Student Performance Prediction

Isha D Shetty
Department of Information
Technology
VESIT
Chembur, India

Dipshi   Shetty
Department of Information
Technology
VESIT
Chembur, India

Sneha Roundhal
Department of Information
Technology
VESIT
Chembur, India

Abstract:Predicting Students performance beforehand can be very beneficial for educational institutions to improve their teaching quality. This paper proposes to predict students performance by considering their academic details. For classification, decision tree, Random forest, Support vector machine, Linear regression, Naive Bayes are used. This paper also focuses on the use of Artificial Neural network algorithms such as MLP and Convolutional Neural network for predicting the performance.

1.   INTRODUCTION

Today every educational institution handles and deals with large amount of student data which can be beneficial for a number of reasons. One of the important application of such data is predicting student performance. Such a prediction can be useful not only for the students but also for teachers/mentors. Mentors can provide special assistance to the students who are on the verge of failing. In order to determine which category a student lies, such data can be quite helpful. This application can be used by any prominent school or colleges. It can be used to predict the pointer ranges or percentage range for future semester exams. These ranges can be predicted using a number of data mining algorithms such as classification algorithms, rule-based algorithms, ensemble methods, and neural networks. The main aim of this project is the selection of features that show a strong relationship with a target attribute that is to be predicted from a high dimensional dataset. We have evaluated and compared the number of algorithms such as decision tree, random forest, support vector machine, naive Bayes and neural networks by applying them on the dataset The rest of the paper provides an explanation on nature of neural networks along with the results of our evaluation.

1.1   Objective

Today every educational institution handles and deals with large amount of student data which can be beneficial for a number of reasons. One of the important application of such data is predicting student perfo

2.   OVERVIEW OF DATASET

This dataset includes actual academic details of the students of one of the prominent engineering institute of Mumbai University. Engineering students of mumbai university are classified as Regular students (who join engineering colleges after their Higher Secondary education) and Direct Second Year (DSE-who join after completing their diploma degree). The SGPI(Student Grade Performance Index) of a particular student varies from .01 to .99 and can be difficult to analyze and predict, hence to get efficient prediction the academic performance data is categorized.

It has 6050 instances and 55 attributes. The dataset contains students Personal details like parents name, date of birth, address etc and academic features like marks of all semesters, HSC percentages, SSC percentages etc. There is a mix of the numeric and nominal attribute.

Some of the attributes considered while training the model are :

Table -1: Dataset

Attributes
Student Type (Regular or DSE)
SSC Board
SSC Percentage
HSC Board
HSC Percentage

DSE Board
DSE Percentage
Semester Marks (Graduation)
Number of KT's
Education Gap
Branch
Admission Type(CAP or Minority)

3.   DATA PREPROCESSING

The dataset is prepared to apply data mining techniques. Preprocessing methods such as data cleaning, transformation of variables, data balancing and data partitioning had to be applied.

There are 17 numeric attributes with null values. To deal with null values, semester pointers and 10th percentage columns are replaced with their respective column medians. On the other hand, the columns such as the number of KTs were replaced with 0 as in the dataset there is proper mention of whether the person has a kt. So in case if the value of this attribute is NAN, then it means the student has no KT. The SSC percentages which are in terms of CGPI were converted into percentages.

The dataset was split into the training and testing data. The ratio of the training and testing data is 8: 2.

3.1   Data Balancing

Unbalanced data typically refers to a problem with classification problems where the classes are not represented equally. The unbalanced data is solved using SMOTE also called as synthetic sampling. In both 3 and 5 category section, the SGPI range<=4 i.e class 0 has less number of instances comparatively. Since this class is important to identify weak students, it cannot be neglected. This technique is similar to oversampling where the instances of class 0 are increased to handle imbalance.

3.2   Feature Selection

The attributes such as name,email-id, phone-number, mother's name, father's name, address are unique and are dropped as it shows the least relationship with the target attribute. The Semester pointers and 12th board results and 10th board results showed a linear relationship with the target variable. Correlation states how the features are related to each other or the target variable.Thus, the correlation matrix was calculated for finding the relationship. For instance, the semester 5 SGPI range was to be predicted. So, it showed a strong relationship with the following features.

Fig -1: Result of correlation matrix using heatmap
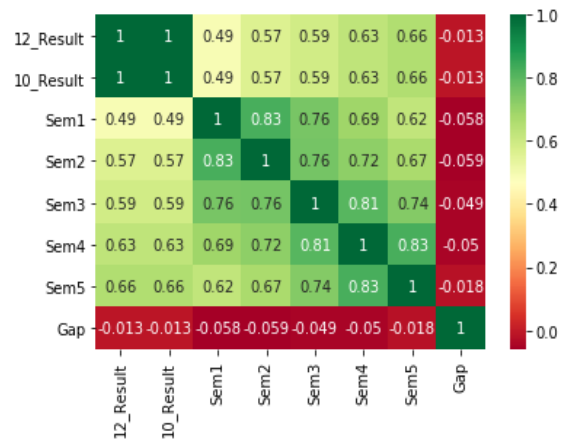
The features such as 12_Result,10_Result,Sem1,Sem2, Sem3, Sem4 were selected.Thus, the dataset was trained using these features. This is applicable for the dataset of both regular and Diploma students.

4.   METHODOLOGY

3 category

The student's SGPI is converted into 3 categories. These categories are:

if sgpi<=4 the 0
If sgpi<=7 then 1
else 2

The categories for semester wise SGPI are predicted by applying ensemble algorithms -Random forest, Bagging and Boosting, rule-based algorithms- Naive Bayes, classification algorithms - decision tree and artificial neural networks- Convolutional neural network and Multilayer Perceptron.

5 category

The student's SGPI is converted into 3 categories. These categories are:

if sgpi<=4 the 0
If sgpi<=5 then 1
If sgpi<=6 then 2
If sgpi<=7 then 3
else 4

The categories for semester wise s are predicted by applying ensemble algorithms -Random forest, Bagging and Boosting, rule-based algorithms- Naive Bayes, classification algorithms - decision tree and artificial neural networks- Convolutional neural network and Multilayer Perceptron.

5. RESULTS

For Regular Students

The results of these algorithms were compared and the results are:

Table -1: Results of Algorithms

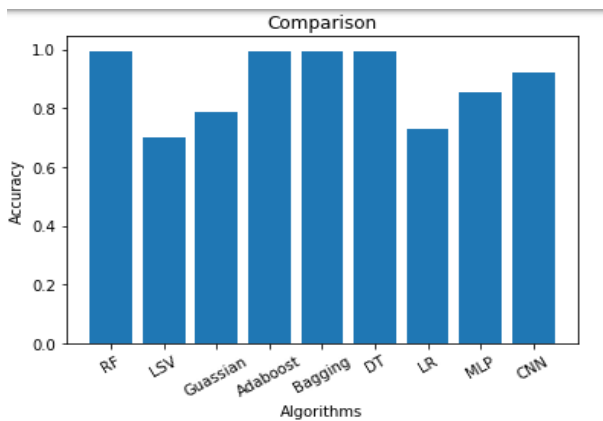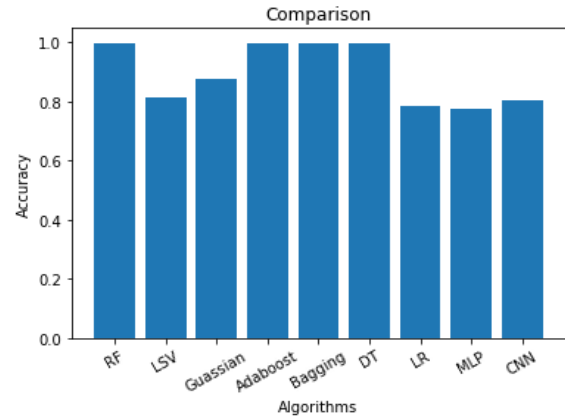| Algorithms	Accuracy	
	3 Category(*100)	**5 Category(*100)**
Random Forest	0.99	0.99
Gaussian Naive Bayes	0.78	0.87
Support vector machine	0.70	0.81
Bagging	0.99	0.99
Boosting	0.98	0.99
Decision Tree	0.98	0.98
Linear Regression	0.73	0.78
MLP	0.85	0.77
CNN	0.92	0.80

Chart -1: For 3 Category

Chart -2: For 5 Category

For Diploma Students

The results of these algorithms were compared and the results are:

Table -1: Results of Algorithm

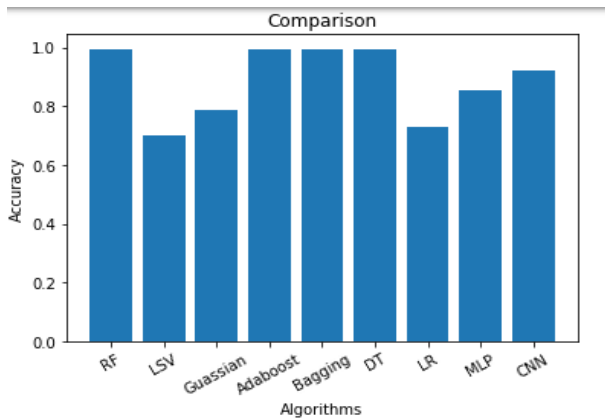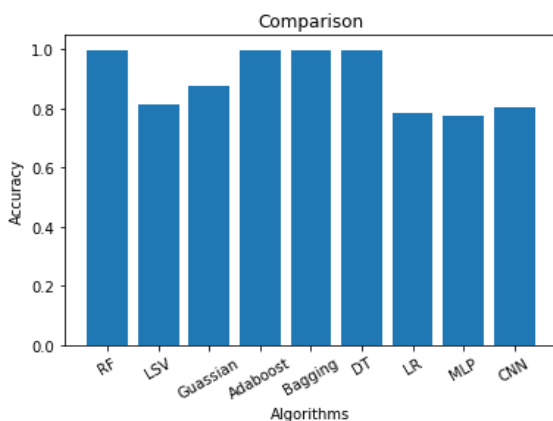| Algorithms	Accuracy	
	3 Category(*100)	**5 Category(*100)**
Random Forest	0.99	0.99
Gaussian Naive Bayes	0.97	0.95
Support vector machine	0.93	0.88
Bagging	0.99	0.99
Boosting	0.98	0.99
Decision Tree	0.98	0.99
Linear Regression	0.73	0.79
MLP	0.62	0.77
CNN	0.88	0.78

Chart -1: For 3 Category

Chart -2: For 5 Category

6. CONCLUSION

In this paper, the classification, rule-based learning, ensemble methods, and neural network based algorithms are employed in student information to predict the students' division on the premise of previous information. The accuracy was calculated using a confusion matrix. The confusion matrix proved to be a good metric for unbalanced data.

Ensemble methods work best with such structured data as compared to its counterpart i.e neural networks. The neural networks require a large amount of data for training.Since the dataset used is less complex and is in a structured format, ensemble methods work best as compared to Convolutional neural networks and Multilayer perceptron for both 3 categories and 5 categories prediction.This study may be quite helpful for both students and teachers for improving the performance and improving the future results of the weak students.

7. ACKNOWLEDGEMENT

8. REFERENCES

[1] Prediction of student performance using educational mining by Ms.Tismy Devasia, Ms.Vinushree T, Mr.Vinayak Hegde Department of Computer Science Amrita Vishwa Vidyapeetham University, Mysuru Campus -2016 International Conference on Data Mining and Advanced Computing.

[2] Proposed academic students' performance prediction model: A Malaysian case study.

[3] Performance Prediction of Engineering students using Decision trees .-R. R. Kabra, R. S. Bichkar -International Journal of Computer Application.

[4] Student performance analysis using classification techniques -International Journal of Pure and Applied Mathematics

[5] Data Mining Techniques for Predicting Student Performance -2015 IEEE International Conference on Engineering and Technology (ICETECH), 20th March 2015, Coimbatore, TN, India

[6] GritNet: Student Performance Prediction with Deep Learning -2018 Educational Data Mining Organisation,Mountain View , California

[7] Students' Performance Prediction Using Deep Neural Network -International Journal of Applied Engineering Research ,Volume 13, Number 2 (2018)

[8] Student Performance Prediction using Machine Learning -International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 IJERTV4IS030127 www.ijert.org (This work is licensed under a Creative Commons Attribution 4.0 International License.) Vol. 4 Issue 03, March-2015

[9] Early Prediction of Students Performance using Machine Learning Techniques-International Journal of Computer Applications (0975 – 8887) Volume 107 – No. 1, December 2014

[10] Using Machine Learning to Predict Student Performance-University of Tampere Faculty of Natural Sciences Software Development M. Sc. Thesis Supervisor: Jorma Laurikkala,June 17

[11] Machine Learning Approach to Predict Student Academic Performance-International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887 Volume 6 Issue IV, April 2018

[12] Predicting Students' Academic Performance in the University Using Meta Decision Tree Classifiers -Authors : Shanthini, A., G. Vinodhini and R.M. Chandrasekaran

[13] Machine Learning Based Student Grade Prediction: A Case Study -Authors : Zafar Iqbal , Junaid Qadir, Adnan Noor Mian and Faisal Kamiran