# Analysis of Students' Performance using Modified K- Mean Algorithm

[1]Mr. A. T. Sonawane, [2]Prof. Pritesh Jain, [3]Mr. Upendra singh

[1]PG Student, [2]Assistant Professor, PCST College Indore, MP, India.

[1]ajay.sonawane10121993@gmail.com, [2]Pritesh.jain@gmail.com, [3]Upendrasingh49@gmail.com

**Abstract—. Machine Learning is a branch of artificial intelligence that provides computers with the ability to learn without being explicitly programmed. Machine learning is a field that is used in every system. Machine learning is used in educational system, pattern recognition, Games, Industries, Social media services, online customer support, Product Recommendation Etc. In education system its importance becomes more because of the future of the students. Educational data mining is very useful disciplines, 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 approaches for measuring students' performance. With the help of data mining the hidden information in the database is get out which help for improvement of students' performance. In recent years, the biggest challenges that Educational institutions are facing the explosive growth of informative data and to use this data to improve the value of administrative decisions. Clustering is one of the basic techniques often used in analysing data sets. Many clustering techniques are there but modified K-means is one of most efficient and used method. Classification techniques are also there and most popular is decision tree. Decision tree is also a method used for analysis of the students' performance but compare to modified K-means, it is less stable. Unsupervised algorithm is discussed. These make use of cluster analysis to segment students in to groups according to their characteristics. Elbow method is there to determine the cluster size; it will help in optimal solution. 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. It is not only useful for students but also for teacher and institute to improve their result.**

**Keywords:-Machine Learning, Clustering Technique, Modified K-means, EDM, Decision tress, and Students data.**

## I. INTRODUCTION

Machine Learning is the part of artificial intelligence (AI) that provide the system to ability to learn automatically, it also provide facility to improve from experience without being explicitly programmed. It can be defined as a branch of artificial intelligence that provides computer with the ability to learn without having all the information with respect to a domain in the program itself. It has the study of Pattern Recognition. Machine Learning is related to Computational statistics, which also emphasis on prediction using the computer. Data mining is term in the machine learning, in which it focuses on the data analysis.

The feature of machine learning is significant because as the models are exposed to new data, they are able to independently adapt. They learn from earlier computations to produce consistent, repeatable decisions and result. Machine learning is field of AI in which there is lots of data and it fit the data into modules so that can be utilized by people. Machine learning and traditional computational method both are different and also machine learning is field of computer science. In traditional computing, algorithm is sets of programmed instruction used by computers to calculate. Machine learning algorithm permit computer to train on data input and use statistical analysis for the output fall within specific range.

The main focus of any higher institution is to improve decision making at the administrative level and to impart education. Analysis of students' success in high institutions is one of the bases for improving the quality of education. Student's performance is an important and integral part in higher institutions. This is because the quality of education in universities is based on its excellent record of academic achievements. Predicting students' performance has become an important task due to the huge amount of data in educational databases. So educational data mining is important and useful now days.

Machine learning is a field that can learn from past experience to improve the future performance. Here the learning means improvement of the algorithm and then use these algorithm in the future. The systems are there and it

 DOI : 10.35291/2454-9150.2019.0254 

is cannot be consider intelligent if it does not have ability to learn. So learning ability is most important feature for the intelligent system [1].

The self-driving Google car, fraud detection, online recommendation system, friend suggestions on social network, Netflix showcasing the movies are the example of applied machine learning. Facial recognition technology is so used method now days which allows user to tag and share photos in social medial platform. OCR (Optical Character Recognition) technology converts images of text into movable type.

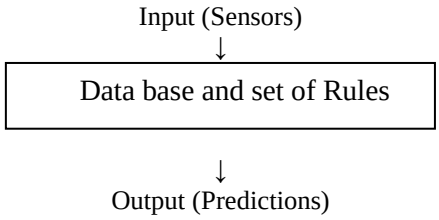


**Figure 1.1 Machine Learning**

## II. LITERATURE WORK

**M.Durairaj [10]-**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 behaviour 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 navie bayes techniques.

**Mr.Shashikant pradip borgavakar [11]-**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.

**Edin Osmanbegovic [12]-**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 Bayes algorithm, multilayer Perceptron, J48 issued. The result indicates that the naïve Bayes classifier outperforms in predication decision tree and neural network method. These will help the student for future.

**E.venkatasan et.al [13]-**In this 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.seetharam Nagesh [14]-**Prediction of students' performance is so important but if it is predicted at early stage it become so useful for the students Here they applied k means clustering algorithm for analysing the students result data and predicting the students' performance. Unsupervised techniques are also called clustering techniques. The k means is partition based clustering algorithm. Euclidean distance is the distance which is measure in k means clustering algorithm. 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 [15]-**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 factors affect 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.

## III. METHODOLOGY

A. Development of k-mean clustering algorithm Given a dataset of n data points x1, x2, …, an such that each data point is in Rd , the problem of finding the minimum variance clustering of the dataset into k clusters is that of finding k points $\{m_j\}$ (j=1 , 2..... K) in $R^d$ such that

$$\frac{1}{N}\sum_{i=1}^{n}[min_j d^2(x_i, m_j)] \quad (1)$$

is minimized, where $d(x_i, m_j)$ denotes the Euclidean distance between $x_i$ and $m_j$ The points $\{m_j\}$ (j=1 , 2..... K) Are known as cluster cancroids'. The problem in Eq.(1) is to find k cluster cancroids', such that the average squared Euclidean distance (mean squared error, MSE) between a data point and its nearest cluster cancroids is minimized.

Those k-means calculation gives a simple strategy should execute estimated answer for eq. (1). The purposes behind

 DOI : 10.35291/2454-9150.2019.0254 

the Notoriety from claiming k-means are simplicity Also Straightforwardness for implementation, scalability, pace from claiming merging What's more versatility to meager information.

Those k-means algorithm camwood be considered perfect Likewise An gradient plummet procedure, which starts during beginning bunch cancroids', and iteratively updates these cancroids' should diminishing the destination capacity Previously, eq. (1). The k-means dependably meet with An neighborhood base. Those specific nearby least found relies on the beginning group cancroids'. The issue for finding the worldwide least will be NP-complete. Those k-means calculation updates group cancroids' till neighborhood least is found. Fig. 1 indicates the summed up pseudo codes of k-means algorithm; and customary k-means calculation is introduced clinched alongside fig. 2 separately.

Preceding the k-means calculation converges, separation What's more cancroids calculations need aid carried out same time loops need aid executed An amount about times, say l, the place the certain basic l may be known as the amount from claiming k-means iterations. The exact esteem for l differs relying upon the starting beginning bunch cancroids' much on the same dataset. With the goal those computational period intricacy of the calculation is O(nkl), the place n will be those aggregate number about Questions in the dataset, k will be those required amount of groups we recognized and l will be the amount from claiming iterations, $k \leq n$, $l \leq n$ [6].

Step 1: Accept the number of clusters to group data into and the dataset to cluster as input values
Step 2: Initialize the first K clusters
- Take first k instances or
- Take Random sampling of k elements
Step 3: Calculate the arithmetic means of each cluster formed in the dataset.
Step 4: K-means assigns each record in the dataset to only one of the initial clusters
- Each record is assigned to the nearest cluster using a measure of distance (e.g Euclidean distance).
Step 5: K-means re-assigns each record in the dataset to the most similar cluster and re-calculates the arithmetic mean of all
the clusters in the dataset.

**Figure 3.1: Generalized Pseudo Code of Traditional K-Means**

```
1 MSE = large number;
2 Select initial cluster cancroids' {mj}j
K = 1;
3 Do
4 Old MSE = MSE;
5 MSE1 = 0;
6 For j = 1 to k
7 mj = 0; NJ = 0;
8 end for
9 For i = 1 to n
10 For j = 1 to k
11 Compute squared Euclidean
Distance d2 (xi, mj);
12 end for
13 Find the closest cancroids mj to xi;
14 mj = mj + xi; nj = nj+1;
15 MSE1=MSE1+ d2(xi, mj);
16 end for
17 For j = 1 to k
18 NJ = max (NJ, 1); mj = mj/nj;
19 end for
20 MSE=MSE1;
while (MSE<OldMSE)
```

**Figure 3.2: Traditional K-Means Algorithm**

 DOI : 10.35291/2454-9150.2019.0254 

## IV. SIMULATION AND RESULT

We have done our work by using the modified K-means clustering algorithm. Modified K-Means algorithm is used for analysis of students' performance. It is stable and efficient as compare to decision tree. Elbow method is used for the cluster size. 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 the respective semester.
**Sem Result (Sgpa) –**The percentage of those students in that particular semester.
**Gender** – The gender of the students.
**HSC 10th** – The percentage of the 10th class of each student.
**HSS 12th** – The percentage of the 12th class of each student.
**Stay Location** – The location of the student where they are staying i.e. hostel, PG, room.
**Section Id** – The section Id is the section of the student i.e. A or B.
**Student Absence Days** – The number of day's student was absent.
**Raised Hands** – The number of times student raised the hand for any problem.
**Discussion** – The number of times student do the discussion.

The attribute taken in the data set have unique id for every student and there is a marks and result with correspond to each and every students. We have taken 50 unique ids for implementation and there is data of two semesters. Some parameter are useful and some are not like result is important.

### 4.1 Graph between Id and sgpa

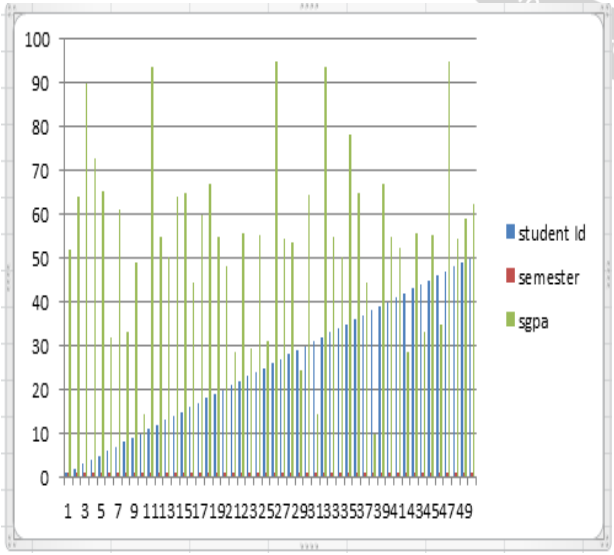


**Figure 4.1 Bar Graph between Id, Semester, Sgpa**

The figure 4.1 shows the bar graph between id, sem, sgpa. These shows that for different it different sgpa are there in semester 1, so semester is same and green line shows the sgpa foe every student. According to these it can be found out the number of students in between the different sgpa.

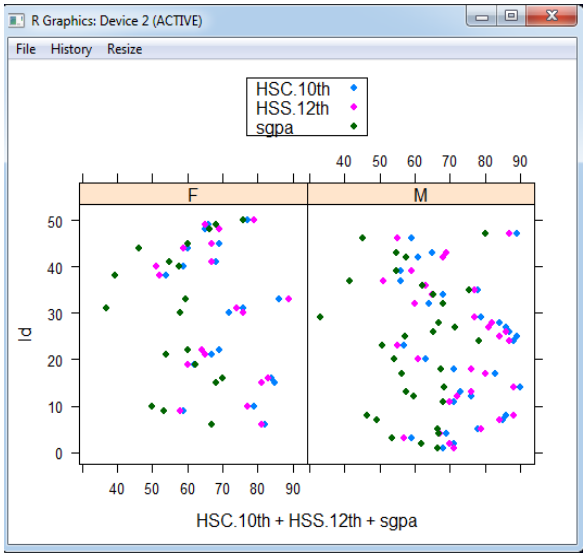


**Figure 5.4 XY Conditioning Plot**

**Table: 4.1 Compression Study Existing Work versus Proposed Work**

| Parameters | Existing Work | Proposed Work |
|---|---|---|
| | Decision Trees | Modified K-Mean |
| Accuracy | Above 58% | Above 71 % |
| Source Code Execution | | 10.32 Seconds |

## V. CONCLUSION

**5.1 Conclusion :** Analysis of student's performance is done with the help of modified k means algorithm. Machine learning is very emerging technology that every placed it used. Now days in bank, labs, telecom, industrial, and in every place machine learning is used. Data mining is part of it which helps in prediction, future prediction is very important in much system, and in the education system it becomes more important. Many algorithm is build and more and more research is going on every technology used the concept of it. Modified K-means algorithm is there in which elbow pint is used. Here in the modified k means the cluster are made of similar type which helps in more accuracy.
Finally, the result indicates that modified K-means algorithm is provide better result than other algorithm and it has high accuracy and stable.

**5.2 Future Scopes :** In this work modified K-means has been implemented and analysis of the students' performance. This work can be extended by taking more parameter and can be comparing with different algorithm it can also be compare with unsupervised algorithm. So in future, different techniques can be used for prediction of

 DOI : 10.35291/2454-9150.2019.0254 

student's performance. There can be more parameter for the data set.

## REFERENCES


[1] K. Govindasamy, T.Velmurugan, "Analysis of student academic performance using clustering techniques" International Journal of Pure and Applied Mathematics Volume 119 No. 15 2018.

[2] Shaymaa, E. Sorour, Tsunenori Mine, Kazumasa Goda, Sachio Hirokawa, "Efficiency of LSA and K-means in Predicting Students' Academic Performance Based on Their Comments Data" CSEDU, International Conference on Computer Supported Education, 2014.

[3] Neetesh Kumar, Deo Prakash Vidyarthi , "A Green SLA Constrained Scheduling Algorithm for Parallel/Scientific Applications in Heterogeneous Cluster Systems" , Sustainable Computing: Informatics and Systems , Elsevier , 2019, pp. 1-18.

[4] Rashmi Chaudhry, Shashikala Tapaswi, Neetesh Kumar , "FZ enabled Multi-objective PSO for multicasting in IoT based Wireless Sensor Networks" , Information Sciences , Elsevier , Volume 498 , 2019 , Pp. 1-20.

[5] Neetesh Kumar, Deo Prakash Vidyarthi , "A GA based energy aware scheduler for DVFS enabled multicore systems" , Computing , Springer Vienna , Volume 99 , Issue 10 , 2017, pp. 955-977

[6] Neetesh Kumar, Deo Prakash Vidyarthi , "A novel hybrid PSO–GA meta-heuristic for scheduling of DAG with communication on multiprocessor systems" , Engineering with Computers , Springer London , Volume 32 , Issue 1 , 2016 , pp. 35-47.

[7] Rashmi Chaudhry, Shashikala Tapaswi, Neetesh Kumar , "Forwarding zone enabled PSO routing with network lifetime maximization in MANET" , Applied Intelligence , Volume 48 , Issue 9 , Springer US , 2018 , pp. 3053-3080

[8] Dinesh Bhuriya ,Girish Kaushal ,Ashish Sharma, Upendra Singh ," Stock Market Predication Using A Linear Regression ", Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of IEEE , 20-22 April 2017 ,pp. 1-4.

[9] Ashish Sharma ,Dinesh Bhuriya ,Upendra Singh, "Survey Of Stock Market Prediction Using Machine Learning Approach" , Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of IEEE , 20-22 April 2017 ,pp.1-5.

[10] M. Durairaj,C. Vijitha, "(IJCSIT) International Journal of Computer Science and Information Technologies" , Vol. 5 (4) , 5987-5991, 2014.

[11] Mr. Shashikant Pradip Borgavakar, Mr. Amit Shrivastava, "Evaluating student's performance using k-means clustering" International Journal of Engineering Research & Technology (IJERT), Vol. 6 Issue 05, May – 2017.

[12] Edin Osmanbegovic, Mirza Suljic, "Data mining approacg for predicting student performance" Economic Review – Journal of Economics and Business, Vol. X, Issue 1, May 2012.

[13] E.Venkatesan, S.Selvaragini, "Prediction of students' academic performance using classification and clustering algorithms" International Journal of Pure and Applied Mathematics, Volume 116 No. 16 , 327-333, 2017.

[14] A Seetharam Nagesh, Ch V S Satyamurty, "Application of clustering algorithm for analysis of student academic performance" IJCSE International Journal of Computer Sciences and Engineering, Volume-6, Issue-1, 31/Jan/2018.

[15] Qasem A. Al-Radaideh, Emad Al-Shawakfa, Mustafa I. Al-Najjar, "Mining students data using decision trees", The 2006 International Arab Conference on Information Technology (ACIT'2006), Jordan, Nov. 2006.

[16] Vineeta Prakaulya ,Roopesh Sharma ,Upendra Singh, "Railway Passenger Forecasting Using Time Series Decomposition Model ", Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of IEEE , 20-22 April 2017 ,pp.1-5.

[17] Sonal Sable ,Ankita Porwal ,Upendra Singh , "Stock Price Prediction Using Genetic Algorithms And Evolution Strategies ", Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of IEEE , 20-22 April 2017 ,pp.1-5.

[18] Rohit Verma ,Pkumar Choure ,Upendra Singh , "Neural Networks Through Stock Market Data Prediction" , Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of IEEE , 20-22 April 2017 ,pp.1-6..

[19] T.Velmurugan, Dr.T.Santhanam, "Performance analysis of k-means and K-medoids clustering algorithm for a randomly generated data set" International Conference on Systemics, Cybernetics and Informatics, volume 1, January 2018.

[20] Pooja Kewat , Roopesh Sharma , Upendra Singh , Ravikant Itare, "Support Vector Machines Through Financial Time Series Forecasting ", Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of IEEE , 20-22 April 2017 ,pp. 1-7.

[21] Yashika Mathur ,Pritesh Jain ,Upendra Singh, "Foremost Section Study And Kernel Support Vector Machine Through Brain Images Classifier ", Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of IEEE , 20-22 April 2017 ,pp.1-4.

[22] Pooja Kewat , Roopesh Sharma , Upendra Singh , Ravikant Itare, "Support Vector Machines Through Financial Time Series Forecasting ", Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of IEEE , 20-22 April 2017 ,pp. 1-7.

[23] Yashika Mathur ,Pritesh Jain ,Upendra Singh, "Foremost Section Study And Kernel Support Vector Machine Through Brain Images Classifier ", Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of IEEE , 20-22 April 2017 ,pp.1-4.

[24] Vineeta Prakaulya ,Roopesh Sharma ,Upendra Singh, "Railway Passenger Forecasting Using Time Series Decomposition Model ", Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of IEEE , 20-22 April 2017 ,pp.1-5.

[25] Sonal Sable ,Ankita Porwal ,Upendra Singh , "Stock Price Prediction Using Genetic Algorithms And Evolution Strategies ", Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of IEEE , 20-22 April 2017 ,pp.1-5.

[26] Rohit Verma ,Pkumar Choure ,Upendra Singh , "Neural Networks Through Stock Market Data Prediction" , Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of IEEE , 20-22 April 2017 ,pp.1-6.

[27] Dinesh Bhuriya ,Girish Kaushal ,Ashish Sharma, Upendra Singh ," Stock Market Predication Using A Linear Regression ", Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of IEEE , 20-22 April 2017 ,pp. 1-4.

[28] Ashish Sharma ,Dinesh Bhuriya ,Upendra Singh, "Survey Of Stock Market Prediction Using Machine Learning Approach" , Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of IEEE , 20-22 April 2017 ,pp.1-5.


 DOI : 10.35291/2454-9150.2019.0254 