Advanced Prediction of Performance of Student in a University using Machine Learning Techniques

Shreya Singh[1], Aditi Bhardwaj[2], Gazala Malik[3]
[1,2,3]*Student, Meerut Institute of Engineering and Technology, Meerut*

Abstract **- Machine Learning is a field of computer science that makes the computer to learn itself without any help of external programs. These machine learning techniques can be used to predict the output for certain inputs. There are two approaches for machine learning techniques one is supervised learning and the other one is unsupervised learning. In unsupervised learning, K-means and Hierarchical clustering are being used and in supervised, Naive Bayes and Decision Trees are used. 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 used for recruiting process. Here the performance of K L University students is evaluated by applying all the above algorithms.**

Index Terms **- Clustering, Classification, K-Means, Hierarchical, Naïve Bayes, Decision Trees.**

1.INTRODUCTION

Machine Learning is all about designing and developing algorithms through which computers can predict behaviours based on the given data. It is a branch of artificial intelligence. Nowadays in the area of education, there is a significant growth in using these algorithms. These machine learning algorithms mainly focus on pattern recognition and decision making. From the given input data first patterns are recognized, rules are generated. According to that rule, the behaviour is predicted, and decision making is done.[1]

In normal traditional programming, the data and program are given as input to the computer to get the output, which was different from machine learning. Mainly Machine Learning algorithms are classified into three types. They are:

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Here, Supervised and Unsupervised Learning was discussed.

A. Supervised Learning:

Supervised Learning is mainly used in classification problems. The main goal of this learning is to build a classification model through which the computer learns about the data and predicts the output. The main steps involved in supervised learning are collecting the data, selecting an appropriate algorithm, building the model, using the model for prediction.In supervised learning, it takes a set of inputs and known responses to build a model. The model that was built is used for mapping new data to the desired responses or outputs. The supervised learning algorithms also give the probabilities of all the given inputs. There should not be any missing values in the dataset. If there are any missing values, it is not possible to predict the output. The algorithms that belong to this type of learning are Naïve Bayes, Logistic Regression, Decision Trees, Neural Network.

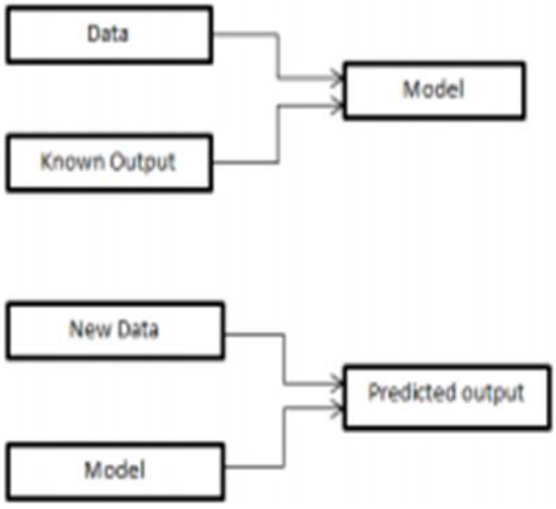


Fig 1: Supervised Learning

Unsupervised Learning:

In Unsupervised learning, only inputs and do not have any outputs. The main aim of this type of learning is to model the structure of the unlabelled data to learn more about the data. As there is no specific output, the data is left to the algorithm itself to analyse the interesting patterns in the data. Here the similarities that present between the data is mainly found. All the outputs are from the same set of latent variables. But in the case of the supervised learning input set of data is the reason for an output set of data. The difference between both types of learning is [2]

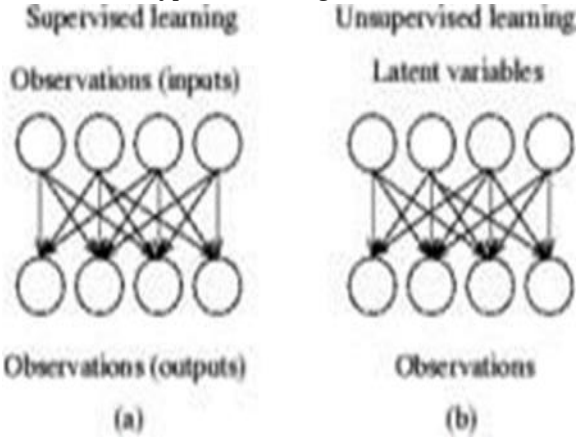


Fig 2: Difference between supervised and unsupervised learning

2.LITERATURE SURVEY

Several methods are introduced by various researchers for prediction of student performance. They are:

Ishwank Singh et al. had proposed a model for analysing student performance based on clustering algorithms. The parameters considered are tenth and inter marks, graduation marks, internships, backlogs and projects. First academic performance of students is analysed by applying k-means clustering on Tenth, Inter and B. Tech and they are compared.

Next co-curricular performance is analysed by applying k-means on projects, internships and skills and they are compared. Finally, both academic and co-curricular performances are compared.[3]

Shaymaa E. Sorour et al. had proposed a new method for predicting the performance of students based on comment data mining. This new method consists of Latent Dirichlet Allocation (LDA) and Support Vector Machine (SVM) techniques and these are used to predict the student grades. Later majority vote method is applied to improve the predicted results and the reliability of predicted grades is also evaluated.[4]

Pratiyush Guleria et al. had used decision trees to evaluate student performance. Here an educational data set is considered and entropy and information gain of all the attributes present in the dataset is calculated. The attribute which consists of the highest information gain is considered as the root node of the tree. This classification algorithm is used to identify students with poor performance.[5]

Zhenpeng Li et al. had proposed a novel approach to predict student performance. This approach consists of fuzzy clustering and multi-variable regression and an offset value mechanism. This method mainly concentrates on the attributes that are related to student courses only. It had analyzed a large number of data which are normally or continuously distributed when compared to other models.[6]

M. Maha lakshmi et al. had proposed a scrum methodology to track the performance of the student in web-based education. In this methodology, the teacher describes the learning objectives, and that teacher is responsible for monitoring the progress. Later evaluation of all the members is done.[7]

Krina Parmar et al. had proposed a method of predicting student performance by using distributed data mining. They had used classification models for prediction and the input was given in a distributed environment. The data was given to several data mining algorithms and several local models were made. And all the local models are aggregated.[8]

Priyanka Anandrao Patil et al. had proposed a model of using the frequent tree to predict student performance. The frequent patterns of the student data are found by applying generalized sequential pattern mining. From that frequent pattern, a tree was built by using frequent pattern tree algorithm. This tree helped predict whether a student will pass or fail in the exam. [9]

Shaymaa E.Sorour et al. had proposed a method on comment data mining to predict the performance of a student. The comments that are given by students for all questions are collected and classified into six attributes. Decision Tree and Random Forest models are applied to the collected data.[10]

3. PROPOSED SYSTEM

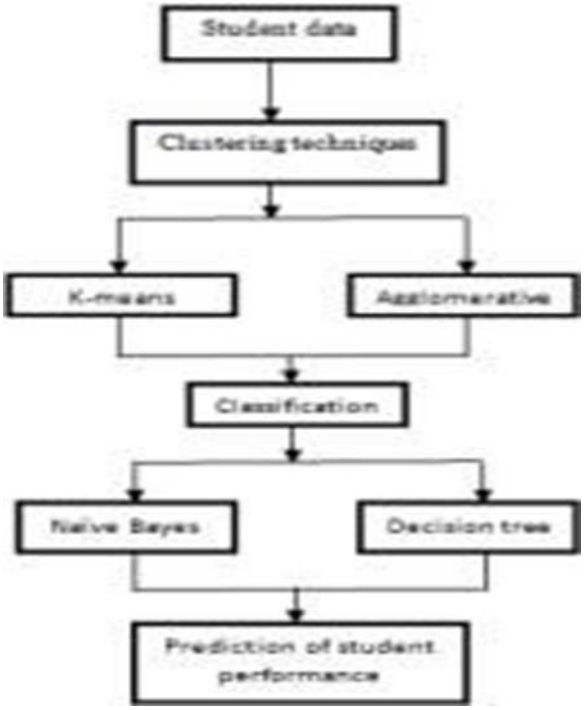


Fig 3: Proposed System to predict the student performance

4.METHODOLOGY

[11] The real data of K L University students who are studying B.Tech (2013-2017 batch) are taken. First based on their Intermediate marks and B.Tech percentage up to 3-2 semester, clustering is performed. [12]Three clusters are formed by using K-means and Hierarchical clustering. Table:1 clusters

Intermediate Marks	Cluster
below 85 85-90 90 and above	Poor Good Excellent

A. K-means Clustering

K-means is one of the unsupervised algorithms which is used to Agglomerative hierarchical clustering is mainly classified into Single linkage clustering, Average linkage clustering, Complete linkage clustering. Here, complete linkage clustering is used. In this algorithm, each point acts as a cluster itself. Depending form clusters in a large dataset. The inputs are the number of clusters to be formed represented by 'k' and the dataset. Initially, some centroids are randomly chosen and clusters are formed by calculating Euclidean distance between the centroid and the instance. Every instance will be assigned to a cluster where it has the least Euclidean distance. After the formation of initial clusters again centroids are updated based on the mean of the cluster until convergence occurs.

K-means clustering is performed on Inter marks and B.Tech percentage of 200 students. The clusters that are formed are

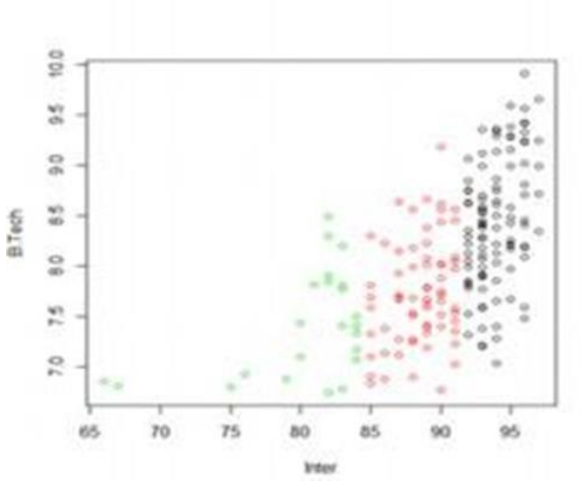


Fig-4: Three clusters formed by K-means

[13][14]The cluster which contains black dots indicate that the student performs excellently in their academics. The cluster which contains red dots indicates that the student performs well in their academics and the cluster which contains green dots indicate that the student performs average poor in their academics.

Clusters	Cluster Size	Cluster Means	
		Inter	B.Tech
1	108	94.0	8.4
2	68	88.5	7.7
3	24	80.3	7.4

Table-2: Clusters size and cluster means formed by k-means

B. Hierarchical Clustering

[16] [17] Hierarchical clustering is classified into two types. They are Agglomerative and Divisive. In Agglomerative small clusters are combined recursively to form large clusters on the distance between the sequential clusters, the clusters that are separated by the shortest distance are merged into a single cluster. Hierarchical clustering is performed on

Inter marks and B. Tech percentage of 200 students. The clusters that are formed are

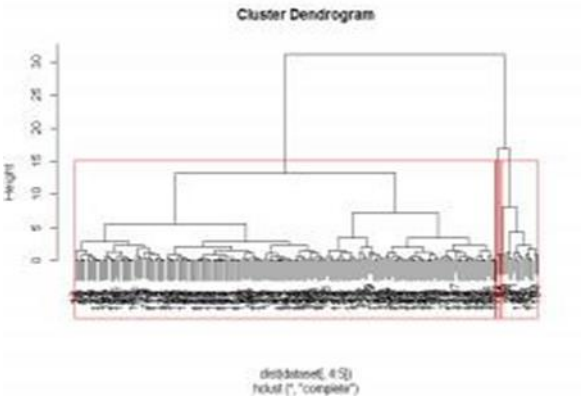


Fig-5: Cluster Dendrogram

While comparing k-means and hierarchical clustering, the performance of k-means is high and hierarchical is moderate. K-means is good with large data sets and hierarchical is good with small data sets. As a maximum number of tuples belongs to one cluster in hierarchical, the clusters that are formed by k means are only considered for further analysis.

The clusters of k-means are divided into training and testing data and are given for classification. Through classification, whether a student will pass or fail in a technical exam conducted by any company as a part of the recruiting process will be predicted. As it is a technical exam the marks of subjects like C-language (C), Operating Systems (OS), Computer Networks (CN) and Database Management Systems (DBMS) are considered. Classification Techniques that are used are Decision Trees and Naïve Bayes algorithms.

C. Decision Trees

The decision tree has a tree structure, where the leaf node represents label, branch represents the outcome and the internal nodes represent the attributes. The root node is the topmost node. ID3, C4.5 and C5.0 are some of the decision tree algorithms. C5.0 is the extension of the C4.5 algorithm and C4.5 is an extension of the ID3 algorithm. C5.0 handles continuous and categorical data along with the date, time and timestamps It deals with missing data.

[18][19]Decision tree classification is performed on technical subjects C, OS, CN, DBMS. The data of 200 students is considered. That data is divided into training and testing data. Training data consists of 100 tuples and testing data consists of 100 tuples. The tree that was formed is The predictions of testing data are 81% (35+46) accurate.

D. Naïve Bayes

It is also a classification algorithm and is used for predictions. Naïve Bayes is a simple classifier that uses Bayesian theorem. Let M be the tuple and N be the hypothesis that A belongs to a certain class. Then Bayesian theorem is represented by

P(N|M) = P(M|N) P(N) / P(M) where

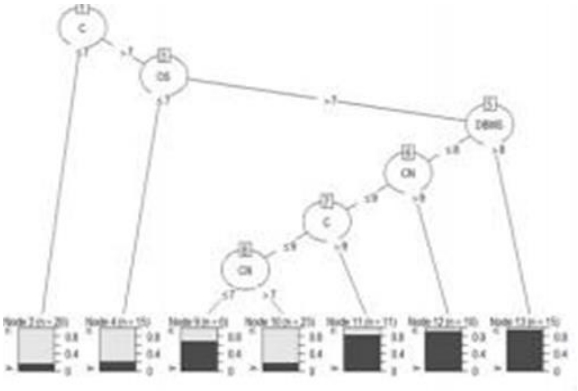


Fig-7: Description of the decision tree

After training of the data is done then testing is performed. By using testing data predictions are done

Predicted type Actual type	N	Y	Row Total
N	35	10	45
Y	9	46	55
Column Total	44	56	100

Table-3: Predictions of decision tree

```
Class specified by attribute 'outcome'

Read 100 cases (5 attributes) from undefined.data

Decision tree:

C <= 7: n (20/3)
C > 7:
:...OS <= 7: n (15/3)
    OS > 7:
    :...DBMS > 8: y (15/1)
        DBMS <= 8:
        :...CN > 9: y (10/1)
            CN <= 9:
            :...C > 9: y (11/2)
                C <= 9:
                :...CN <= 7: y (6/2)
                    CN > 7: n (23/4)
```

P(M/N) is the posterior probability of M conditioned on NP(N) is the prior probability of N P(M) is the prior probability of M

The same data what is used for decision tree is used here for results are compared. Here in the below figure

A, B, C, X, O represents the grades that were scored by the students. O represents the grades that were scored by the students. O represents 10 points. Similarly X=9, A=8, B=7, C=6.

Y= probability to pass in the exam

represents the grades that were scored by the students. O represents 10 points. Similarly X=9, A=8, B=7, C=6.

Y= probability to pass in the exam.

P(N|M) is the posterior probability of N conditioned on M

Fig-8: Apriori and Conditional Probabilities

Predicted type Actual type	N	Y	Row Total
N	40	13	53
Y	15	32	47
Column Total	55	45	100

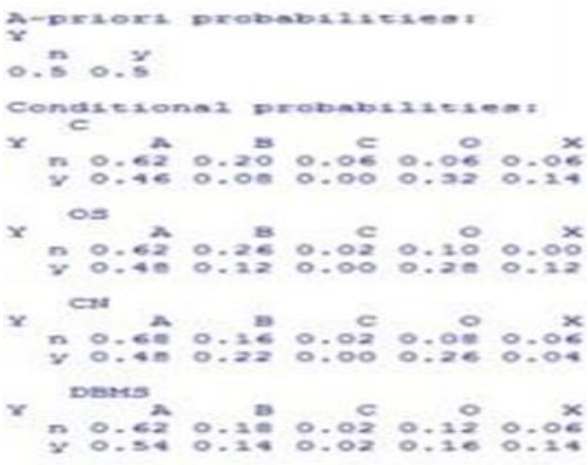
Table-4: Predictions of naïve Bayes

The predictions are 72% (40+32) accurate. While comparing both algorithms decision predicts more accurately than the naïve Bayes algorithm for this dataset.

5.CONCLUSION

In this paper, clustering and classification are applied to a dataset of K L University (2013-2017 batch) students to evaluate the student's performance and had predicted whether a student will pass or fail in a technical exam that was conducted as a recruiting process. The results that are obtained are accurate. Some technical subjects like Computer Networks, Operating Systems, C language and Database Management Systems are only considered for classification. In future, this can be extended with more number of subjects and a large data set.

Here, we have used two supervised and unsupervised algorithm ,k-means is used for the clustering procedure, agglomerative algorithm is used here as the extended version of the k-means algorithm .we have used the random forest algorithm for the advanced version of decision trees to perform the data mining and the xgboost algorithm woks with the highest efficiency and accuracy a form of supervised learning algorithm .Our testing modules help us to select the percentage of data to work with and the training module allows to select the algorithm to work and predict the performance .



6 REFERENCES

[1] [C.M.A.K. Zeelan Basha, T. Maruthi Padmaja and G.N. Balaji, “Automatic X-ray Image Classification System”, Smart Computing and Informatics Proceedings of the First International Conference on SCI 2016, Volume 2

[2] S. Sai Niveaditha, V. Pavithra, R. Jayashree, T. Tamilselvi, “Online diagnosis of X-ray image using FLDA image processing algorithm”. IRF International Conference, March 2014, pp. 76–80.

[3] Jaskirat Kaur, Sunil Agrawal, Renu Wig, “A comparative analysis of thresholding and edge detection segmentation techniques.” International journal of computer applications. Vol. 39, No. 15, pp. 29–34, 2012.

[4] K.Thangavel, R. Manavalan, and I. Laurence Aroquiaraj, Eliminating Speckle Noise from ultrasound medical images: a non-linear approach, Intelligent Computing Models Narosa Publishing House, New Delhi, , 2009 173-180

[5] Jaskirat Kaur, Sunil Agrawal, Renu Wig, "A comparative analysis of thresholding and edge detection segmentation techniques." International journal of computer applications. Vol. 39, No. 15, pp. 29–34, 2012.
[6] S.K. Mahendran, "A comparative study on edge detection algorithms for computer aided fracture detection systems". International journal of engineering and innovative technology. Vol. 2, No. 5, pp. 191–193, November 2012.
[7] Subodh Kumar, Prabat Pandey, "Implementation of X-ray image segmentation by using edge detection using sobel edge operator", International journal of innovative research and studies, Vol. 3, No. 2, pp. 191–202, February 2014.
[8] Rafeal. C. Gonzalez & Woods, "Chapter 10: Image segmentation", Digital image processing, Pearson Education, Inc., 2005, pp. 392–439.
[9] Balaji, G. N., T. S. Subashini, and N. Chidambaram. "Automatic Classification of Cardiac Views in Echocardiogram Using Histogram and Statistical Features." Procedia Computer Science 46 (2015): 1569– 1576.
[10] Basha, C. Z., Krishna, A., & Savarapu, P. R. (2019). Automatic detection of lung infection. International Journal of Recent Technology and Engineering, 8(3), 200–203.
[11] Kim, Seong-Hoon, et al. "X-ray image classification using random forests with local binary patterns." Machine Learning and Cybernetics (ICMLC), 2010 International Conference on. Vol. 6. IEEE, 2010
[12] M.V.Bramhananda Reddy, Varadala.Sridhar, M.Nagendra and Roy, " Dental X-Ray Image Analysis by Using Image Processing Techniques", Volume 2, Issue6.
[13] Ghofrani, Fatemeh, et al. "Fuzzy-Based Medical X-ray Image Classification." Journal of medical signals and sensors 2.2 (2012): 73.
[14] Akshay Trimbak Chikhalekar, "Analysis of Image Processing for Digital X-Ray", International Research Journal of Engineering and Technology (IRJET), Volume: 03 Issue: 05
[15] C.M.A.K. Zeelan Basha, T. Maruthi Padmaja and G.NC.M.A.K. Zeelan Basha, T. Maruthi Padmaja and G.N. Balaji, "Automatic X-ray Image Classification System", Smart Computing and Informatics Proceedings of the First International Conference on SCI 2016, Volume 2
[16] S. Sai Niveaditha, V. Pavithra, R. Jayashree, T. Tamilselvi, "Online diagnosis of X-ray image using FLDA image processing algorithm". IRF International Conference, March 2014, pp. 76–80.
[17] Jaskirat Kaur, Sunil Agrawal, Renu Wig, "A comparative analysis of thresholding and edge detection segmentation techniques." International journal of computer applications. Vol. 39, No. 15, pp. 29–34, 2012.
[18] K.Thangavel, R. Manavalan, and I. Laurence Aroquiaraj, Eliminating Speckle Noise from ultrasound medical images: a non-linear approach, Intelligent Computing Models Narosa Publishing House, New Delhi, , 2009 173-180
[19] Jaskirat Kaur, Sunil Agrawal, Renu Wig, "A comparative analysis of thresholding and edge detection segmentation techniques." International journal of computer applications. Vol. 39, No. 15, pp. 29–34, 2012.
[20] S.K. Mahendran, "A comparative study on edge detection algorithms for computer aided fracture detection systems". International journal of engineering and innovative technology. Vol. 2, No. 5, pp. 191–193, November 2012.