Machine Learning and Data Mining Approach to Predict Academic Performance of Students

B.Tech. Project Report-II

By

Reet Roy
Souhardya Mandal
Manthan Chowdhary
Subhajit Bokshi

Under Supervision of
Dr. Partha Ghosh
and
Prof. Ranjit Kr. Mandal

Department of Computer Sc. and Engineering

Government College of Engineering and Ceramic Technology
Kolkata

December 2021

Machine Learning and Data Mining Approach to Predict Academic Performance of Students

A Project Report

Submitted in partial fulfillment of the
requirements for the award of the degree
of

Bachelor of Technology

In

Computer Sc. and Engineering

By

Reet Roy, GCECTB-R18-3019
Souhardya Mandal, GCECTB-R18-3038
Manthan Chowdhary, GCECTB-R18-3013
Subhajit Bokshi, GCECTB-R18-3031

Department of Computer Sc. and Engineering

Government College of Engineering and Ceramic Technology
Kolkata

December 2021

DECLARATION

We hereby declare that the project entitled **“Machine Learning and Data Mining Approach to Predict Academic Performance of Students”** submitted for the B. Tech. (CSE) degree is our original work and the project has not formed the basis for the award of any other degree, diploma, fellowship or any other similar titles.

Name and Roll No. of the Students **Signature of the Students**

1. Reet Roy, GCECTB-R18-3019 ………………………………………..
2. Souhardya Mandal, GCECTB-R18-3038 ………………………………………..
3. Manthan Chowdhary, GCECTB-R18-3013 ………………………………………
4. Subhajit Bokshi, GCECTB-R18-3031 ……………………………………….

Place:

Date:

Government College of Engineering and Ceramic Technology

73, A. C. Bannerjee Lane, Kolkata, West Bengal 700010

………………………………………………………………………………

BONAFIDE CERTIFICATE

Certified that this project report titled **“Machine Learning and Data Mining Approach to Predict Academic Performance of Students”** is the authentic work carried out by **Reet Roy(GCECTB-R18-3019), Souhardya Mandal (GCECTB-R18-3038), Manthan Chowdhary (GCECTB-R18-3013), Subhajit Bokshi(GCECTB-R18-3031)** who carried out the project work under **my / our** supervision. Certified further, that to the best of my knowledge the work reported herein does not form part of any other project report or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

………………………………………

Dr. P. Ghosh
SUPERVISOR
Assistant Professor
Department of Computer Science and Engineering
Government College of Engineering and Ceramic Technology
Kolkata-700010

………………………………………

Prof. Ranjit Kr. Mandal
JOINT SUPERVISOR
Assistant Professor
Department of Computer Science and Engineering
Government College of Engineering and Ceramic Technology
Kolkata-700010

………………………………………

Dr. K. Saha Roy
HEAD OF THE DEPARTMENT
Assistant Professor & Head
Department of Computer Science and Engineering
Government College of Engineering and Ceramic Technology
Kolkata-700010

…………………………………………

External Examiner

Abstract

Data mining offers strong techniques for different sectors involving education. In the education field the research is developing rapidly increasing due to huge number of student's information which can be used to invent valuable pattern pertaining learning behavior of students. The institutions of education can utilize educational data mining to examine the performance of students which can support the institution in recognizing the student's performance. In data mining classification is a familiar technique that has been implemented widely to find the performance of students. The prediction of academic performance of students has been an important and developing research domain in educational data mining (EDM), in which data mining and machine learning techniques are used for deriving data from educational warehouse. Relevant research study reveals that numerous methods for academic performance forecasting are built to carryout improvements in administrative and teaching staff of academic organizations. In the put forwarded approach, the acquired data-set is pre-processed to purify the data quality, the labeled academic historical data of student is utilized to train KNN Classifier, Support Vector Classifier, Random Forest and DT-classifier.

ACKNOWLEDGEMENT

First and foremost, we would like to thank **Dr. Partha Ghosh** sir and **Prof. Ranjit Kr. Mandal** sir for guiding us throughout the project.

I would like to thank **Head of the Computer Science and Engineering Department Mrs. Kalpana Saha (Roy),** who helped us whenever required.

Besides, we would like to thank all the teachers who helped by giving us valuable advice and the equipment which we needed.

At last but not least, I would like to thank everyone who helped and motivated us to work on this project.

Reet Roy, GCECTB-R18-3019

Souhardya Mandal, GCECTB-R18-3038

Manthan Chowdhary, GCECTB-R18-3013

Subhajit Bokshi, GCECTB-R18-3031

Table of Contents

1. INTRODUCTION

1.1. PROBLEM DEFINITION

The purpose of this project is to take various features of student as input feature and predict the grade of the student based on this factor using various ML classification technique to classify students and help the institute to take necessary actions and effort to improve the performance of students of each category and support them to success.

1.2. PROJECT OVERVIEW & SCOPE

The educational institutes over the world are concentrating on generating graduates with better academic performances.

For this they are keeping track on how the students are performing in a specific sector and in what field they need much training .

However what if we could take some of this huge data of student, kept by the university and predict the performance of students based on it.

This will enable educational institutes to pay greater attention to the predicted underperforming student beforehand and based on their effort they can achieve better grades.

2. LITERATURE SURVEY

This section reviews on how researchers of the past have made use of several standard classification algorithms in predicting student academic performance. Here we will mainly talk about two approaches: the first one **using a Hybrid Data Mining Approach** by Bindhia K. Francis & Suvanam Sasidhar Babu and the second one using **hybrid educational data mining model (HEDM) by V. Ganesh Karthikeyan, P. Thangaraj and S. Karthik .**

● The first paper used a 2-layer approach for classifying: first it fed the data to the supervised learning classifiers and then used unsupervised learning clustering algorithm (K means clustering plus majority voting) in order to increase the accuracy. The dataset includes several features like the Demographic, Academic, Behavioural and Extra features. This data has been taken with regard to the interaction of the student with the online website of 2 institutions from Kerala.The researchers took a subset of different features and used 4 supervised classifying algorithms :Naive Baye's, Decision Tree (Iterative Dichotomiser 3 algorithm), Support Vector Machine, and Neural Network (Multilayer Perceptron) in first layer and then the K means clustering algorithm grouped them into clusters indicating high, medium and low performing students. As such they used four performance metrics in order to compare the performance of various classifiers under different criteria: Precision,Recall, Accuracyand F1-score. Through various combination of feature and machine learning models they finally arrived at the conclusion that Academic and Behavioral features of students contributed largely to the grade received by the student and Decision Tree was the most accurate model for predicting grades with an accuracy of 75%

● The model proposed in the second paper combines the efficiencies of Naive Baye's classification technique and J48 Classifier for deriving the results and categorizing the student performance in precise manner.Data used in this paper is collected during the 2005- 2006 school year from two public schools, from the Alentejo region of Portugal.The paper proposes to categorize data into pass and fail list using Naive Baye's and then uses J48 Decision Tree Classifier to categorize the data further to identify the Student_Nature under the classes such as, Excellent_List, Good_List, Average_List and Low_List. To compare the performance of this approach four performance metrics are used:Precision, Recall,F1-Score and Accuracy. The final accuracy achieved by Naïve Baye's + J48 Decision Tree classifier_was 98.6% using the WEKA tool.

3. METHODOLOGY

This section explains the methodology which involves the phases used for the process of prediction of academic performance of the students through classification and clustering student data. The phases are depicted in the below figure (Fig. 1): Each phase of the methodology proposed is explained below:

3.1 Data Flow Diagram

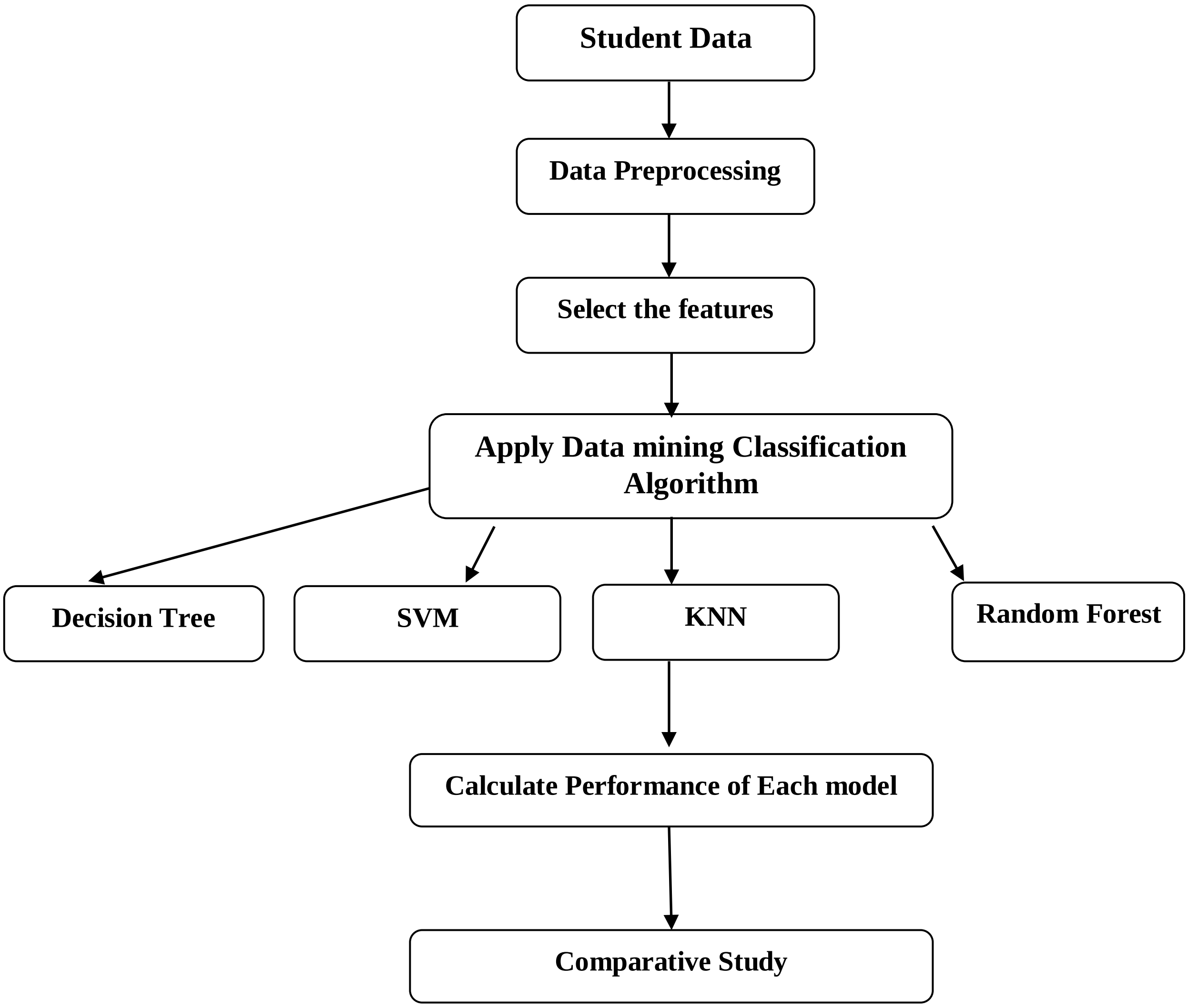


Fig1: Data flow diagram of our model

3.2 IMPLEMENTATION

3.2. 1: Understanding of Business

The main aim of this study is to develop a model of predication for academic performance of students using data mining classification and to decide which classifier performs good with the gathered data set of education.

3.2.2: Data collection

Data used in this paper is collected during the 2005- 2006 school year from two public schools, from the Alentejo region of Portugal.

The database was built from following sources:

○ school reports, based on paper sheets
○ attributes (the three period grades and number of absences)
○ questionnaires, used to complement the previous information.

Attribute Information:

Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:
1. school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2. sex - student's sex (binary: 'F' - female or 'M' - male)
3 age - student's age (numeric: from 15 to 22)
4. address - student's home address type (binary: 'U' - urban or 'R' - rural)
5. famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6 Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7. Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€ “5th to 9th grade, 3 â€ “secondary education or 4 â€ “higher education)
8. Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€ “5th to 9th grade, 3 â€“secondary education or 4 â€“ higher education)
9. Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g., administrative or police), 'at_home' or 'other')
10. Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g., administrative or police), 'at_home' or 'other')
11. reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course'preference or 'other')
12. guardian - student's guardian (nominal: 'mother', 'father' or 'other')

13. traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14. studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15. failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16. schoolsup - extra educational support (binary: yes or no)
17. famsup - family educational support (binary: yes or no)
18. paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19. activities - extra-curricular activities (binary: yes or no)
20. nursery - attended nursery school (binary: yes or no)
21. higher - wants to take higher education (binary: yes or no)
22. internet - Internet access at home (binary: yes or no)
23. romantic - with a romantic relationship (binary: yes or no)
24. famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25. freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26. goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27. Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28. Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29. health - current health status (numeric: from 1 - very bad to 5 - very good)
30. absences - number of school absences (numeric: from 0 to 93)

# these grades are related with the course subject, Math or Portuguese:
31. G1 - first period grade (numeric: from 0 to 20)
31. G2 - second period grade (numeric: from 0 to 20)
32. G3 - final grade (numeric: from 0 to 20, output target)

3.2.3: Data Pre-processing:

After the collection of dataset pre-processing methods are applied to develop the data set quality. The data preprocessing is regarded as an essential step in the process of knowledge discovery which involves cleaning of data, feature selection, data transformation and data reduction. Before applying the data mining algorithm the data preprocessing is the step which transforms the actual information into an applicable shape to be used by a specific algorithm of mining.

3.2.4: Applying Data Mining Classification Algorithms

This study carried out the experiments using SVM, KNN, Decision tree, Random Forest classifiers. These five classifiers have been chosen to evaluate the measures of dataset. Support vector machine is used for solving the non linear function estimation and pattern recognition issues. The Support vector machine is used for representing the training information nonlinearly into a higher dimensional feature space then builds an isolating hyper plane with maximum margin. This yields a non linear boundary of decision in input space. Support vector machine solutions are acquired from issues of quadratic programming possessing a global solution.

I) Decision Tree Algorithm

Decision trees create a model or a tree that predicts the value of a target variable (in this cased passed) by learning simple decision rules from the data. For example in our data set it is most likely that children with only one guardian are more likely to drop out of school. In this case a simple decision tree with one node would classify pupils as dropouts if one of the guardian variables are zero.

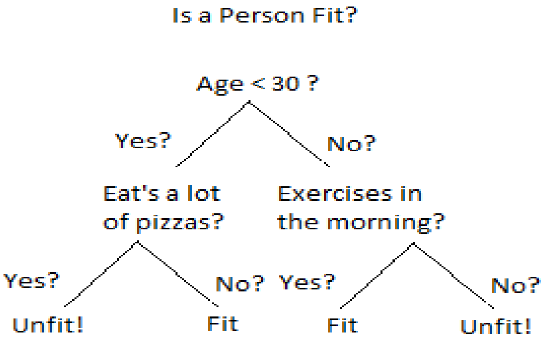


Fig 2: Working of Decision Tree

Advantages and Disadvantages of Decision Tree

Decision trees have many advantages. First of all, they can be easily visualized. This is not possible in case of many other algorithms. For example, in SVM, it is basically impossible for us to imagine data in 10-dimensional space. Secondly, they have logarithmic run times, therefore they are very fast. When examining the performance table above you can see that they are the fastest. Another advantage would be that decision trees can use categorical data even without any formatting, however in our examples we are already preprocessing data, so we are taking advantage of this feature.

Several disadvantages exist. First it is easy to see from the performance table that the decision tree that we have created is over fitting and there is huge difference between F1 scores of training and testing data sets. Second disadvantage is that decision trees do not perform well if one the classes dominates, but I think in this case it is 30% vs. 70%, so they are roughly equal.

Given this discussion, I tried decision trees for this particular problem because they are usually used for classification, easy to understand and visualize and suitable for our data where most of the features are categorical.

II) Support Vector Machine

Support vector machines are a set of classifiers that try to find the optimum linear separator between two classes. Although in reality most data would not be linearly separable, SVM uses a kernel trick that adds news features into the data by combining existing features in various ways. For example, a set of data points containing x and y features may not be linearly separable in two-dimensional space, but they may be in three-dimensional space where $x^2 + y^2$ is the third variable.

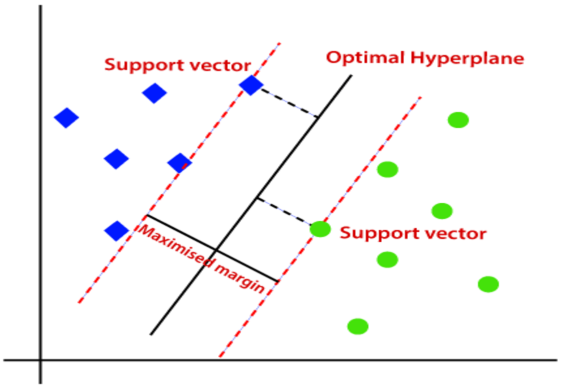


Fig 3: Working of SVM

Advantages and Disadvantages of Support Vector Machine

There are two advantages of using support vector machines. One is that they use only a small subset of data in order to train model and they are called support vectors. This is a more memory efficient method in situations where we have tons of data. A second advantage is that they are robust in cases where the numbers of data points are on the order of the number of features. In other words they suffer less from the curse of dimensionality.

There may be two disadvantages. One is that it may difficult to explain how they work. The concept of high dimensionality may be difficult to grasp. SVM also does not work when the number of features is much bigger than data, but this is rarely the case. I think it is suitable in this case because it is a widely used classification algorithm that performs best right out of the box.

III) K Nearest Neighbours Algorithm

Classification based on nearest neighbours algorithm is one where no explicit model for the data is constructed. Classifier simply stores the data and whenever a new point comes, classifier assigns the new point to a class based on its closest k neighbours.

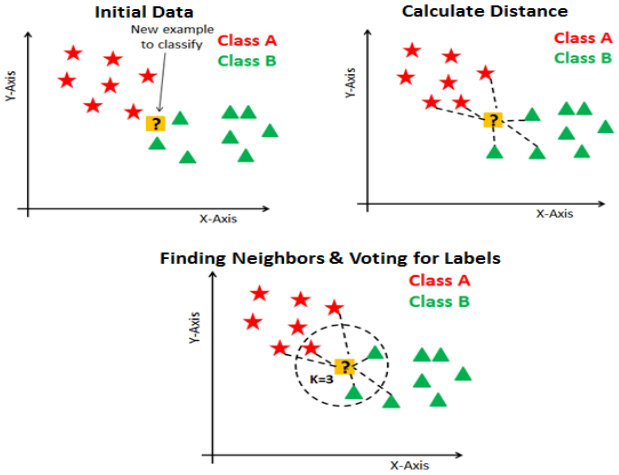


Fig 4: Working of KNN

Advantages and Disadvantages of K Nearest Neighbours Algorithm

The main advantage of KNN is that training times are short. There is essentially no training and no model. Therefore it is easy to update KNN as the new data comes. Another advantage is that it will be very easy to explaining this model in laymen's terms.

However the disadvantage is that predicting each new data point takes a long time. For each data point we have to calculate its nearest neighbours. Another disadvantage is the curse of dimensionality as explaining by Charles Isbell during the lectures. As the dimensions of the feature space increases each data point is going to represent a bigger volume of space for which it may not be representative. I wanted to try KNN also as it is a very simple algorithm that makes no assumption about the data.

IV) Random Forest Algorithm

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

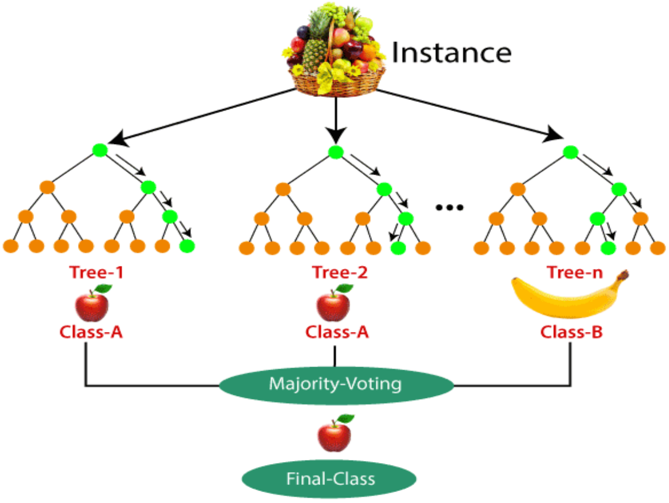


Fig 5: Working of Random Forest Classifier

Advantages and Disadvantages of Random Forest Algorithm

Random Forest is capable of performing both Classification and Regression tasks. It is capable of handling large datasets with high dimensionality. It enhances the accuracy of the model and prevents the over fitting issue.

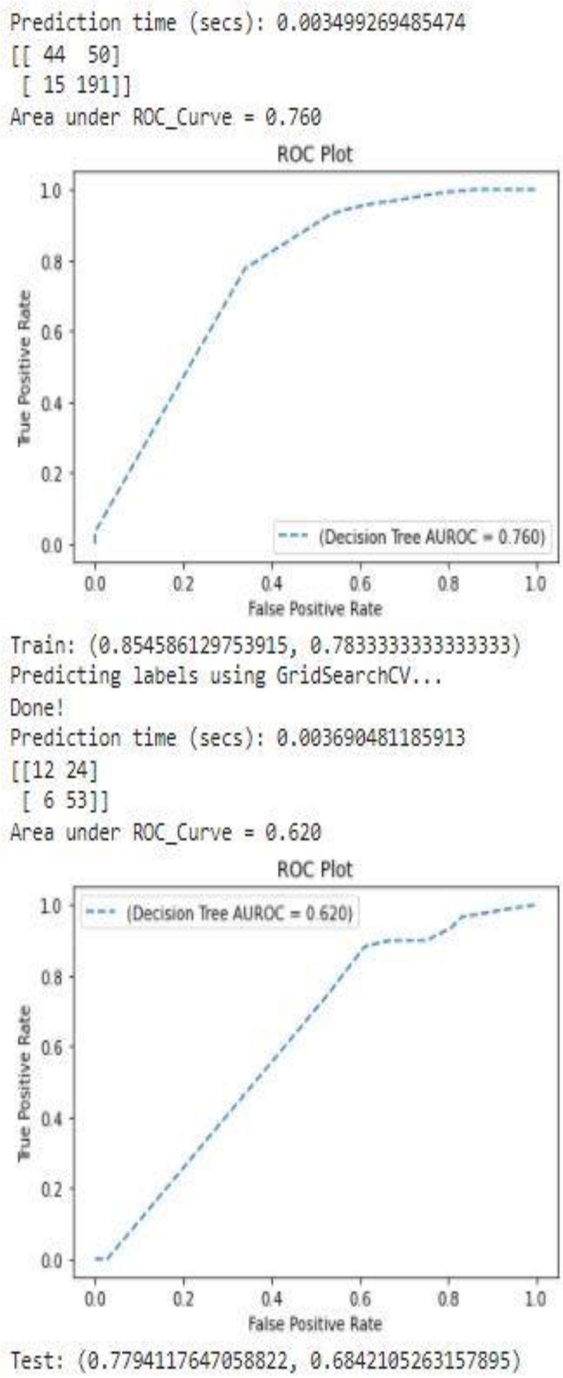
Although random forest can be used for both classification and regression tasks, it is not more suitable for Regression tasks.

.

4. RESULT

We have applied all the above mentioned classifiers on the dataset one by one and have calculated the performance metrics.

i) Decision Tree



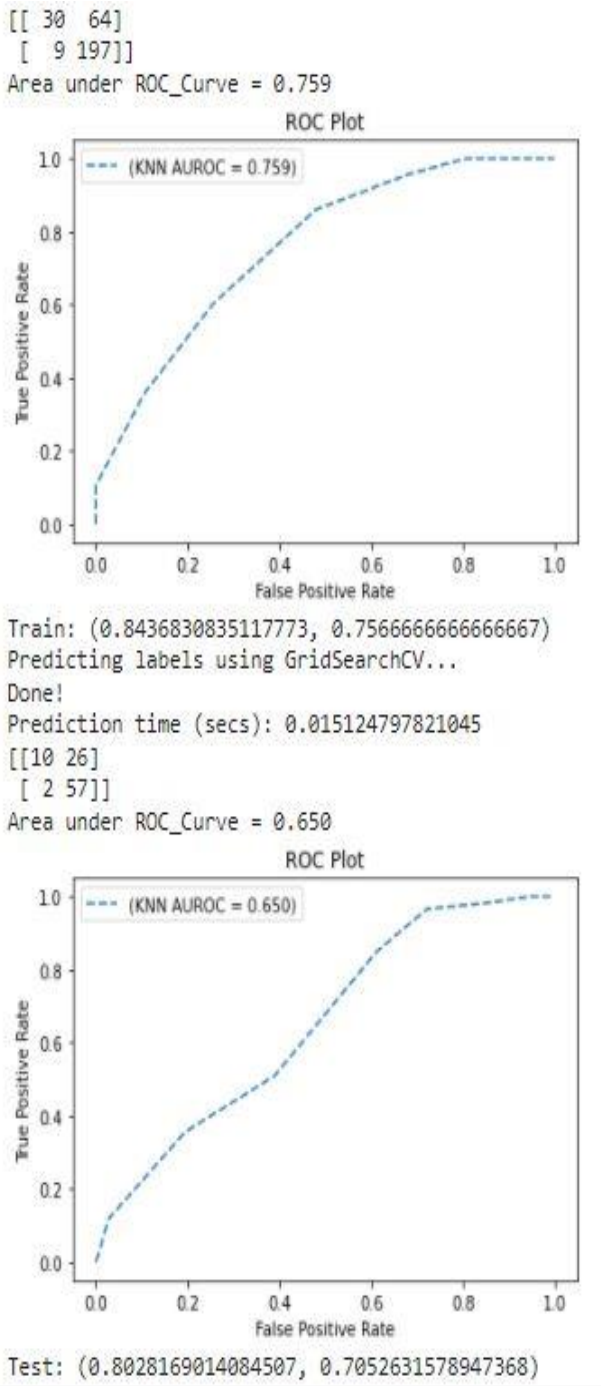
```
Prediction time (secs): 0.003499269485474
[[ 44  50]
 [ 15 191]]
Area under ROC_Curve = 0.760
```

```
Train: (0.854586129753915, 0.7833333333333333)
Predicting labels using GridSearchCV...
Done!
Prediction time (secs): 0.003690481185913
[[12 24]
 [ 6 53]]
Area under ROC_Curve = 0.620
```

```
Test: (0.7794117647058822, 0.6842105263157895)
```

Fig 6: Performance of Decision Tree

ii) KNN



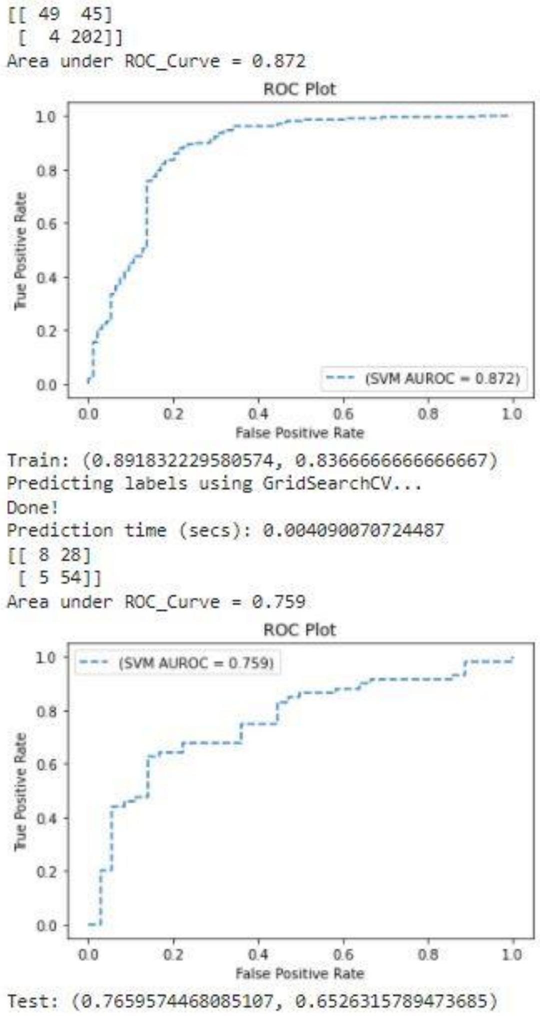
```
[[ 30  64]
 [  9 197]]
Area under ROC_Curve = 0.759
```

```
Train: (0.8436830835117773, 0.7566666666666667)
Predicting labels using GridSearchCV...
Done!
Prediction time (secs): 0.015124797821045
[[10 26]
 [ 2 57]]
Area under ROC_Curve = 0.650
```

```
Test: (0.8028169014084507, 0.7052631578947368)
```

Fig 7: Performance of KNN

iii) SVC



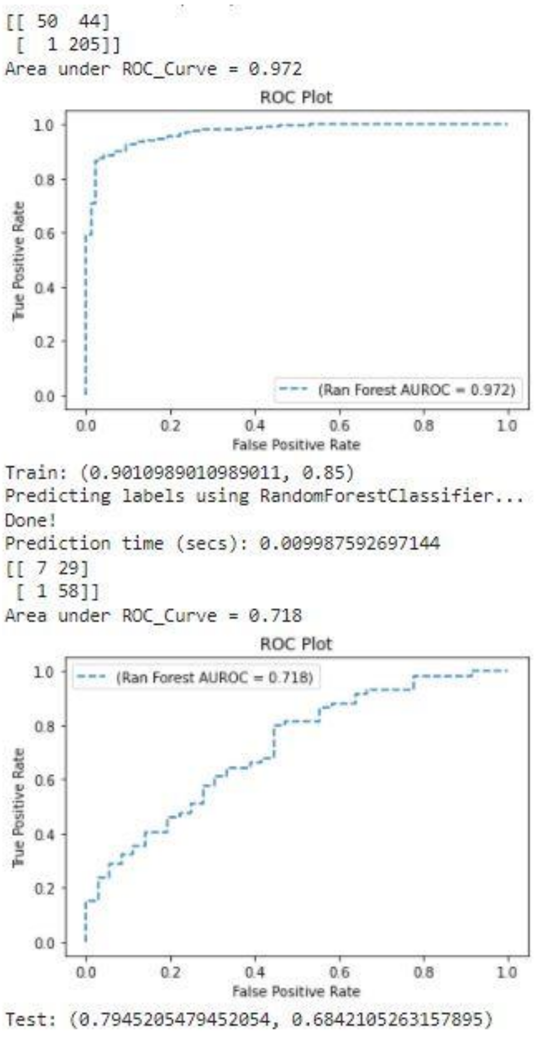
```
[[ 49  45]
 [  4 202]]
Area under ROC_Curve = 0.872
```

```
Train: (0.891832229580574, 0.8366666666666667)
Predicting labels using GridSearchCV...
Done!
Prediction time (secs): 0.004090070724487
[[ 8 28]
 [ 5 54]]
Area under ROC_Curve = 0.759
```

```
Test: (0.7659574468085107, 0.6526315789473685)
```

Fig 8: Performance of SVC

iv) Random Forest



```
[[ 50  44]
 [  1 205]]
Area under ROC_Curve = 0.972
```

```
Train: (0.9010989010989011, 0.85)
Predicting labels using RandomForestClassifier...
Done!
Prediction time (secs): 0.009987592697144
[[ 7 29]
 [ 1 58]]
Area under ROC_Curve = 0.718
```

```
Test: (0.7945205479452054, 0.6842105263157895)
```

Fig 9: Performance of Random Forest

Performance of each model on test data is mentioned in the table below:

Classifier Name	**F1 score**	**Accuracy**	**Area under ROC curve**
Decision Tree	0.78	0.68	0.62
Support Vector Machine	0.76	0.65	0.76
K Nearest Neighbors	0.80	0.71	0.65
Random Forest	0.79	0.68	0.72

Below is the graph in which accuracy of each model on training set and test set is plotted side by side.

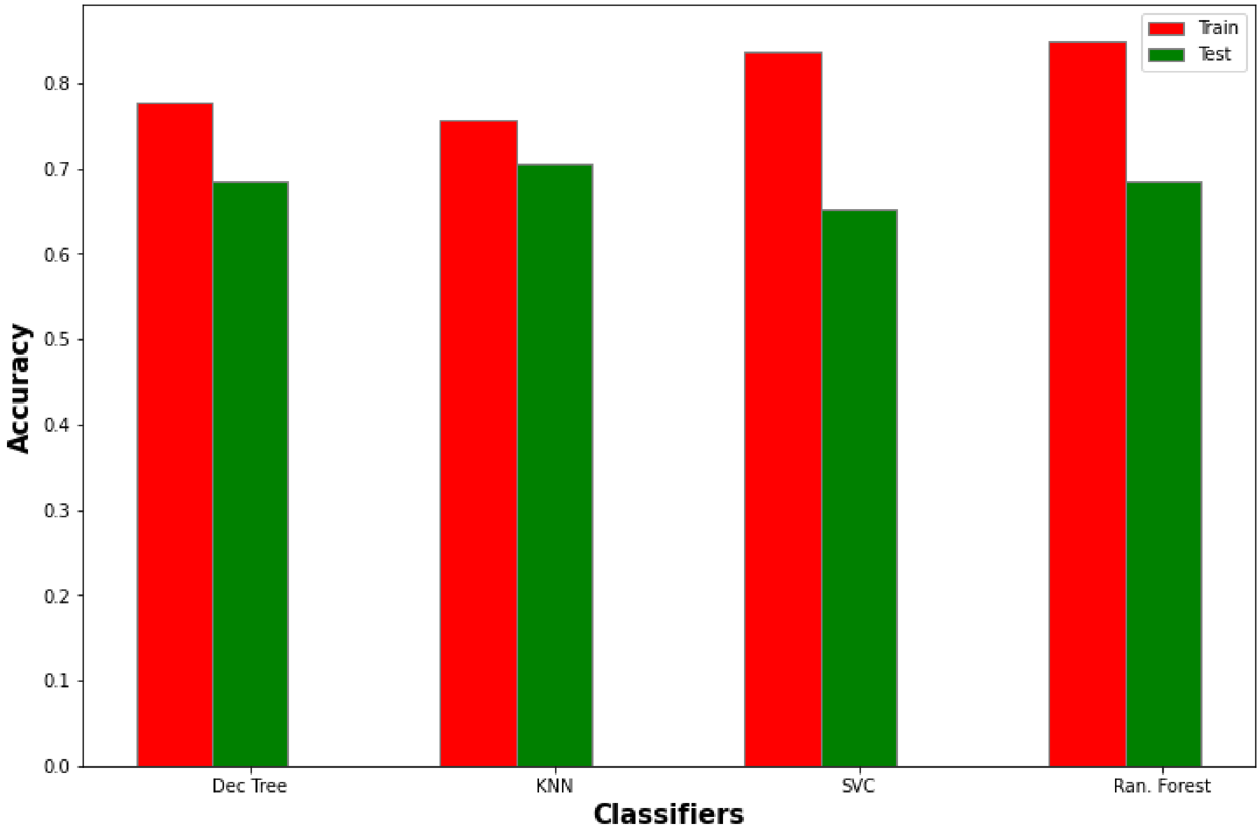


Fig 10: Comparison of Performance of Different Classifiers

5. CONCLUSION & FUTURE WORKS

In this project we applied various classification techniques on the student dataset to determine the student’s final result (i.e. Pass or Fail) and we have found that the KNN can be a reliable model to predict student performance accurately because it shows consistent accuracy, f1 score and auroc curve performance. However, we plan to extend this project further by training models on features which affect student grades more than the others, by converting string data (such as fam_size) to integer as well as we will try to implement loss function to compare the performance of these classification models and reach a final decision.