Prediction of Students Performance Using K-Means Algorithm (Machine Learning Techniques)

Mr. Shubham Agrawal[1], Mr. Kapil Sahu[2]
[1]M.tech Scholar, [2]Asst. Professor
[12](CSE) SIMS INDORE

Abstract- Machine Learning is a field that is used in every system. Machine learning is used in educational system, In pattern recognition, Games, Industries.In education system its importance becomes more because of the future of the students. Education data mining is very usefuldisciplines, because the amount of data in education system is increasing day by day .in higher education is relatively new but its importance increases because of increasing database. There are many approach for measuring students' performance .K-means is one of most efficient and used method .With the help of data mining the hidden information in the database is get out which help for improvement of students' performance. Decision tree is also a method used to predict the students' performance.In recent years, the biggest challenges that Educational institutions are facing the more growth of data and to use this data to improve the quality so it can take better decisions.Clustering is one of the basic techniques often used in analyzing data sets. This study makes use of cluster analysis to segment students in to groups according to their characteristics.Unsupervised algorithm like K-means is discussed. Education data mining is used to study the data available in education field to bring the hidden data i.e. important and useful information from it. With the help of these it is easy to improve the result and future of students.

Key words- Clustering Technique, K-means, EDM, Decision tress,and Students data.

I. INTRODUCTION

Clustering methods in machine learning have been applied in many applications such as fraud detection, banking, academic performance and instruction detection. Data mining is data analysis methodology used to identify hidden patterns in a large data set. Higher education is very important for student'slife. Higher education institute are focus on analysis of every objects because of private participation. Machine Learning provides various methods these include classification,association,k-means,decision tree, regression, time series, neural network,etc.

Application of data mining in the educational system is directly help to analysis of participants in the education system. The students also recommend many activities and task.Many factors could act as a barrier to student for maintaining a high percentage that reflects the overall academic performance in college. These factors could be targeted by the faculty members in developing strategies to improve student learning and academic performance by the way of monitoring and analyzing the progression of their performance. Data mining is also used to show how students use material ofparticular course. In teaching environment trainer are able to obtain feedback on students.

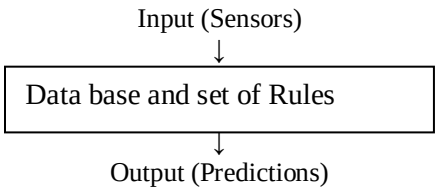


Fig.1: Machine Learning

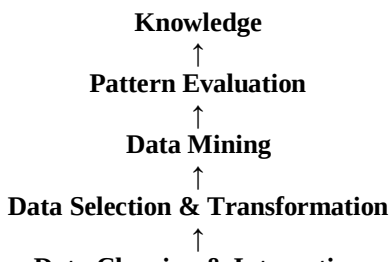


Fig.2: Different stages of data mining process

II. RELATED WORK

FatmaChiheb[1]- Decision tree method is used in these paper. Decision tree is build using the J48 algorithm. Weka toolkit is used and CRISP-DM model is applied. They collect the data about graduates and post-graduates students. It is a case of an Algerian university. The data is taken from, from computer science department. They test decision tree and analyzed the error rates in order to choose the best input and output. Different grades are taken as attribute and student's performance is predicated.

V.Shanmugarajeshwari[2]-They evaluate the students' performance using the classification techniques. The input data is collected from ayya nadir janakiAmmal College,sivakasi, from the computer science department. For feature selection number of methods is discussed. Training data is applied on the data set and the classifier model has

been developed. Decision tree classification was used to predict the students' performance.

M.Durairaj[3]-Educational details and performance is based upon various factors like personal details, social etc. WEKA toolkit is used they collect the data set of college students real time data that describe the relationship between learning behavior of students and their academic performance, the data set contain students detail of different subject marks in semester which is subjected to the data mining process. In these K-means clustering is used and from the total number of 300 student record dataset, they choose 38 students record for our analysis .The confusion matrix is there to shows pass, fail, and absence for the exam. They compare the weighted average for decision tree and naviebayes techniques.

Mr.Shashikantpradipborgavakar[4]-Here the data clustering is used as k-means clustering to evaluate students' performance. Their performance is evaluated on the basic of class test, mid test, and final test. In their model they measured by internal and external assessment, in which they tale class test marks, lab performance, quiz etc. and final grade of students is predicted They generate the graph which shows the percentage of students getting high, medium, low gpa.

EdinOsmanbegovic[5]-In these paper supervised data mining algorithm were applied.Different method of data mining was compared.The data were collected from the survey conducted during the summer semester at the University of Tuzla. Many variable like Gender,GPA,Scholarships,High school,Entrance Exam,Grade,etc. are taken for the performance.Naive Bayesalgorithm, multilayer Perceptron, J48issued. Theresult indicates that the naïve Bayes classifier outperforms in predication decision tree and neural network method. These will help the student for future.

E.venkatasanet.al[6]-In these article the clustering and classification algorithm were compared using matrix laboratory software, for the initial data WEKA software is utilized. Data set of students was picked up from private arts and science colleges from Chennai city. Near about 573 students are there in the database. In the details they take the internal exam and end semester exam details. Algorithm such as J48 were used allows the input attribute to get classification model. Matrix Laboratory is used for measuring the operational of several data mining algorithm. There is a table for error measure.

A.seetharamNagesh[7]-Prediction of students' performance is so important but if it is predicted at early stage it become so useful for the studentsHere they applied k means clustering algorithm for analyzing the students result data and predicting the students' performance. Unsupervised techniques are also called clustering techniques. The k means is partition based clustering algorithm. The distance measure in k means clustering is Euclidean distance. Here the data set used was obtained from the information department of the engineering college. The attribute are aggregate and attendance for experiment. They create the final output after clustering, They shows by red, green, blue to differentiate the poor, average, good students.

Qasem A.Al-Radelideh [8]-The title of the paper is "Mining student data using decision tree".They use data mining process for student performance in university courses to help the higher education management.Many factorsaffect the performance.They use classification technique for building the reliable classification model,the CRISP-DM (cross-industry standard process for data mining) is adopted .These method consist of five steps i.e. collecting the relevant features of the problem, Preparing the data, Building the classification model, Evaluating the model and finally future prediction. The data were collected in table in proper format, the classification model were building using the decision tree method. Many rules were applied. The WEKA toolkit is used Different classification methods were used like ID3,C4.5 and naïve Bayes and accuracy were in the table as result.

MashaelA[9]-These researches has applied decision tree for predicting students final GPA.It used WEKA toolkit .It collect the data from C.s. College at king save university in the year 2012 were collected from the institute.Each student record with different attributes.Student name,student id, final GPA,semester of graduation etc.It is important to improve the final GPA of the student.

Ryan S.J.D.Baker[10]-"The state of educational data mining in 2009:A review and future vision"In these paper author review the trend in 2009 in field of educational data mining. The year 2009 finds research communizing of EDM and these moment in EDM bring unique opportunity.EDM categories in web mining, Statistics and Visualization, Clustering, Relationship mining i.e. Association rule mining and datamining. There are many application of edm.These papers discuss about the EDM.

Pooja M.Dhekankar[11]-"Analysis of student performance by using data mining concept "Data mining technique is used in many area and in the educational field it become so important for future of the students .Students classification is done on the basic of students mark.Association rule,clustering outlier detection,classification is discussed in this paper.

Anjad Abu saa[12]-It applies c4.5, CART, ID5 algorithm for analysis of students' performance. It takes various parameters for the accuracy. Decision tree is build and based on it student performance is predict. Naive Bayes classification is also applied which assumes that all given attribute in a dataset is independent. It create different quantities predictive model by using different data mining tasks which is effective to predict student grades .various decision tree algorithm were implemented . Finally we can say that it help the university as well as students.

YoavBergner[13] et.al-It used collaborative filtering analysis of student data.There is logistic regression as collaborative filtering.There is parameter estimation. There is simulated

skill response.It applied numerical method for analyzing student response matrix with the goal of predicting response; it showed of naturally parameterizes series of models and multidimensional IRT.

III. EXISTING SYSTEM

Decision tree is supervised techniques and there are many methods to build the decision tree and to predict the performance. There are huge amount of data produced in educational system. These can be exploited in order to extract the useful knowledge. In today's system lots of technique is used to predict the students' performance. In the existing system decision tree is build using J48 algorithm. There is a case of Algerian university in which student's performance is predicting using decision tree. Decision tree method is unstable because decision tree give many possible answers. On changing the root node it change the tree and have different prediction. There is the huge amount of data in the educational system in the existing system they predict the performance on the basic of previous semester result. Decision tree is build using the J48 algorithm which is very hard to build because of its splitting. Tree algorithm use many test to determine particular split. But even before that has been determined the algorithm has tried much combination of variables to get the best split. Weka toolkit is used and crisp-dm model is applied.

IV. PROBLEM STATEMENT

There is problem that, there is huge amount of data in the educational system, For predicting the students' performance there should be method which is more efficient and produced useful result. Decision tree is a classification technique which is less efficient as compare to clustering techniques J48 is a decision tree algorithm which is used for predicting student performance but it is less efficient as compare to k-means clustering techniques. Decision trees examine only a single field at a time, leading to rectangular classification boxes. This may not perform well with the actual records in the decision tree.Calculations can get very complex, particularly if many values are not certain and if many outcomes are linked. Decision tree are not stable it means that small change in the data can lead to a large change in the structure of the optimal decision tree.

V. PROPOSED WORK

Prediction of students' performance can be done using Machine Learning algorithm. Clustering is a technology in which there is cluster with group of similar data. K means algorithm is used to predict the performance of students. K means is a unsupervised machine learning algorithm. K means clustering set the partition of n clarifications into k clusters in which each observations belongs to cluster with nearest mean.Cluster is slow with the mean value of the objects in a cluster, which can be viewed as the cluster centroid. The idea is to define K centers and one for each cluster. These middles should be placed I proper way because different location give different result.So better choice is to place them far away from each other.The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is finished and anprimary group age is done. At this point we need to re-calculate k new centroids by watchingfrom the previous step. After we have these k new centroids, a new procedure has to be done amongst the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may announcement that the k centers change their location step by step until no more changes are done or in other words centers do not move any more.

VI. COMPARATIVE STUDY OF EXISTING AND PROPOSED WORK

Decision Tree (J48 Algorithm)-A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node represents a test on an attribute, each branch means the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node.J48 is an extension of ID3. The additional structures of J48 are accounting for missing values, decision trees pruning, constant attribute value ranges, derivation of rules, etc. In the WEKA data mining tool, J48 is an open source Java implementation of the C4.5 algorithm. The WEKA tool provides a number of options associated with tree pruning. In case of potential over fitting pruning can be used as a tool for précising. In other algorithms the classification is performed recursively till every single leaf is pure, that is the classification of the data should be as perfect as possible. This algorithm it produces the rules from which particular identity of that data is generated. The objective is gradually generalization of a decision tree until it gains equilibrium of flexibility and accuracy.

Disadvantages of decision tree algorithm-

- For data including definite variables with different number of stages information gain in decision treeis biased in favor of those attributes with more levels.
- Tree structure prone to sampling – While Decision Trees are mostly robust to outliers, due to their tendency to over fit, they are prone to sampling errors. If sampled preparation data is somewhat different than evaluation or scoring data, then Decision Trees tend not to create great results.
- Tree splitting is locally greedy – At each level, tree looks for binary divided such that impurity of tree isreduced bymaximum amount.
- They are often relatively inexact. Many other predictors perform better with similar data. This can be remedied by changing a single decision tree with random forest of

decision trees, but a random forest is not as easy to interpret as a single decision tree.

K-means Algorithm- clustering is a method of grouping a set of objects of similar type i.e. in such a way that objects in the same group are more similar to each other as compare to those in other groups. It is also known as cluster analysis. It is not specific algorithm, but the general task to be solved. It can be achieved by many algorithms that differ in their understanding of what makes a cluster and how to efficiently find them. The main advantage of clustering techniques is that it is adaptable to changes which are less in classification techniques and it also helps single out useful features for different groups.Clustering method or cluster analysis is mainly used in applications such as prediction, market research, pattern recognition, data analysis, and image processing.

Advantages of K-means algorithm-

- It is easy to implement.
- When there is large number of variables, K-Means may be computationally faster than other clustering techniques.
- K-Means may produce higher clusters.
- An instance can change cluster (move to another cluster) when the centroids are recomputed.

VII. IMPLEMENTATION DETAILS

K means algorithm is used to predict the performance of students.k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. **Algorithm is as follow-**

Step 1-Accept the number of cluster to group data and the dataset to cluster as input values.

Step 2-Inilize the first K cluster (Choose random K element)

Step 3-Calculate the arithmetic mean of each cluster formed in the dataset.

Step 4-K mean assign each record in the dataset to only one of the initial cluster (the nearest cluster using a distance measure).

Step 5-K means re-assigns each record in the dataset to the most similar cluster and recalculate the mean of the entire cluster in the dataset.

VIII. RESULT ANALYSIS

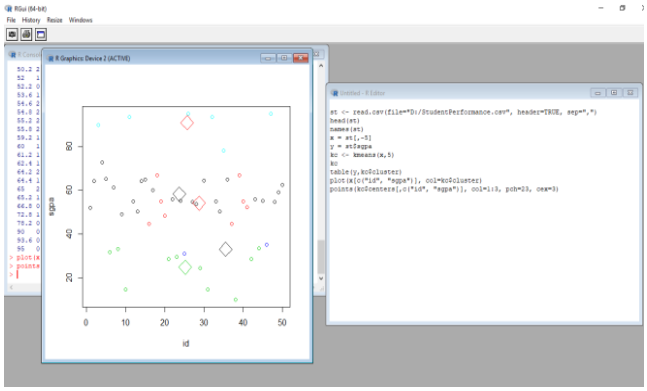
K-Means algorithm is used to predict the students' performance. It is stable and efficient as compare to decision tree.In the dataset we take the attribute-

Student_id,-Unique id correspond to every student.

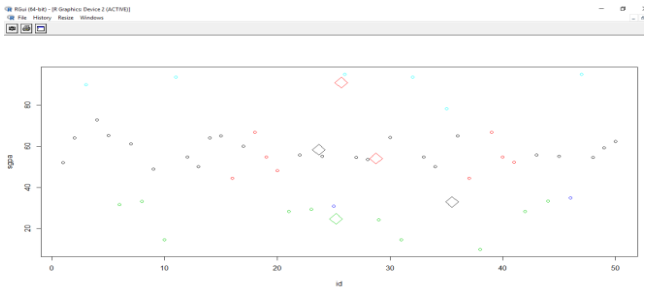
Semester (sem1-sem2)-Semester id correspond to semester i.e.(sem 1 or sem 2).

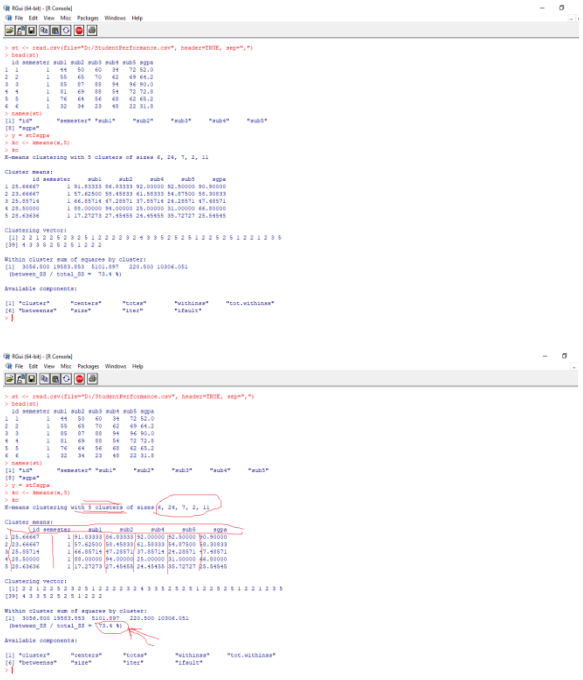
Subject-marks (sub1-sub5)-Each subject marks correspond to every student in both the semester.

Sem Result (Sgpa)–The percentage of those students in that particular semester.

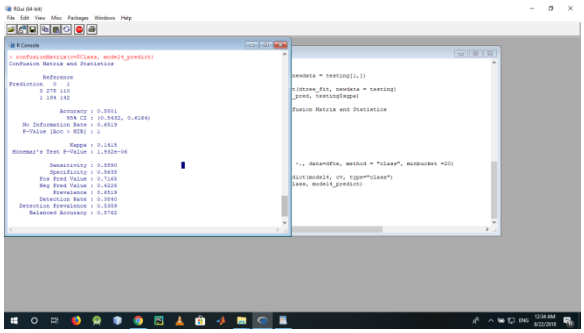


When k=5 and the graph between id and sgpa





We also run the decision tree with same dataset for a better compression.



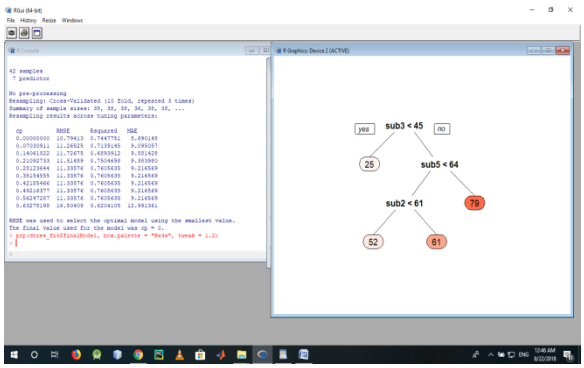


Table: Compression Study Existing Work Versus Proposed Work

Parameters	Existing Work	Proposed Work
	Decision Trees	K-Mean
Average(Sgpa)	Above 58%	Above 71 %
Source Code Execution		10.32 Seconds

IX. CONCLUSION

Machine learning is very emerging technology that every placed it used. Now days in bank, labs, telecom, industrial each and every place machine learning is used. Data mining is part of it which helps in prediction, future prediction is very important in many place which help so much. Many algorithm is build and more and more research is going on every technology used the concept of it. We survey many papers for prediction of students' performance.Decision tree method is used in many place but on comparing to clustering techniques i.e. k means it is less efficient, K means is more efficient and stable.Students'performance is so important for their future it not only help student but also help teachers institute parents. Many big institutes used the concept of AI for prediction.

X. REFERENCES

[1]. Fatmachiheb,FatimaBoumahdi- Predicting students' performance using Decision trees: Case of an Algerian University.2017 International conference on Mathematics and information technology, Adrar, Algeria –Dec 4-5,2017.

[2]. V.Shanmugarajeshwari,Analysis of students performaqnce evaluation using classification techniques,978-1-4673-8437-7 IEEE.

[3]. M.Durairaj-Educational data mining for prediction of students' performance using clustering algorithm, M.durairaj et.al (IJCSIT) International journal of computer science and information technologies vol.5(4),2014.

[4]. Mr.ShashikantpradipBorgavakar-Evaluating students' performance using K means clustering, International journal of engineering Research and technology(IJERT) vol.6 issue 05 May 2017.

[5]. EdinOsmanbegovic, MirzaSuljic-Data mining approach for predicting student performance Economic Review – Journal of Economics and Business, Vol. X, Issue 1, May 2012.

[6]. E.Venkatesan et.al-Prediction of students' academic performance using classification and clustering algorithm,International journal of pure and applied mathematics volume 116 no. 16 2017.

[7]. A.SeetharamNagesh –Application of clustering algorithm for analysis of student academic performance,International journal of computer sciences and engineering volume-6,issue-1.

[8]. Qasem A. Al-Radaideh, Emad Al-Shawakfa - Mining Student Data Using Decision Trees, Research Gate Article 2006.

[9]. Mashael A. Al-Barrak and Muna Al-Razgan "Predicating students final GPA using decision tree: A case study "International Journal of Information and Education Technology, Vol. 6, No. 7, July 2016.

[10]. Ryan S.J.D. Baker KalinaYacef "The state of educational data mining in 2009 A review and future vision" Journal of Educational Data Mining, Article 1, Vol 1, No 1, Fall 2009.

[11]. Miss .Pooja M .Dhekankar Dinesh S. Datar "Analysis of Student Performance by using Data Mining Concept" International Journal on Recent and Innovation Trends in Computing and Communication Volume: 3 Issue: 5 2942 – 2944 IJRITCC | May 2015.

[12]. Amjad Abu Saa , "Educational Data Mining & Students' Performance Prediction" , (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 5, 2016.

[13]. Yoav Bergner , Stefan Droschlery , GerdKortemeyerz , SaifRayyan, Daniel Seatonand David E. Pritchard "Model Based Collaborative Filtering Analysis of Student Response Data: Machine Learning Item Response Theory" Proceedings of the 5th International Conference on Educational Data Mining.