EARLY PREDICTION OF ACADEMIC PERFORMANCE USING MACHINE LEARNING APPROACH

Manisha Bhimrao Mane
ME Student, Department of Computer Engineering, D.Y Patil College of Engineering, Ambi, Pune,

Dhanashri Kulkarni
Professor, Department of Computer Engineering, D.Y Patil College of Engineering, Ambi, Pune.

Abstract

Predicting Students performance heretofore can be extremely helpful for educational institutions to improve their instructing quality. This paper proposes to predict students performance by thinking about their scholarly subtleties. Educational associations are extraordinary and assume utmost significant job for the improvement of any country. As Education changes the lives of people, families, networks, social orders, nations and at last the world! This is the reason we live agreeable lives today. Presently a day’s training isn’t restricted to just the homeroom instructing however it goes past that like Online Education System, Web-based Education System, Seminars, Workshops, MOOC course turns into It’s all the more testing to Predict understudy’s performance in view of the colossal main part of information put away in the conditions of Educational databases, Learning Management databases. Students’ performance can be assessed with the assistance of different accessible techniques. It is advancing territory of concentrate that accentuation on different strategies like characterization, prediction, include determination. It is utilized on learning records or information identified with training to predict the students’ performance and learning conduct by extricating the hidden knowledge.

Keywords: —College Education, Machine Learning, Result Prediction, supervised learning.

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 data-set. 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 data-set. The rest of the paper provides an explanation on nature of neural networks along with the results of our evaluation. Machine learning is used for analyzing data based on past experience and predicting future performance. Reinforcement machine learning algorithms is a branch of artificial intelligence. It automatically determines the behaviour of environment and maximizes its performance.

1.1 Motivation

We will be focusing on the improvement of Prediction classification techniques which are used to analyze the skill expertise based on their academic performance by the scope of knowledge. Measuring student performance using classification technique such as decision tree. The task can be processed based on the several attributes to predict the performance of the student activity respectively.

1.2 Problem Statement

In any form of higher education it is necessary to predict a student’s academic performance. There are two reasons for this: it is essential to identify which set of students would do well in semester end examination so that they can be awarded scholarships and more importantly to identify the students who may fail in semester end examinations viewed by user. The current system is maintaining academics records manually. Manual maintenance of records involves burden and it is quite tedious task. In general existing system there is no security.

2. Related Work

In the [1] work, author has used DT and BN classes of MLAs for predicting the undergraduate and post graduate results of two universities in Thailand. The total number of student records used for this prediction is 20492 and 932 respectively. Algorithms used for this prediction are C4.5, MSP and Naive Bayes. They concluded that for all classes of predictions DT yields better results than BN by 3 to 12 percent. Re sampling was used to improve the prediction accuracy.

In the [2] work, Kotsiantis et al [4] described a model to predict student results for a distance learning course in Hellenic Open University. Predictions were done on the basis of marks
obtained in written assignments. The algorithms used for this prediction are C4.5, Naïve Bayesian Network (NBN), Back Propagation (BP), 3-Nearest Neighborhood (3-NN) and Sequential Minimal Optimization (SMO). A set of 510 students of the university was chosen for experimental purpose. It was found that the NBN algorithm generates the best results.

In the [3] work, Rama-swami et al [5] developed a predictive data mining model for student performance to identify the factors causing poor performance in higher secondary examination in TamilNadu. A data set for 772 students collected from regular students and school offices were used for this prediction. Algorithm used for this prediction is Chi-Square Automatic Interaction Detection (CHAID) DT. This tree was used to generate a set of decision rules used for predicting student grades.

In the [4] work, Menaei-Bidgoli et al[6] applied data mining algorithms on “logged data” in a educational web based learning system. The system was tested with a data set of 227 students enrolled in a physics course in Michigan State University. Classification was initially performed using Quadratic BN, 1-NN, Prazen Window, Multi layer Perception (MLP) and C5.0 DT. It was seen that combining these classifiers increases prediction accuracy. Genetic Algorithms (GA) were further used to improve prediction accuracy by 10 percent.

Kovacic [7] explores the “socio-demographic” and “study environment” factors that results in student dropout in a polytechnic college in New Zealand. He uses student enrollment data like age, gender, ethnicity for this purpose. The total number of student records used for purpose was 450. Algorithms used for this prediction are CHAID and Classification and Regression Trees (CART). It was found that CART obtained a higher degree of accuracy (60.5). Based on the results of Confusion Matrix and ROC curve he concluded that decision trees based on enrollment data alone are not sufficient to classify students accurately.

Karamohzis and Vrettos [8] have used ANN for predicting student graduation outcomes at Waubonsee College. The prediction model was constructed from a profile of 1407 students of which 1100 were used for training and 307 were used for testing purpose. The average predictive efficiency for training and test sets were 77 percent and 68 percent respectively.

3. Proposed Method

This approach is based on machine learning so we work on the parameters specified by teacher for individual students.

Admin will be the central authority for User activation purpose then once teacher logged in, he/she can update the each and every student’s records and can generate the results as per parameters.

Machine learning and Deep learning classifiers are given to the data set collected from educational environments. Data is preprocessed and check for missing values. Classifiers are applied on the data set to build the models. Models are tested with test data to predict the students’ performance and the best models yielding high accuracy are considered.
4.1 Naive Bayes Algorithm:

Naive Bayes algorithm is the algorithm that learns the probability of an object with certain features belonging to a particular group/class. In short, it is a probabilistic classifier. The Naive Bayes algorithm is called ‘naïve’ because it makes the assumption that the occurrence of a certain feature is independent of the occurrence of other features. Here we classify the heart disease based on heart checkup attributes. Naive Bayes or Bayes’ Rule is the basis for many machine learning and data mining methods. The rule (algorithm) is used to create models with predictive capabilities. It provides new ways of exploring and understanding data.

Why to prefer naive Bayes implementation:

• When the data is high.

• When the attributes are independent of each other.

• When we expect more efficient output, as compared to other methods output. Based on all these information and steps we classify to predict the heart disease depending on heart checkup attributes.

Steps:

1. Given training dataset D which consists of documents belonging to different class say Class A and Class B

2. Calculate the prior probability of class A=number of objects of class A/total number of objects Calculate the prior probability of class B=number of objects of class B/total number of objects

3. Find N1, the total no of frequency of each class Na=the total no of frequency of class A Nb=the total no of frequency of class B

4. Find conditional probability of keyword occurrence given a class:

\[ P(\text{value } 1/\text{Class } A) = \frac{\text{count}}{\text{ni}(A)} \]

\[ P(\text{value } 2/\text{Class } A) = \frac{\text{count}}{\text{ni}(A)} \]

\[ P(\text{value } n/\text{Class } A) = \frac{\text{count}}{\text{ni}(A)} \]

5. Avoid zero frequency problems by applying uniform distribution


7. Assign document to class that has higher probability.

4.2 Mathematical Model

Mathematical equation:

The algorithm implemented in this project is describe as:

Algorithm \[ P(\text{class/features}) = P(\text{class}) * P(\text{features/class}) P(\text{features}) \]

\[ P(\text{class/features}) : \text{Posterior Probability} \]

\[ P(\text{class}) : \text{Class Prior Probability} \]

\[ P(\text{features/class}) : \text{Likelihood} \]

\[ P(\text{features}) : \text{Predictor Prior Probability} \]

A. Normal distribution

The probability density of the normal distribution is:

\[ f(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]

Conclusion

Machine learning techniques can be useful in the field of student’s performance prediction considering that they helps to identify from the beginning of academic year. The aim of this paper is to apply machine learning algorithms for prediction of student performance. An early analysis of student having poor performance helps the management take timely action to improve their performance through predicting their academic details. Accurately predicting student performance based on their ongoing academic records is predicted. Also we conclude that proposed system is helping us to make the student performance better. In this paper machine learning can prove to be powerful tool and all algorithms we used increases with increase in data set size.

References


[9] Stefanowski, J., An Experimental Study of Methods Combining Multiple Classifiers - Diversified both by Feature Selection and Bootstrap Sampling.

[10] Zhao, Y., and Zhang, Y., Comparison of decision tree methods for finding active objects.