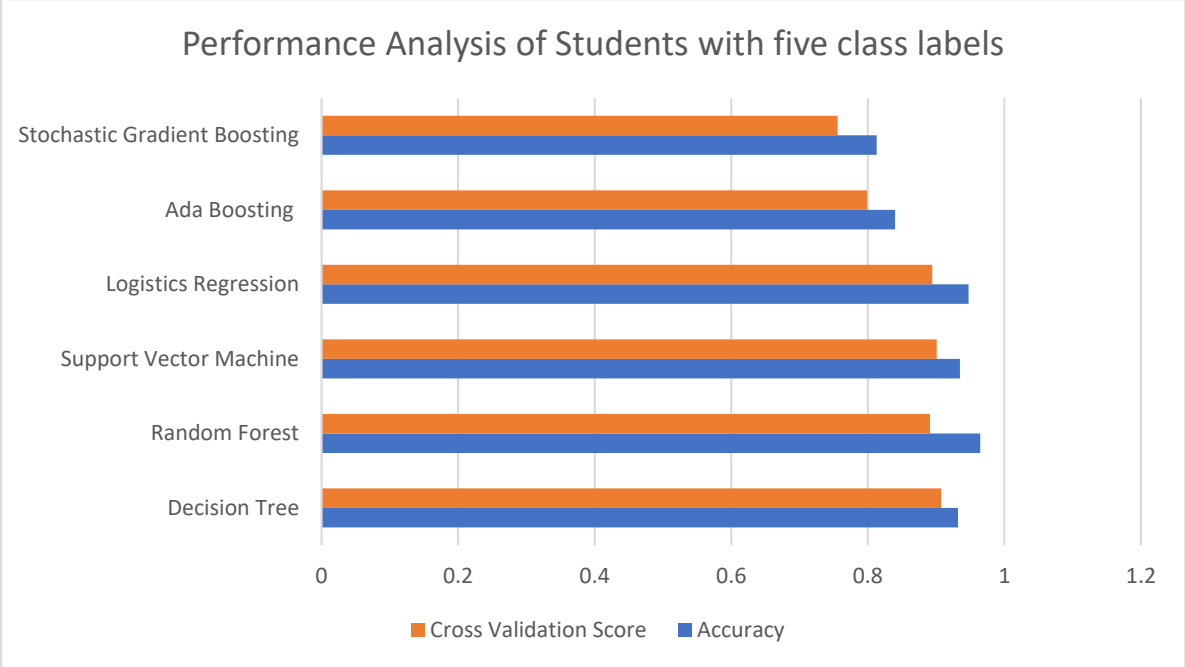

Figure 6.4 Performance Analysis of Students with five class labels

6.7 Chapter Summary

This chapter focused on the work to find the best classifier for student performance prediction. Different classifier models were built, such as LR, DT, SVM, AdaBoost, RF and GB. The target variable was categorized based on different class labels. This study classified the dataset into three categories: two-labels, three-labels and five-labels. This chapter concludes that the accuracy of the two-labels class was always found to be greater than the multi-labels classes. The models' performance was compared on the basis of accuracy. The result showed that RF got higher accuracy than other models.

The next chapter concludes the thesis and gives suggestions for future work in this research area.