International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/IJSRCSEIT

Students Performance Prediction in Online Courses Using Machine Learning Algorithms

Guntumadugu Sravani[1], Kattamanchi Nagendra Rao[2]
M. Tech Student[1], Associate Professor[2]
Department of Computer Science, Chadalawada Ramanamma Engineering College, Andhra Pradesh, India

ARTICLEINFO

Article History:

Accepted: 01 April 2023
Published: 12 April 2023

Publication Issue
Volume 10, Issue 2
March-April-2023

Page Number
765-771

ABSTRACT

Online learning has attracted a large number of participants because it has no limit to enrolment and regardless of personal background and location. Predicting academic performance is an important task for the students in university, college, and school, etc. Machine Learning is a field of computer science that makes the computer to learn itself without any help of external programs. The dataset used in this project is stored in a SQL database and accessed using queries as and when required. There are two approaches for machine learning techniques one is supervised learning and the other one is unsupervised learning. In unsupervised learning, K-means clustering are being used and in supervised, ensemble techniques like Random Forest and XG Boost algorithm are implemented. Nowadays evaluating the student performance of any organization is going to play a vital role to train the students. All of the above algorithms were combined and used for student evaluation and a possible suggestion to the student is provided to improve their career.

Keywords : Predicting Academic Performance of Students, Machine Learning, K-Means, XG Boost, Random Forest, Decision tree Ensemble method.

I. INTRODUCTION

Student's academic performance is a crucial part of an academic institution. This is considered as one of the important measures for many superior universities. Some researchers stated that the student's academic performance can be measured through learning assessment and co-curriculum activities. Though, the majority of researchers have mentioned that the student's past performances, achievements, and grades can play a vital role to predict the student's success rate. Online learning provides lecture videos, online assessments, discussion forums, and even live video discussions via the internet. Its environments have presented convenient learning opportunities and enormous learning resources for various types of participants from all over the world.Predominantly, most of the

higher level institutions use grade as the main measure to assess student's performance. In addition, course structure, student behaviour and extracurricular activities will affect the student's academic performance. The student's academic program can be well planned during their sophomore period of studies in an institution to analyse the performance of students.

At present, machine learning algorithms are most popular to evaluate student's academic performance that has been extensively applied in the education sector. The topic of explanation and prediction of academic performance is widely researched. The ability to predict student performance is very important in educational environments. Increasing student success is a long-termgoal in all academic institutions. If educational institutions can predict students' academic performance early before their final examination, then extra effort can be taken to arrange proper support for the low performing students to improve their studies and help them to success.On the other hand, identifying attributes that affect course success rate can assist in courses improvement. Newly developed web-based educational technologies and the application of quality standard offer researchers' unique opportunities to study how students learn and what approaches to learning lead to success.

II. RELATED WORK

[1] Karimi, Hamid et al. "A Deep Model for Predicting Online Course Performance." (2020).

Online learning has attracted a large number of participants because it has no limit to enrolment and regardless of personal background and location. One of main goals of education is improving students' learning gain. However, the completion rates for online learning are notoriously low. We focus on predicting students' learning performance early and help instructors to provide intervention in-time. We propose a deep online learning performance prediction model incorporate clickstream and demographic data of students. The experiments on the Open University Learning Analytics Dataset (OULAD) show that fusion of learner demographic information can make up for inadequate online learning behaviour data early and improve prediction performance. And our model can achieve reliable performance both in intra-course and inter-course outcome prediction.

Summary: This journal discusses about scoring and performance predictions in online courses.

[2] Sharma, Himani & Kumar, Sunil. (2016). A Survey on Decision Tree Algorithms of Classification in Data Mining. International Journal of Science and Research (IJSR). 5.

As the computer technology and computer network technology are developing, the amount of data in information industry is getting higher and higher. It is necessary to analyze this large amount of data and extract useful knowledge from it. Process of extracting the useful knowledge from huge set of incomplete, noisy, fuzzy and random data is called data mining. Decision tree classification technique is one of the most popular data mining techniques. In decision tree divide and conquer technique is used as basic learning strategy. A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node. This paper focus on the various algorithms of Decision tree (ID3, C4.5, CART), their characteristic, challenges, advantage and disadvantage.

Summary: In this paper, we learn about Decision Tree, types of Decision tree (ID3, C4.5, CART etc). It also discusses about the advantages and disadvantages of Decision Tree.

[3] Kaushik, Manju & Mathur, Bhawana. (2014). Comparative Study of K-Means and Hierarchical Clustering Techniques. International journal of Software and Hardware Research in Engineering. 2. 93-98.

Clustering is a process of keeping similar data into groups. Clustering is an unsupervised learning technique as every other problem of this kind; it deals with finding a structure in a collection of unlabeled data. Many types of clustering methods are— hierarchical, partitioning, density –based, model-based, grid –based, and soft-computing methods. In this paper compare with k-Means Clustering and Hierarchical Clustering Techniques. Strength and weakness of both Clustering Techniques and their methodology and process.
Summary: In this paper, we learn clustering algorithms like Kmeans and Agglomerative clustering and their comparisons.

[4] Kabakchieva D (2012) Student performance prediction by using data mining classification algorithms. IJCSMR 1: 686-690.

This paper presents the results from data mining research, performed at one of the famous and prestigious Bulgarian universities, with the main goal to reveal the high potential of data mining applications for university management and to contribute to more efficient university enrolment campaigns and to attracting the most desirable students. The research is focused on the development of data mining models for predicting student performance, based on their personal, pre-university and university-performance characteristics. The dataset used for the research purposes includes data about students admitted to the university in three consecutive years. Several well-known data mining classification algorithms, including a rule learner, a decision tree classifier, a neural network and a Nearest Neighbor classifier, are applied on the dataset. The performance of these algorithms is analyzed and compared.

III. EXISTING METHOD

Earlier works involves using older Machine Learning algorithms like Logistic regression. They suffer from very low accuracies. There are Deep Learning based predictions as well which uses neural networks for prediction but they have high complexities and they fail to individually identify the important features for prediction.

DISADVANTAGES:

- Low accuracy.
- High Variance.
- Incurs bias in classification.
- High Complexity.

IV. PROPOSED SYSTEM

In our research & extensive literature survey, we found that Random Forest, Decision tree works fine for Student's performance prediction with a great accuracy but it can be further increased by other tree based algorithm like XGboost, when tuned properly can generate significant increase in performance. Also, we have used both unsupervised and supervised learning methods. The data is stored in a cloud server which makes it easy to access from anywhere and this system generates predictions and provides suggestions to the student.

Advantages of proposed system:

- Higher Accuracy.
- Low variance in classification.
- Bias due to assumption about dataset are minimum or even nonexistent.
- Low Complexities.
- Easy access of data.

BLOCK DIAGRAM

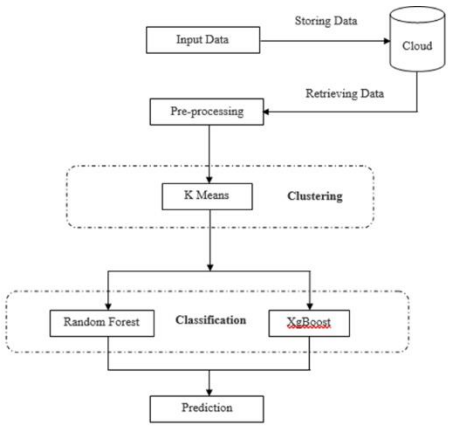
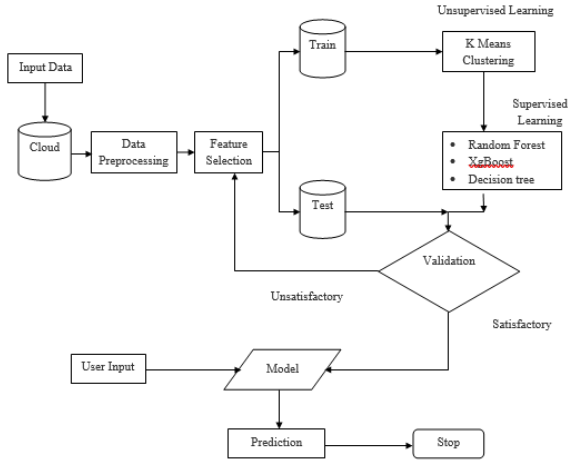


Fig. Block diagram of proposed method

ARCHITECTURE:



Algorithm:

K Means Clustering:

There is an algorithm that tries to minimize the distance of the points in a cluster with their centroid – the k-means clustering technique.

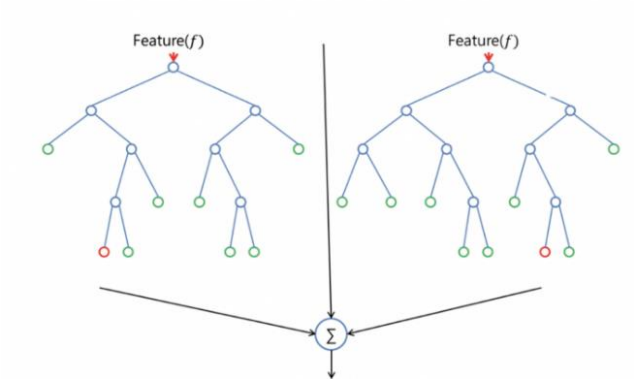
The main objective of the K-Means algorithm is to minimize the sum of distances between the points and their respective cluster centroid.

- Choose the number of clusters k.
- Select k random points from the data as centroids.
- Assign all the points to the closest cluster centroid.
- Recompute the centroids of newly formed clusters.
- Repeat steps 3 and 4.

Random Forest:

Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result.

Random forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because of its simplicity and diversity. It can be used for both classification and regression tasks. One big advantage of random forest is that it can be used for both classification and regression problems, which form the majority of current machine learning systems. Let's look at random forest in classification, since classification is sometimes considered the building block of machine learning. Below you can see how a random forest would look like with two trees:



XgBoost:

XG Boost is the most widely used algorithm in machine learning, whether the problem is a classification or a regression problem. It is known for its good performance as compared to all other machine learning algorithms.

XG Boost or extreme gradient boosting is one of the well-known gradient boosting techniques (ensemble) having enhanced performance and speed in tree-based (sequential decision trees) machine learning algorithms. It is the most common algorithm used for applied machine learning in competitions and

has gained popularity through winning solutions in structured and tabular data. It is open-source software. Earlier only python and R packages were built for XG Boost but now it has extended to Java, Scala, Julia and other languages as well.

XG Boost falls under the category of Boosting techniques in Ensemble Learning. Ensemble learning consists of a collection of predictors which are multiple models to provide better prediction accuracy. In Boosting technique the errors made by previous models are tried to be corrected by succeeding models by adding some weights to the models. Unlike other boosting algorithms where weights of misclassified branches are increased, in Gradient Boosted algorithms the loss function is optimized. XG Boost is an advanced implementation of gradient boosting along with some regularization factors.

Decision Tree

A tree has many analogies in real life, and turns out that it has influenced a wide area of machine learning, covering both classification and regression. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions. Though a commonly used tool in data mining for deriving a strategy to reach a particular goal.

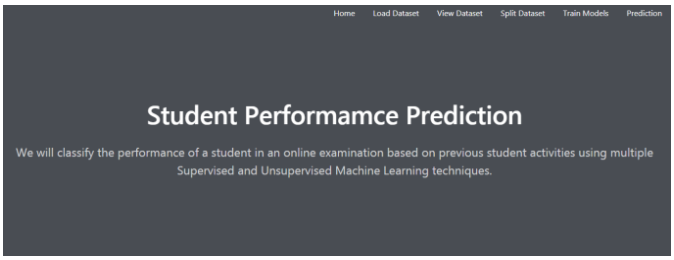
A decision tree is drawn upside down with its root at the top. In the image on the left, the bold text in black represents a condition/internal node, based on which the tree splits into branches/ edges. The end of the branch that doesn't split anymore is the decision/leaf, in this case, whether the passenger died or survived, represented as red and green text respectively.

Although, a real dataset will have a lot more features and this will just be a branch in a much bigger tree, but you can't ignore the simplicity of this algorithm. The feature importance is clear and relations can be viewed easily. This methodology is more commonly known as learning decision tree from data and above tree is called Classification tree as the target is to classify passenger as survived or died. Regression trees are represented in the same manner, just they predict continuous values like price of a house. In general, Decision Tree algorithms are referred to as CART or Classification and Regression Trees.

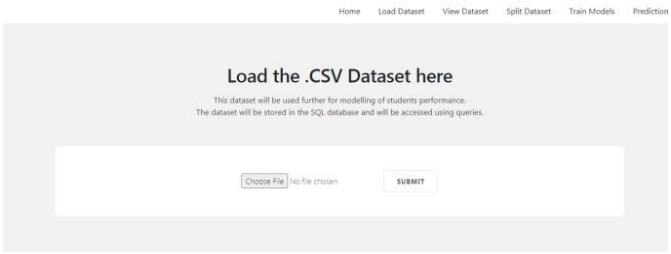
So, what is actually going on in the background? Growing a tree involves deciding on which features to choose and what conditions to use for splitting, along with knowing when to stop. As a tree generally grows arbitrarily, you will need to trim it down for it to look beautiful. Let's start with a common technique used for splitting.

Screen Shots:

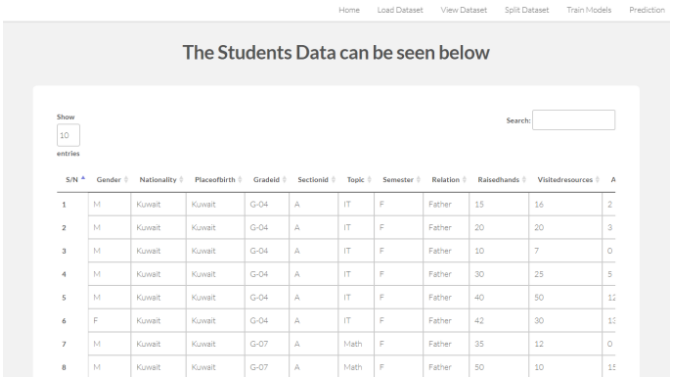
Home Page:



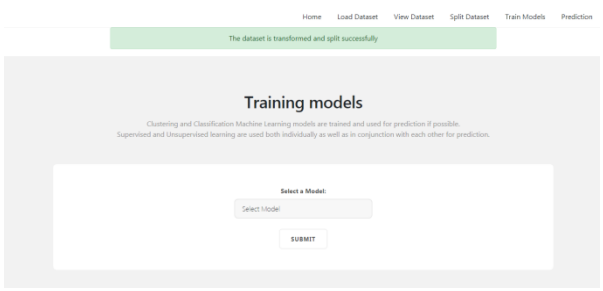
Load Data:



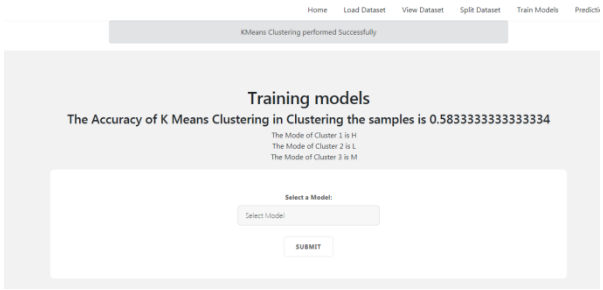
View Data:

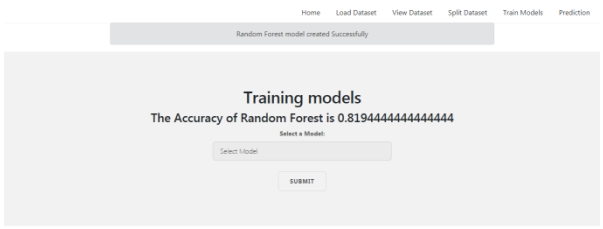


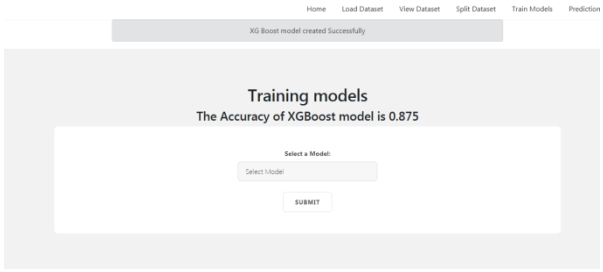
Split Data:



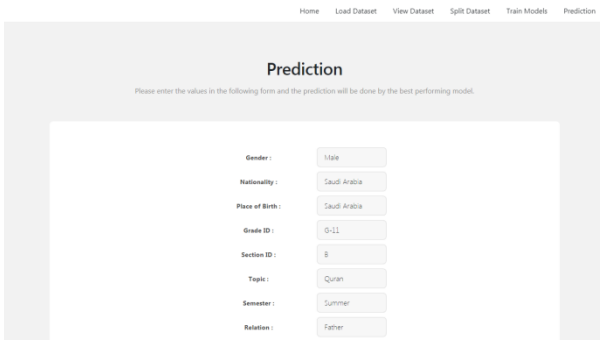
Train Dataset:

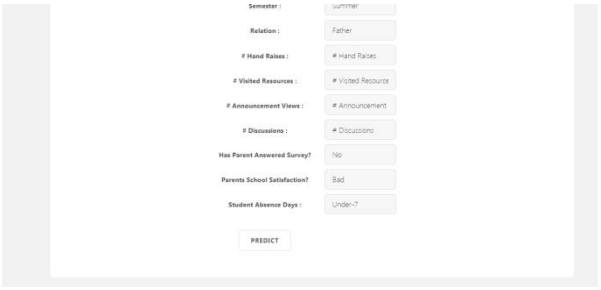


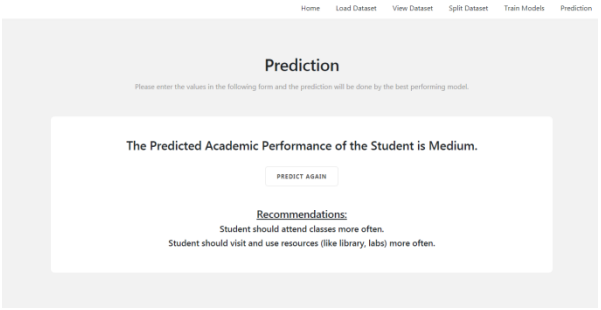




Predictions:







V. CONCLUSION

In this application, we have preprocessed the data by removing the null values and encoding all the variables. We used an unsupervised and 2 supervised learning methods.

KMeans clustering is the unsupervised algorithms which we have used here. Random Forest and XgBoost are the 2 supervised algorithms used for actual classification of the students' performance.

The best model was the XgBoost model with hyper parameters tuning. Clustering algorithms cannot be explicitly used for classification. But, we can use them in conjunction with supervised techniques to be used for prediction. The dataset used was the dataset with students online course performance against their activities previously.

VI. FUTURE SCOPE

We should consider students' performance prediction using neural network which are high in complexities but offers high accuracy and automation of feature selection. Neural networks (even though a bit complex), can provide huge performance gains in students' performance classification.

VII. REFERENCES

[1]. Karimi, Hamid et al. "A Deep Model for Predicting Online Course Performance." (2020).

[2]. Kaviani, Pouria &Dhotre, Sunita. (2017). Short Survey on Naive Bayes Algorithm. International Journal of Advance Research in Computer Science and Management. 04.

[3]. Sharma, Himani & Kumar, Sunil. (2016). A Survey on Decision Tree Algorithms of Classification in Data Mining. International Journal of Science and Research (IJSR). 5

[4]. Altaher A, BaRukab O (2017) Prediction of student's academic performance based on adaptive Neuro-fuzzy inference. IJCSNS 17: 165-169.

[5]. Kaushik, Manju & Mathur, Bhawana. (2014). Comparative Study of K-Means and Hierarchical Clustering Techniques. International journal of Software and Hardware Research in Engineering. 2. 93-98.

[6]. Kabakchieva D (2012) Student performance prediction by using data mining classification algorithms. IJCSMR 1: 686-690.

[7]. Baker, Ryan SJD, Yacef K (2009) The state of educational data mining in 2009: A review and future visions. JEDM 1: 3-16.

[8]. Ramesh V, Parkavi P, Ramar K (2013) Predicting student performance: A statistical and data mining approach. IJCA 63: 35-39

[9]. R. R. Kabra,R. S. Bichkar, "Performance Prediction of Engineering Students using Decision Trees", International Journal of Computer Applications (0975 – 8887), Volume 36– No.11, December 2011

[10]. Ajay Kumar Pal, Saurabh Pal, "Data Mining Techniques in EDM for Predicting the Performance of Students", International Journal of Computer and Information Technology (ISSN: 2279 – 0764),Volume 02– Issue 06, November 2013

[11]. Agavanakis, Kyriakos &Karpetas, George & Taylor, Michael & Pappa, Evangelia &Michail, Christos &Filos, John &Trachana, Varvara &Kontopoulou, Lamprini. (2019). Practical machine learning based on cloud computing resources.

[12]. Santhanam, Ramraj&Uzir, Nishant & Raman, Sunil & Banerjee, Shatadeep. (2017). Experimenting XGBoost Algorithm for Prediction and Classification of Different Datasets.