Advanced Prediction of a student in a university using Machine Learning techniques

Bannet Tumuhimbise
M.Tech(CSE) JNTUACEA,Anantapur Andhra Pradesh, India
Dr. A.P. Siva Kumar
Associate Professor JNTUACEA, Anantapur Andhra Pradesh, India
Chidananda K
Research Scholar JNTUA, Anantapur Andhra Pradesh, India

Abstract—— Prediction of the academic performance of a student is a major element in their education. Nowadays, the education of a student in an organization plays a vital role. Which is difficult to predict manually. We thus opt for machine learning techniques to evaluate student performance. Machine learning which is subpart of Artificial Intelligence that which helps the computer to learn on own without any external support. Machine learning techniques are used to predict the outputs for the certain inputs that are given. There are two approaches in machine learning. They are, supervised learning and unsupervised learning. From supervised learning we are using K-means algorithm and from unsupervised learning we are using XgBoost on of the algorithm from supervised learning and Random Forest as another algorithm to predict students' performance. All these machine learning algorithms are combined for evaluating student performance. And based on the predicted outputs we can provide suggestions to student for better improvements.

Keywords— Predicting student performance, Machine Learning, K-Means, XgBoost, Random Forest

I. INTRODUCTION

Now a days, student's academic performance plays a crucial role in institutions. The performance is considered as one of the important measurements for superior universities. Some researchers stated that the academic performance can be measured through learning assessments and co-curriculum activities. These all includes their performance, achievements and grades, which helps to predict the student's success rate. That which helps in further higher studies.

Predominantly, most of the higher-level institutions consider grades as main measure to assess student's performance. In addition, with the grades course structure, student behavior and extracurricular activities will also impact the academic performance. These all can help a student to plan the academic program and can improve their particular activities.

At present Machine learning algorithms are very popular that can be used for predicting the accurate output based on the past inputs**.** And, in the education systems the prediction of student performance is very important. This can help the educational institutions to fulfill their long-term goals of making a student succeed.

If educational institutes are able to predict the academic performances of student's earlier, extra efforts can be taken to arrange proper support if they are having low performance in the previous academics. Where the attributes that are collected after the performance predictions can be useful to the student to improve in their particular areas where they are little bit down. Also increase in the online classes that which needs the unique opportunities to observe the way of learning by the student and to calculate the approaches of learning that which can lead a student success.

Our proposed model is based on Machine learning technique that can measure the performance of a student. Where, Machine learning considered as the sub part of Artificial Intelligence. Where, it is a process of designing a new algorithm and developing that designed algorithm from which a computer can produce the outputs to the given inputs or the data on its own. There are mainly three types where machine learning algorithms are classified 1. Supervised 2. Unsupervised and 3. Reinforcement Learning. In our project we are using Supervised Learning and Unsupervised Learning algorithms.

Supervised learning is one of the machine learning technique used for the problems of classifications for the labeled data. Supervised learning mainly builds a classification model for the provided data from which the outputs are predicted by the computer on own. The process we follow here is collect the data on which we are performing the algorithm and then select that particular algorithm that is used to build model, then building the model and predicting the outputs with the help of our selected models. In supervised learning we are using XgBoost and Random Forest models.

Unsupervised learning is an algorithm which is performed for unlabeled data. The unsupervised learning is classified into two methods. Those are one of them is principal component and the other one is cluster analysis. Cluster analysis is a method which is used to group datasets. In our proposed method, we are using K-means clustering model. Where this model is performed on the data which is unlabeled.

The objective of this research is to improve and validate the accuracy of the student performance predictions by considering other pre selection factors like communication skills, extra-curricular activities, student personality and student family alongside the already identified selection criteria before selection them. This research also seeks to explore the application of clustering data mining techniques like K-means algorithm to multiple data sets of student data based on pattern mining. The results of this new technique will be compared and contrasted with the results of previous findings from other data mining techniques so as to identify the most suitable and accurate technique to be used.

In this framework, Introduction was discussed in section 1. Related work will be discussed in section 2, section 3 describes about our proposed methodology, where section 4 provides results of our work and we will conclude the paper in section 5.

II. RELATED WORKS

From [1] student's academic performance is predicted using mainly two classifications algorithms of machine learning. One of the algorithms is Support Vector Machine (SVM) and other one is Naive Bayes. As, predicting academic performance is important to students. Here we are considering some of the factors through which we can calculate the predictions. Once the predictions are calculated the performance is shared with class teacher by which the teacher can guide student with particular suggestions given by teacher and that helps in increasing the performance.

In this [2] author discuss about an algorithm of Naive Bayes from which they give a brief description about the model and the concept. The concept includes of hidden bayes, text classification, and traditional bayes and about also includes about the machine learning. Also, discuss about augmented bayes by considering with some of the examples. In the end, some advantages, disadvantages and applications are also discussed for better understanding of the algorithm and also discussed how it can be used in predicting student performance. Basically, this naive bayes is a classification algorithm that which is performed with the help of bayes theorem that which consists of assumptions which are independent. Hence, these independent assumptions help the algorithm to perform more accurately.

In [3] authors give brief description about a classification technique named Decision tree which is considered as one of the popular techniques for data mining. Now a days, the amount of data information industry is increasing more because of the computer technology. Where, the analyzing this large data is very important for the extraction of useful data from it. For this process of extraction of the useful data by deleting the extra noises we are using the decision trees. A decision tree is a type of tree structure that which consists of a roots, branches, and leaves. Where nodes are attached with those branches and leaves. Where that node is used for testing the given input or and the outputs of that nodes are given as the input to the nodes of trees

From [4] authors addressed the Adaptive Neuro Fuzzy Inference Method (ANFIS) for student academic success prediction to help students enhance their academic achievements. The solution is made up of two stages. First, students' results in previous exams are pre-processed by standardizing their results for increasing the accuracy of the prediction. Second, ANFIS is being used to estimate the anticipated output of students in the next semester. In this research work, three ANFIS models vz. ANFIS-GaussMF, ANFIS-TriMF and ANFIS-GbellMF that used different membership functions to produce reliable fuzzy rules for the student's success prediction mechanism are used.

In this [5] authors discuss about clustering that which can helps in student performance prediction. Where clustering is considered as one of the types that which can carry or hold the same type of data. Clustering is one of the unsupervised learning techniques that which is used to find the structures or the particular data of the unlabeled set or data. There are different forms of methods in clustering. Namely, hierarchical clustering, partitioning type of clusters, density-based clustering, model-based clustering, grid based, and soft computing clustering. Here, comparison is takes place between k means and hierarchical. We are comparing mechanism, how both the data set and the clustering process should operate.

The authors [6] are mainly focused on developing the models of data for student success predictions based on the some of the information given in the dataset of the student that which may include regarding their previous studies, their activities in the institution and their personal behavior and way of act in the organization. Here mainly they used neural networks and decision tree algorithms that which are considered as one of the best data mining techniques.

.

Author & Year	**Proposed**	**Finding/Outcomes**
Vairachilai S, Vamshidhar Reddy, 2020	An Machine Learning approach for prediction of student's performance	Procedure of using machine learning techniques for student's success prediction
Kaviani, Pouria & Dhotre, Sunita, 2017	A survey regarding the algorithm of Naive Bayes	It explains about the naive bayes and how it performs the clustering in Machine Learning
Sharma, Himani & Kumar, Sunil,2016	Data mining techniques for the classification using decision tree algorithm	Data mining techniques for the classification using decision tree algorithm
Altaher A,	Predicting student's	Predicted student's

BaRukabO, 2017	academic performance based on adaptive Neuro-fuzzy inference	academic performance using adaptive Neuro- Fuzzy interference
Kaushik, Manju, Mathur, Bhavana, 2014	Clustering techniques like K means and hierarchical discussion	Clustering techniques like K means and hierarchical discussion
Kabakchieva D, 2012	Data mining techniques usage for student performance prediction	Data mining techniques usage for student performance prediction

Table 1: Related Works Summary

III. PROPOSED METHODOLOGY

The procedure to develop our system is clearly described in this section.

- First, to predict the student's results, we need to collect the necessary dataset. Source of our dataset is from a college or university that which is included with the some of the information of the student that may include includes gender, nationality, birthplace, stage, grade, section, topic which they are studying, semester marks, raised hands, visited resources, announcement's view, discussion, parent response survey, parent school satisfaction, student absence days and class.
- The entire dataset is uploaded to the cloud server.
- Once the database is uploaded to the cloud, we will be able to access the database via the queries for the next step.
- Once the data is read, the pre-processing of the data is carried out. We're eliminating the Null Values here, and we're going to do the label encoding. This is the pre-process of the dataset considered.
- Once the preprocessing is completed, we divide data into two parts. One part is for training and other part is for validation purpose.
- After splitting, we will train the data for further processing. Here, we use machine learning algorithms for training. In unsupervised learning we choose K-means and in supervised learning, we choose XgBoost and Random Forest.
- After we have completed the training, the student success forecast is shown on the basis of the training phase.
- After the prediction, we can offer some suggestions to the student to further progress their results.

The architecture of our proposed model is shown in below block diagram.

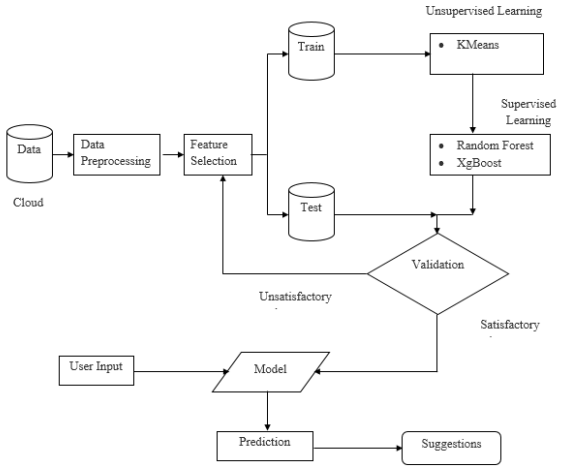

Fig 1: Proposed model Architecture

K-means Clustering

K-means is one of the unsupervised algorithms of machine learning that which is mainly used for clustering. It is considered as the simple algorithm and the commonly used one. Where the unsupervised can work on the inputs that which consists of the unlabeled data or the input.

K-means Clustering is mainly used for identifying the observations that which belongs to the particular group. Here they firstly the partitioned data is given from which it will classify that the particular value belongs to the type of group or the subset. Signal processing is said to be the origin for this clustering model. Here, for the observing the groups or subsets an algorithm is used that which depends on the Euclidean distance calculation. Here, at first a centroid is formed from where the calculation is performed and then the distance is calculated.
Once the centroid is formed, it is given to the k points that which are taken from each subset or group that was divided at the initial. Now from the centroid to the k point the distance calculation is performed. Now the average of all the data points that are calculated are considered and then it will check for the minimal range that which portioned data belongs to which sub group. Then at the end we have our clusters that are combined with their particular sub groups.
The k means algorithm of the unsupervised learning is also classified as nearest neighbor classifier because that will check for the nearest neighbors of the calculated centroid and it classifies the particular group that which belongs to the partitioned one. Where, k is given as a value. For example, we can consider k as 1 that which means for the 1 nearest neighbor. Hence, this algorithm is also named as nearest centroid classifier because mainly algorithm is based on the nearest centroids for the correct classification.

Xg Boost Algorithm
XG Boost is one of the mainly used algorithm in machine learning, whether the problem is a classification or regression problem. It is recognized for its good performance relative to all other machine learning algorithms.
XG Boost is also called as extreme gradient boosting. It is one of the algorithms that which performs operations very quickly. It is a machine learning algorithm that which performs fast and can produce the more accurate values for the considered inputs. This is mainly used for the labeled data. Where, many numbers of different features are given to the data. It performs upon all the featured data and then it provides the high prediction accuracy. Where this is able to perform different tasks. Those are regression, classification and ranking. These are the main three tasks which can be performed by the gradient boosting or XgBoost.
XG Boost will comes in the category of Ensemble Learning Techniques. Where, this learning technique consists of the predictions as a collection that which are used for providing the accuracy when this algorithm is used for the particular dataset. In this technique errors of the previously performed models are considered to which some of the weights are applied by which the present model can rectify those errors and could not perform it twice or can reduce those errors in the present model which we are using. This is an advanced technique when compared with the previous boosting techniques.
XgBoost technique consists of three steps mainly;

1.An initial model X_train is defined to predict the target variable y. The model is then associated with a residual (y-X_train)

2.Then, a new model X_test is fit to the residuals from the previous step

3.Now, X_train and X_test are combined to give y_pred, the boosted version of X_train. The mean squared error from y_pred will be lower than that from X_train:

$$y\_pred_1(x) <- X\_train_0(x) + X\_test_1$$

4.To improve the performance of y_pred, we modeled after the reesiduals of y_pred and created a new model y_pred_2:

$$y\_pred_2 <- X\_train_1 + X\_test_2$$

This is done for several iterations (z), until residuals have been minimized as much as possible thus improving the accuracy.

Algorithm of Random Forest
This is considered as the supervised learning model of the machine learning. Here the name "forest" that creates a collection of trees that are named as decision trees typically completed the process of training part using the "bagging" process. The principle of this baggage approach is a mixture of models from the learnings that helps in improving the results.
It is capable of the operations of labeled data, one of the easy usable algorithms of the machine learning that which provides accurate and great results at the maximum times. The absence of hyper parameter tuning cannot affect this algorithm and can work properly with more accuracies. It is one of the algorithms that which is used more. This can perform the operations of regression and also the operation of classification. It is very simple to use. Where using for the both regression and classification operations is one of the best advantages of the random forest. That which makes one of the popular more usable algorithm. Let us look visualization of random forest from the below diagram.

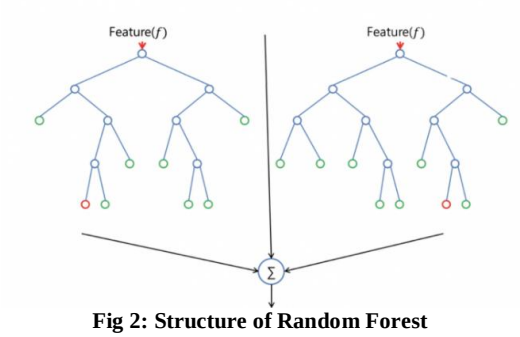

Fig 2: Structure of Random Forest

It is considered same as the decision tree in which both have same type of hyper parameters. It can also be considered as the baggage classifier. Luckily, you don't have to use decision trees that are to be combined with the baggage. Instead of all this long process we can simply consider the random forest model. By using this model, we can perform the regression operation.

In this algorithm the trees or the nodes that are formed randomly as the number of trees or the branches and the nodes grow. When compared with some of the tree algorithms they choose the features which are very important when the nodes break. But in this they choose the features which are best rather than important.

Hence, it considers only the random subsets of the features during the splitting of the nodes as it considers the best features. Here, we can also use renders which are considered as the type of trees with the help of thresholds which are random for more random divisions.

Another one of the advantages of this forest is that it is straightforward for calculating significance that which is relative with each function to forecast. Sklearn offers a one of the best paths for the calculations of the feature values by observing the features from the nodes of trees. That which is used for the calculation of the features from the obtained nodes. Where here the outputs of the calculated nodes of the trees are given as the input to the tree and again the process continues it measures the outputs.

Random Forest technique basically follows the steps as below;

1.The algorithm selects random samples from the dataset that was uploaded.

2.A decision tree is then created for each sample that was selected. The algorithm then gets a prediction result from each decision tree

3.Voting is then performed for every predicted results. The mode is used when classification is being done, while the mean will be used for regression.

4.Finally, the algorithm selects the most voted prediction result as the final prediction.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this session we will discuss about the experimental results obtained by using Supervised and unsupervised machine learning models and comparison of the accuracies obtained by the models is shown.

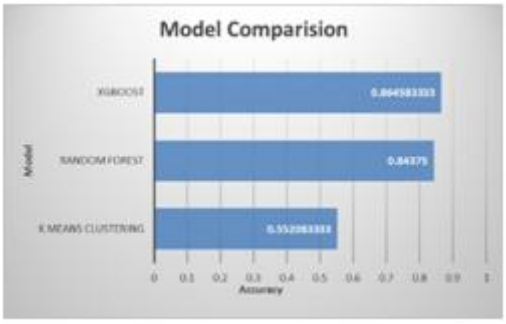

Fig 3. Model Accuracy comparison graph

Model	Accuracy
K Means Clustering	0.552083333
Random Forest	0.84375
XgBoost	0.864583333

Table 2. Model Accuracy comparison graph

From the above graph and table of accuracy comparison of models, we can find the accuracies that are obtained by using the K means clustering from unsupervised learning and Random Forest and XgBoost from the supervised learning of the machine learning.

When compared with K Means clustering and Random Forest, XgBoost provides more accuracy. That means the predictions are given more accurately that are obtained by supervised learning model of XgBoost. Hence, we can use or select the XgBoost for more accurate results and predictions.

Once, the performance is predicted we can provide some suggestions to students based on their performance. Here, we segregated the predicted performances as low, medium and high. If a student gets a low and medium performance then we can give some suggestions to student that which can help a student to further make it as high performance.

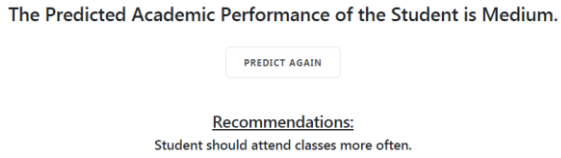

Fig 4. Suggestions if performance is medium

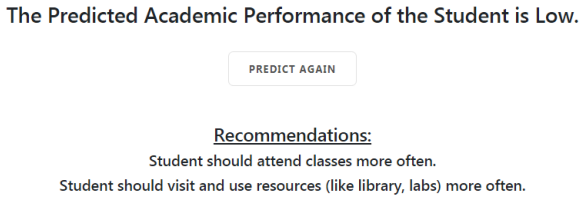

Fig 5. Suggestions if performance is low

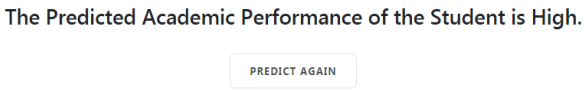

Fig 6. No suggestions if performance is high

From the above shown figures, we provide suggestions to the student if their predicted performance is low or high to improve their performance in the upcoming academics. Where, there is no need of suggestion if student performance is high that which means student is already performing greatly.

V. CONCLUSION

This study provides performance prediction of a student using supervised and unsupervised learning methods of machine learning model. Where, we used Random Forest and XgBoost models from the supervised learning and K Means model from unsupervised. Once the performance is predicted by using our three models we check for the accuracy of the models. Performance is predicted by using the model that which gets more accuracy. In our model XgBoost provides more accuracy when compared with K Means and Random Forest.

After the performance prediction, we provide some suggestions to the students with low and medium performance that which can help student in future to increase their performance.

VI. REFERENCES

[1] Vairachilai S, Vamshidharreddy, "Student's Academic Performance Prediction Using Machine Learning Approach", IJAST, vol. 29, no. 9s, pp. 6731 - 6737, Jun. 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]. Agaoglu M (2016) Predicting instructor performance using data mining techniques in higher education.
[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] Abeer Badr El Din Ahmed1, Ibrahim Sayed Elaraby, "Data Mining: A prediction for Student's Performance Using Classification Method", World Journal of Computer Application and Technology2(2): 43-47, 2014, DOI: 10.13189/wjcat.2014.020203