

Personalized Analytics for Predicting Student Performance

Abhishek Ranjan¹, Dr. Pankaj Richhariya²

¹M.Tech Student, Department of Computer Engineering, Bhopal Institute of Technology & Science, Bhopal, India

²Professor & Guide, Department of Computer Engineering, Bhopal Institute of Technology & Science, Bhopal, India

Abstract-- Machine learning is being employed in the educational system, pattern recognition, games, industries, social media services, online customer care, and product recommendations, among other things. Due to the kids' future, the relevance of the educational system is increasing. Since every student today seeks higher education, there is a vast amount of data in higher education, which increases the need for M.L. procedures in the educational system. There are numerous tools available for analyzing student performance. Data mining is used to uncover hidden information, which will aid in the evaluation of student data. There is a tone of data in the field of education, and all of it is beneficial to both teachers and students. The use of M.L. technology in the educational setting is becoming increasingly important as the institute expands. One of the fundamental methods frequently employed in data analysis is clustering. There are other clustering methods, but modified K-means is one of the most popular and effective ones. There are various classification techniques, with decision trees being the most common. Although less stable than modified K-means, decision trees are frequently used to analyse student performance. The discussion of unsupervised algorithms. To categories students into groups based on their traits, these employ cluster analysis. The elbow approach is available to calculate the cluster size; it will aid in finding the best answer. Within the total of the squares, there is an elbow method that looks over the arm and includes the elbow point. The M.L. approach makes it simple to enhance kids' performance and future. Students can improve their results, but teachers and institutions can as well.

Record Terms – Prediction using SVM, Machine Learning.

I. INTRODUCTION

Artificial intelligence (AI machine)'s learning branch enables systems to learn on their own. They can learn through an automatic mechanism. Additionally, it is possible to use experience to enhance the system. For better results, machine learning identifies patterns in the data. The science of pattern recognition is so present. Computational statistics, which emphasises computer-based prediction, is connected to machine learning. Data analysis is the primary focus of the machine learning concept known as "data mining."

Machine learning is important because it allows models to adapt to new data. In order to create reliable, repeatable decisions and results, they learn from past computations.

AI's machine learning sector uses a large amount of data and organises it into modules that people can use. Machine learning is a branch of computer science, and it differs from standard computational methods in both ways. Traditional computing uses a set of programmes that are used to perform calculations. Machine learning uses numerous analyses, including statistical studies, to produce results from data input.

Any higher organization's primary focus is on recovering result generation at the administrative level. One of the pillars for raising educational standards is the analysis of students' performance in prestigious institutions. Performance by students is an important and crucial component in higher education institutions. This is so because universities' remarkable record of academic accomplishments determines the quality of their knowledge. The educational system generates a lot of data, all of which can be used for research. As a result, it is now even more crucial to analyse the data. Therefore, educational data mining is significant and helpful today.

Machine learning leverages historical data to enhance performance in the future. Here, learning refers to the improvement of the algorithm and subsequent application of the improved algorithm. There are procedures in place, and an object cannot be deemed intelligent if it lacks the capacity to learn. Therefore, the ability to learn is the most crucial component of an intelligent system [1].

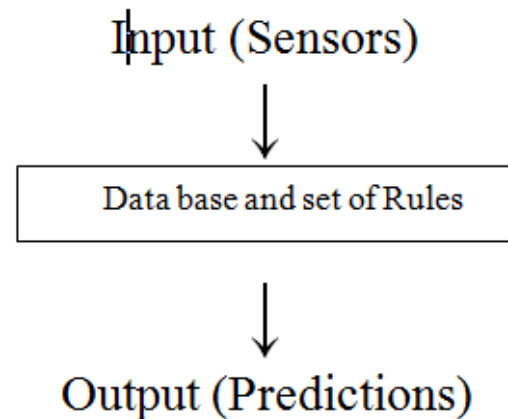


Figure 1.1 Machine Learning



International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 11, Issue 10, October 2022)

There are numerous uses for machine learning, such as fraud detection and product recommendations. This is a very significant application that is utilised by many e-commerce businesses. For instance, if we purchase a phone, we may be encouraged to purchase the phone's case. In social networks, the idea of machine learning is used to provide friend suggestions.

Automatic Because there is so much information in educational databases, predicting student success is a significant task. Educational data mining is being used to tackle this task (EDM). EDM creates techniques for locating data produced from educational settings. Understanding students and their learning environment is accomplished using these techniques. In order to make the appropriate arrangements, educational institutions frequently wonder how many students will pass or fail. It has been noted in earlier studies that many researchers focus on choosing the best algorithm for only classification and neglect to find solutions to issues that arise during the data mining phases, such as data high dimensionality, class imbalance, and classification mistake, among others. These kinds of issues made the model less accurate. Many well-known classification methods are used in this field, however the model this research proposed for predicting student success is based on supervised learning decision trees. Additionally, an ensemble method is used to enhance the classifier's performance. The use of ensemble methods is intended to address classification and prediction issues. This study demonstrates the value of preparing data and fine-tuning algorithms to address problems with data quality. The experimental data set used in this study comes from the UCI Machine Learning Repository and is local to Portugal's Alentejo region. In this study, three supervised learning algorithms—J48, NNge, and MLP—are used experimentally. According to the findings, J48 outperformed all other models with a 95.78% accuracy rate.

II. MOTIVATION

To make prediction of student possibilities to be get selected in company or need of classes.

- ❖ Students can easily get idea of their future possibilities.

- ❖ To make students aware of their future.
- ❖ Enhancement in the completion of work within the constraints of time.

III. PROBLEM DEFINATION

There is often a great need to be able to predict future students' behavior in order to improve curriculum design and plan interventions for academic support and guidance on the curriculum offered to the students. This is where Data Mining (DM) comes into play. DM techniques analyze datasets and extract information to transform it into understandable structures for later use. Machine Learning (ML), Collaborative Filtering (CF), Recommender Systems (RS) and Artificial Neural Networks (ANN) are the main computational techniques that process this information to predict students' performance, their grades or the risk of dropping out of school. Nowadays, there is a considerable amount of research and studies that follow along the lines of predicting students' behavior, among other related topics of interest in the educational area. Indeed, many articles have been published in journals and presented in conferences on this topic. Therefore, the main goal of this study is to present an in depth overview of the different techniques and algorithms proposed that have been applied to this subject

IV. PROJECT SCOPE

- To implement real time system for student performance.
- To perform various operation on student record to check student performance.
- To get prediction of student future possibilities.
- To have the different results in short time

V. USER CLASSES & CHARACTERSTICS

1. Registration: In Registration First, student have to register yourself in portal.
2. Upload Marks: In second phase student should upload their marks as per the academics.
3. Prediction: After uploading marks and details, students will get their prediction details about their career.



Fig 1: Use Case Diagram

VI. SYSTEM ARCHITECTURE

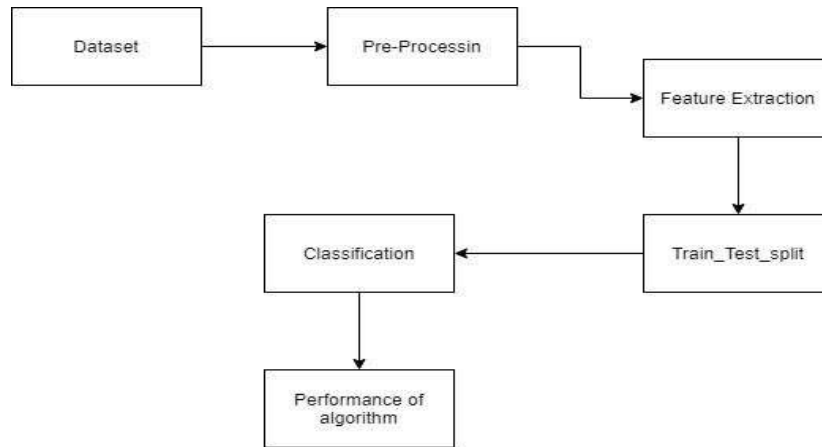


Fig 2: System Architecture

Above diagram shows the Abstract view of System. System has Three Actors

VII. ADVANTAGES

- Student can get the guidance through which he will get idea about in which field he has scope by analyzing his interests and academic performance.
- Student Performance prediction is very important to understand the student progress rate.
- Accessibility from any edge of the world just by having this system. As in this pandemic situation it is useful as no physical analysis will be done by teacher.
- Useful for teacher as she can save time that will be needed to analyze each and every student.



VIII. LIMITATION

- In this we can say, physical analysis will be better than digital analysis.
- It only predicts student on basis of academic performance.

IX. APPLICATION

- Student Performance Prediction can be used in multiple ways by student as well as the teacher.
- From this student can get a suggestion for his future activities. For example, if a Engineering student is using this he will get the suggestions of companies according to his performance and interests.
- Can be used by teacher if she has a huge number of students which may lead to save time.

Implementation

In this chapter we discuss the implementation using the modified K means algorithm on R language and detail about R.

Implementation Detail

Modified K-Means algorithm is used for analysis of student's performance. In these the partition of n object is took place into k cluster. The data is place by look over the nearest mean. It means that data of same type are put in a group and data of other type are there in other group. Analysis is based on the parameter which has their same type of data.

Clustering analysis is broadly used in many applications such as market research, pattern recognition, etc. Modified K means is done by using the elbow method for the cluster size. Using the random number of cluster size may not give optimal solution so by using the elbow method getting number of cluster will give the optimal solution. The idea behind the elbow method is to calculate the sum of square. There is a line chart like plot of the sum of squared values for the range of the value of the K (K may have any number of value). This line chart is look like an arm, and then the valve of the elbow of the arm is the suitable value of the K. The main thing is that choose the small value of K that have a low sum of squared value. It is easy to implement, and give the best suitable result. There are many languages which are widely used for analytic purpose like python, R, SAS etc. Here in the R language to implement the modified K-means algorithm, many packages has to install, then the library of that package which will help to plot the elbow method and also to plot the various graph between parameters. Many parameters are there which used in the analysis of the students' performance, But in the dataset many parameter are there which does not affect the result. Here parameter like result is so important for the analysis, but parameter like gender is not useful for the result but it is also important for the analysis purpose in the way to differentiate the gender and give the actual figures. R studio has lots of option, in which there is an option of help from which anyone can take detail of any library. So there is a lot of option in R studio, and elbow method is useful for taking the cluster size.

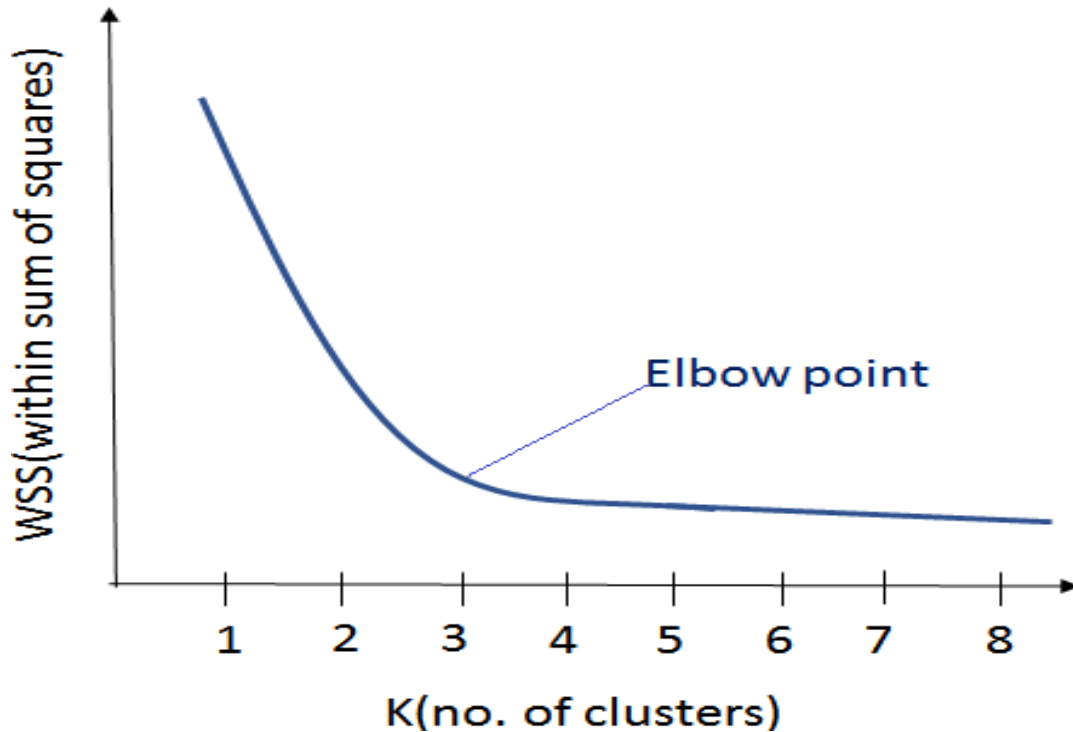


Figure 4.1 Elbow Method

The R setup

R is a programming language that is mainly used in statistical computing. Mostly there is graphical representation in R language. R is available freely. It has an analysis environment. It is available under the general public license and it can be run on various operating systems like windows, Linux, Mac. R is software which also allows the procedures which are in other language like c, c++, python etc. R software is so easy to use and it is simple and it has all programming concept like loops, function etc. In R it has a data handling facility and it also have storage facility. In R there is a function like matrix, vector. It has a large collection of tools for data analysis. The name r was there because it was developed by Ross and Robert and it was managed by core development team. So R is world most widely used statistics language.

Platform required running R studio

- Unix and Unix like systems
- Linux
- Windows XP/7/8
- R studio server.

R Studio

R studio is available freely and it is open source integrated development environment. It is for R programming and for statically computing. There is also a facility of graphical representation in the studio. There are two editions of R studio i.e.

R studio Desktop: Here it is looking like regular desktop application and the program is running locally as regular application. The desktop version is available for windows, Linux, mac operating system.

R studio server: There is a web browser and R studio is accessing through it.

R studio has its importance for graphical use interface and for it, they use the qt framework and it is written in C++ programming language and also in Java language. Its interface is so well organized that user able to view graph, tables of the data, R code, and the output at the same time. It has a lot of feature that allow user to import various files like csv, excel etc. As in below figure it is clear that R studio has four quadrants and all has specific feature. In the first quadrant the script is there, the second quadrant is console.



International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 11, Issue 10, October 2022)

All the running condition generates the result in the console. The third quadrant is environment, in which all the variable are there, they show the environment of different variables.

The last is fourth quadrant in which graph is plotted; more option is there in that like package, files, viewer etc. The entire quadrant has its feature and importance. All the quadrant can be resize as per their needs.

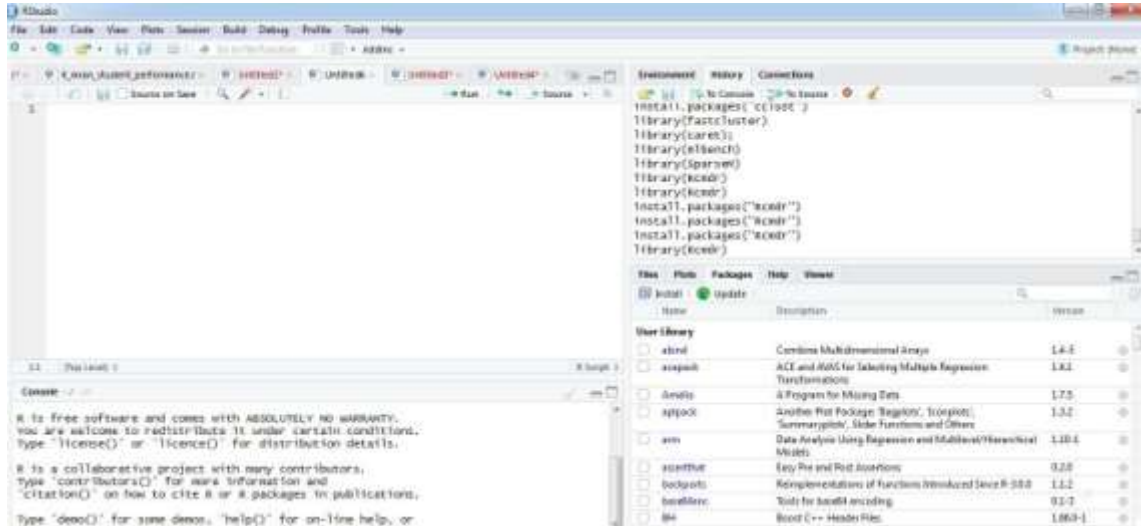


Figure 4.2 R studio

R Packages

There are lots of packages in R and there is an R function and R packages have a collection of it. At the time of installation, R install many packages which is useful at a specific time, Packages can be added later when it need, there is a huge collection of packages and it can be added when there is a specific need. These all are stored in a directory called library. The package which is already installed can be loaded by its use. Some package is default and it is loaded automatically when the console is start. The library location of package is: `.libpaths()`

When we execute the `library()` it will give the list of the packages that are installed. We can install new package as well by: `install.packages("Package name")`

There is huge collection of packages in R library some of the most used packages are `caret`, `ggplot2`, `Rcmdr`, `Stats`. There is lots of option of installing the package in the R. some option are directly and some are by using the script. Many packages are depending on some other package and for it all the packages has to install, then it can be used. If we need the detail of any package, the information of the [package can be seen be `help` option, which can give information about that package.

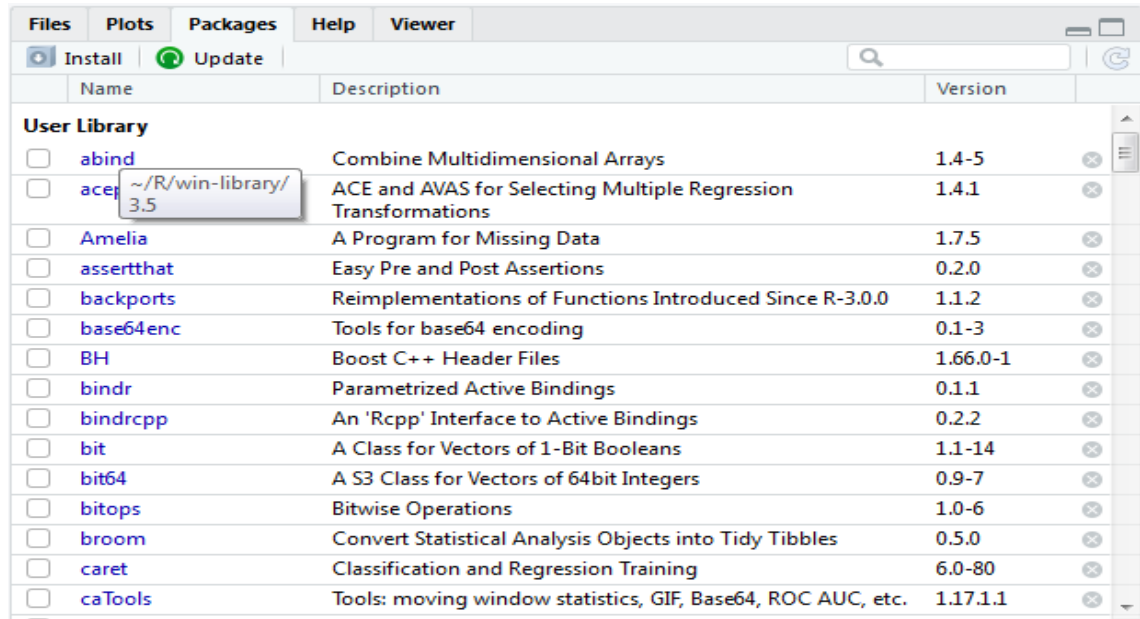


Figure 4.3 Package in R studio

The above figure describe that the in Fourth column in R studio it has a lots of option, there is an option for package, if there needs any help regarding any package it will give, also the use of the package, these feature is so useful because it will help directly when we need any information regarding package.

Dataset of student's

Analysis of students can be done by taking various parameter, but there are some parameter which are so useful and some parameter are there which is not so important or can say which do not affect the result for example here Id, sgpa are important, it has to use in the

dataset, but like gender is not so important means it will not affect the result. The following is the sample figure of the dataset which is taken. In this dataset unique id is there for every student, Here number of id is taken are 50 and has there result like HSC 10th, 12th, sgpa etc. There are some other parameter like raised hands it means the number of times students raised the hands for any problem. In the dataset there are marks of the all five subject and sgpa is there with respect to it. This parameter will help in the analysis of the students' performance. So some parameters are useful and some are not but they are part of the dataset.

id	semester	gender	SectionID	StudentAI	Staylocati	Discusstoi	raisedhan	HSC 10th	HSS 12th	sub1	sub2	sub3	sub4	sub5	sgpa
1	2	M	A	Under-7	Hostel	20	15	68	71	45	45	83	75	85	66.6
2	2	M	A	Under-7	Room	25	20	71	70	64	52	52	85	56	61.8
3	2	M	A	Above-7	Hostel	30	10	59	57	55	53	56	52	51	53.4
4	2	M	A	Above-7	PG	35	30	69	67	65	68	96	53	52	66.8
5	2	M	A	Above-7	PG	50	40	78	79	89	78	57	56	53	66.6
6	2	F	A	Above-7	Room	70	42	82	81	96	45	69	59	65	66.8
7	2	M	A	Above-7	Hostel	17	35	85	84	54	65	54	5	68	49.2
8	2	M	A	Under-7	PG	22	50	86	88	57	32	29	45	69	46.4
9	2	F	A	Under-7	Room	50	12	59	58	96	35	56	15	64	53.2
10	2	F	A	Under-7	Hostel	70	70	79	77	56	62	54	25	52	49.8
11	2	M	A	Under-7	Hostel	80	50	71	70	69	64	56	95	56	68
12	2	M	A	Under-7	Room	12	19	78	72	67	65	58	85	23	59.6
13	2	M	A	Above-7	Hostel	11	5	73	76	69	68	59	53	39	57.6
14	2	M	A	Above-7	PG	19	20	90	88	64	69	52	62	95	68.4
15	2	F	A	Above-7	Room	60	62	85	81	62	68	56	68	86	68
16	2	F	A	Under-7	Hostel	66	30	84	83	59	62	60	94	75	70
17	2	M	A	Above-7	PG	80	36	83	80	55	52	80	50	45	56.4
18	2	M	A	Above-7	PG	90	55	71	76	88	53	75	56	65	67.4
19	2	F	A	Under-7	Room	96	69	62	60	61	84	76	35	56	62.4
20	2	M	A	Under-7	Hostel	99	70	63	61	60	52	71	34	54	54.2
21	2	F	A	Above-7	PG	90	60	67	65	98	51	42	15	64	54
22	2	F	A	Under-7	Hostel	80	10	69	64	87	20	53	85	56	60.2

Figure 4.4 Student dataset

Figure describes that, here it is the dataset in which various parameter is used, in it 50 ids are taken and there is result with respect to it. There is a parameter student absence days which implies the number of day's students are present and absent. Stay location is there which tell that where the student is currently living.

So there are total 50 number of ID and with respect to every there is sgpa and various results.

Operation of Modified K-means Algorithm

K means algorithm run over the R studio by taking above dataset, library is install in the R studio for plotting the graph.

Import of Data

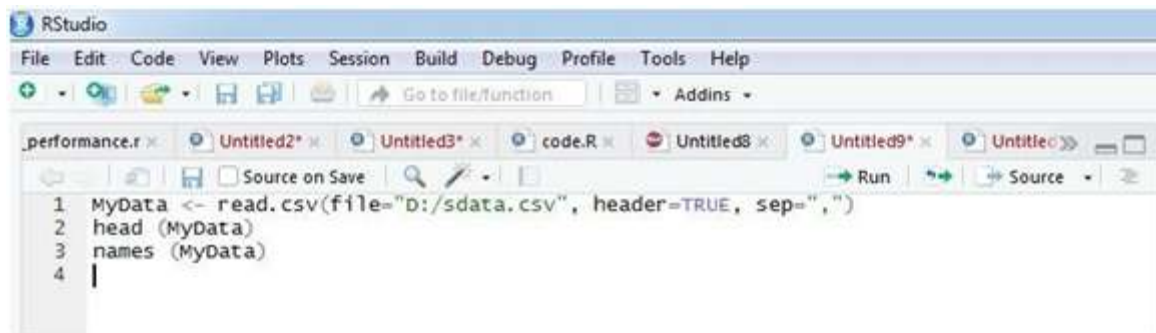
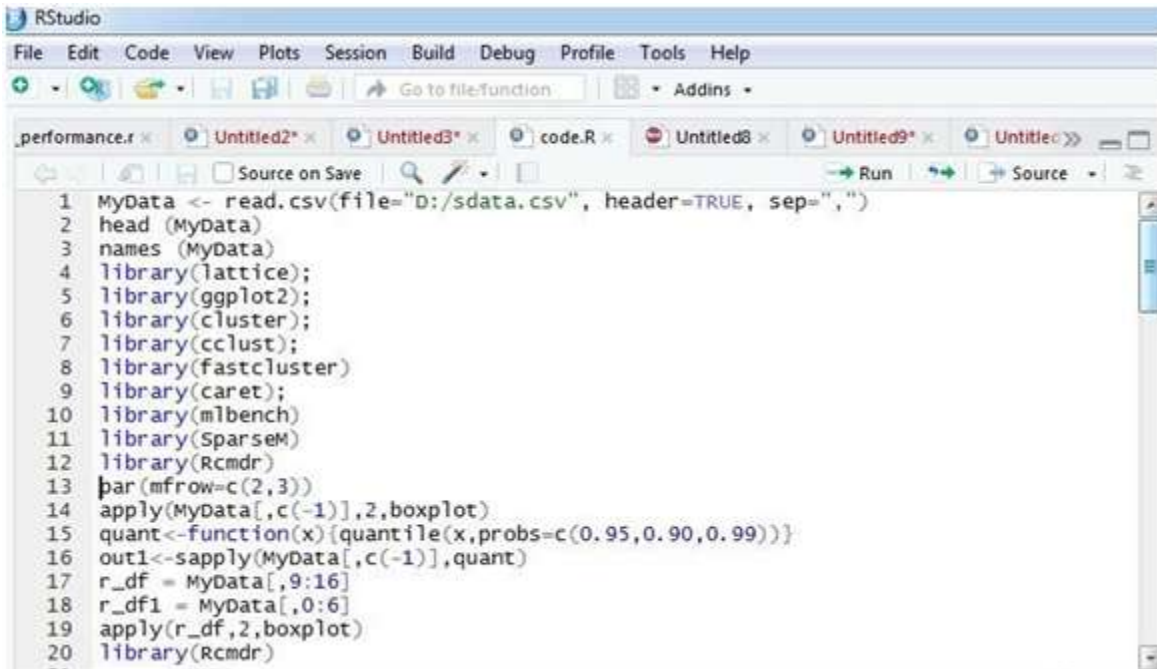


Figure 4.5 Import of data

The above figure describe that it is necessary to read the data from where it is stored and then these data is taken in the data frame which is used in R studio for

implementation, Here it is taken in ("MyData"), then head(MyData) is apply by which all information regarding dataset is run in R console.

Cluster Size By Elbow Method



```

1 MyData <- read.csv(file="d:/sdata.csv", header=TRUE, sep=",")
2 head(MyData)
3 names(MyData)
4 library(lattice);
5 library(ggplot2);
6 library(cluster);
7 library(cclust);
8 library(fastcluster)
9 library(caret);
10 library(mlbench)
11 library(sparseM)
12 library(Rcmdr)
13 par(mfrow=c(2,3))
14 apply(MyData[,c(-1)],2,boxplot)
15 quant<-function(x){quantile(x,probs=c(0.95,0.90,0.99))}
16 out1<-sapply(MyData[,c(-1)],quant)
17 r_df = MyData[,9:16]
18 r_df1 = MyData[,0:6]
19 apply(r_df,2,boxplot)
20 library(Rcmdr)
  
```

Figure 4.6 Cluster size

The above figure describe that it is necessary to install the various packages for implementing the modified K-means clustering, then all the packages library are taken for the elbow method. Here the library ggplot2, caret, Rcmdr are most important because using this library the graph is plotted and the R commander is loaded and using the Rcmdr and the stats library the elbow method is implemented and the line chart is there, by looking the arm

the elbow point is determined and the cluster size is there by elbow method. Preprocessing of the data is there, normalized data is there, raw data is also there then the useful data is determined by the operation. There is a boxplot which describe the data. For the cluster size by elbow method the sum of squared is there and there is graph for elbow method.

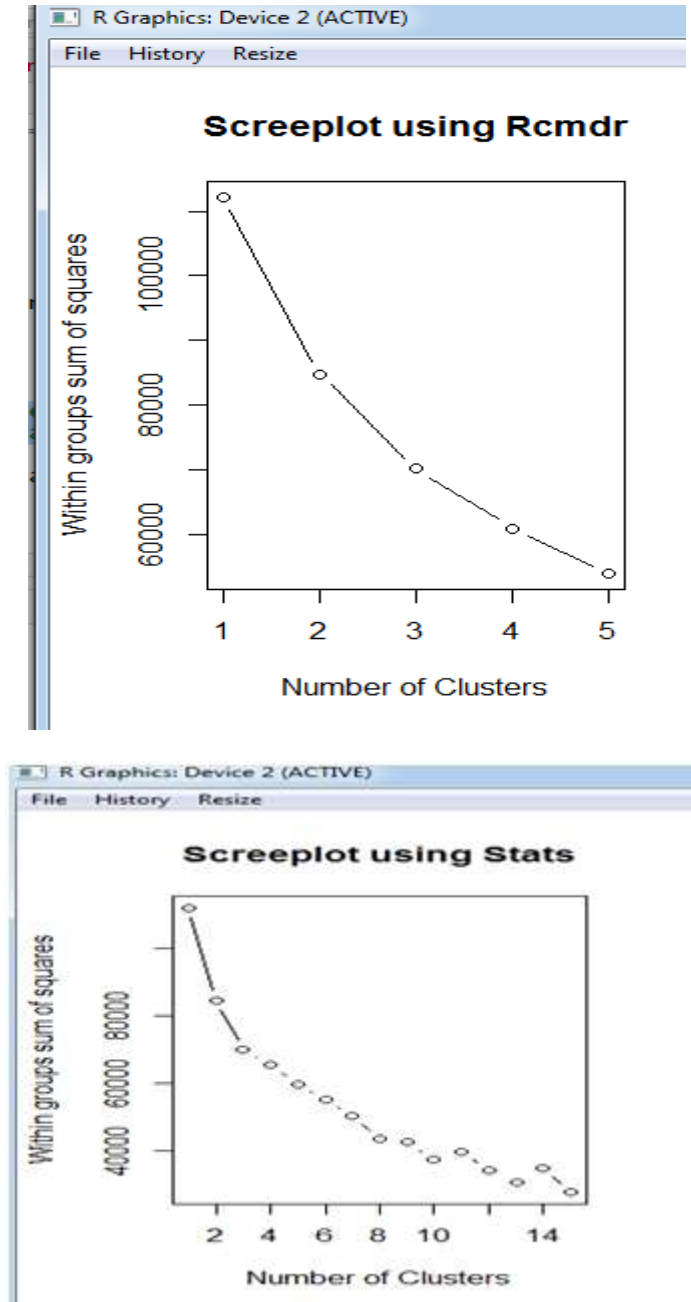


Figure 4.7 Elbow methods Using Rcmdr and Stats

Plotting Graph

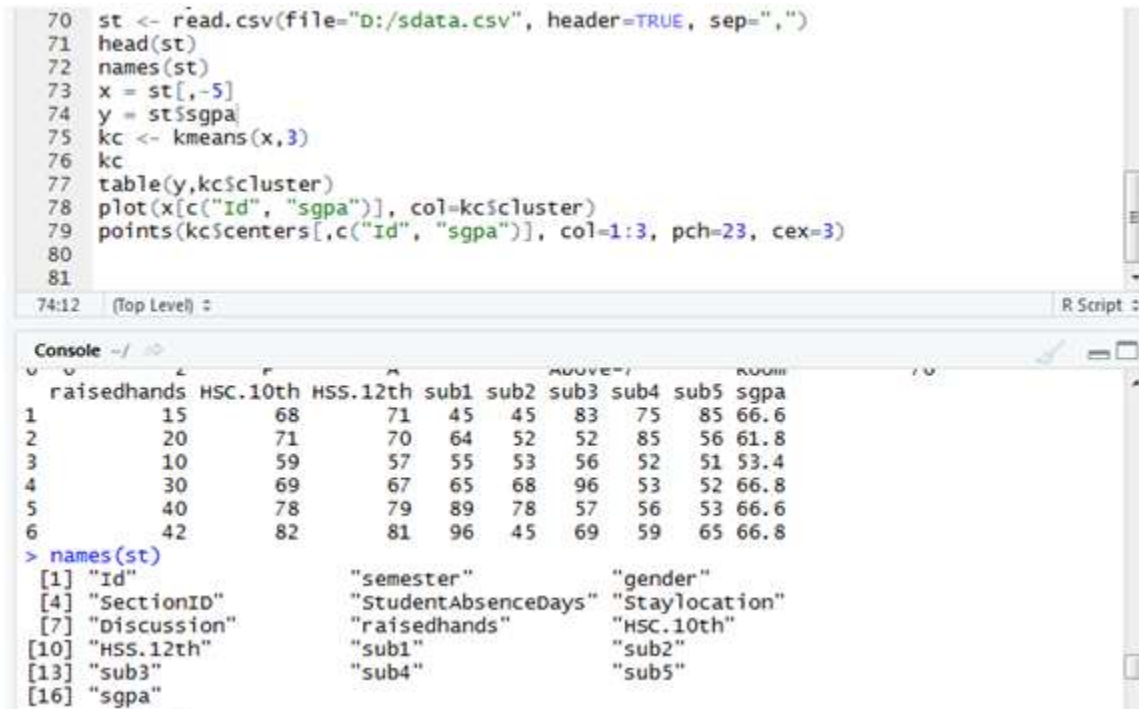
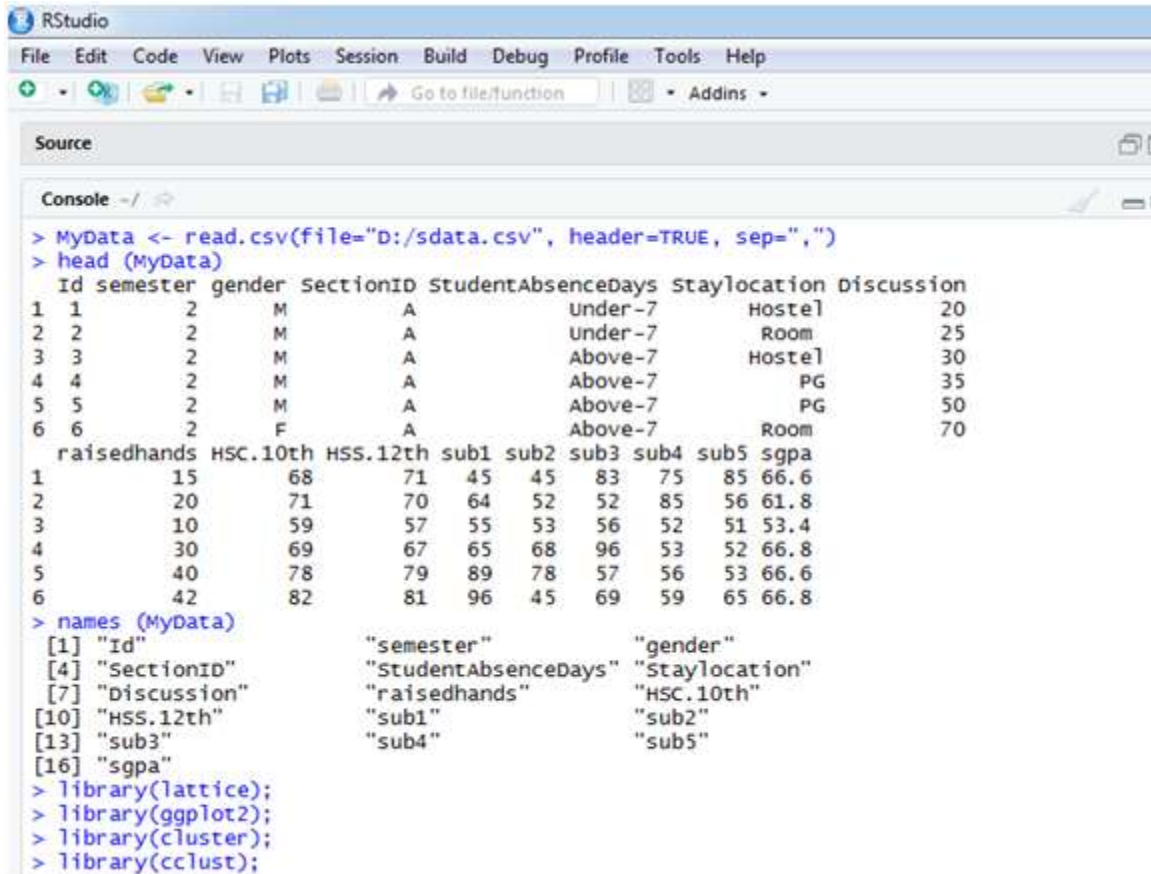


Figure 4.8 Graph Plotting

The above Figure describe that for plotting the graph it is necessary to run the library first, then data is import and then the algorithm is applied, points are taken for the graph and the parameter is taken by which it shown the axis, Here there is graph between id and sgpa by taking k=3, it can be change according to condition. But here there is elbow method using the Rcmdr and Stats and in the line cart it is clear that there is an arm near the K=3, so here the cluster

size is 3, by applying the elbow method it will give optimal solution. In the line chart of elbow method, there is x axis any y axis and in the x axis there is cluster size and in the y axis there is sum of squared error within the data. For the elbow method other packages are also installed and its library is imported as pee need. So the elbow method is important in the clustering method.

Console of R studio



```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
Source
Console -/
> MyData <- read.csv(file="D:/sdata.csv", header=TRUE, sep=",")
> head(MyData)
  Id semester gender SectionID StudentAbsenceDays Staylocation Discussion
1  1         2     M          A           Under-7           Hostel          20
2  2         2     M          A           Under-7           Room            25
3  3         2     M          A           Above-7          Hostel            30
4  4         2     M          A           Above-7           PG              35
5  5         2     M          A           Above-7           PG              50
6  6         2     F          A           Above-7           Room              70
  raisedhands HSC.10th HSS.12th sub1 sub2 sub3 sub4 sub5 sgpa
1           15        68        71   45  45   83   75   85 66.6
2           20        71        70   64  52   52   85   56 61.8
3           10        59        57   55  53   56   52   51 53.4
4           30        69        67   65  68   96   53   52 66.8
5           40        78        79   89  78   57   56   53 66.6
6           42        82        81   96  45   69   59   65 66.8
> names(MyData)
 [1] "Id"
 [4] "SectionID"
 [7] "Discussion"
[10] "HSS.12th"
[13] "sub3"
[16] "sgpa"
      "semester"
      "StudentAbsenceDays"
      "raisedhands"
      "sub1"
      "sub4"
      "gender"
      "Staylocation"
      "HSC.10th"
      "sub2"
      "sub5"
> library(lattice);
> library(ggplot2);
> library(cluster);
> library(cclust);
  
```

Figure 4.9 Console of R

Figure shows the second quadrant of R studio, when any script is run in R it will show in the console, here when K means is run and data is import it shows in console, it imports the data, make the cluster and vector which is very important. There are four quadrants in the R studio and each has its feature.

Here on applying the algorithm it make the cluster of size 12, 6, and 32 now it will plot the graph in the fourth quadrant of the R studio. The cluster size is taken by the elbow method and by the names function all the names of the parameter are taken. There is clustering vector and the sum of squared within the cluster. The console of the R is has its importance because all the operation which is running is there in this quadrant only.

Graph between Id and Sgpa When K=3

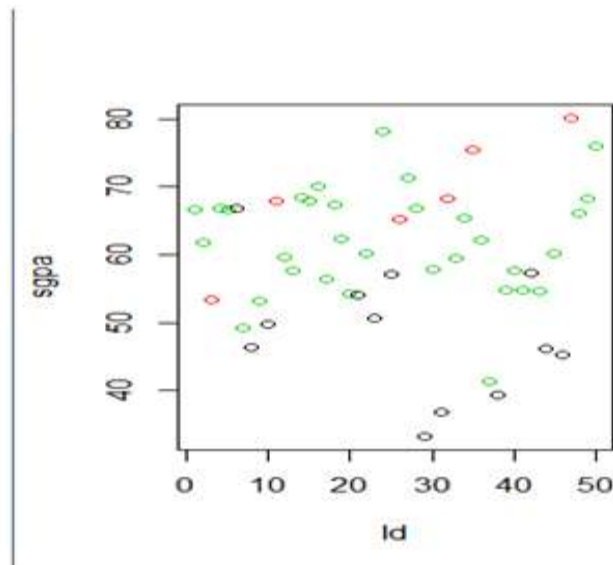
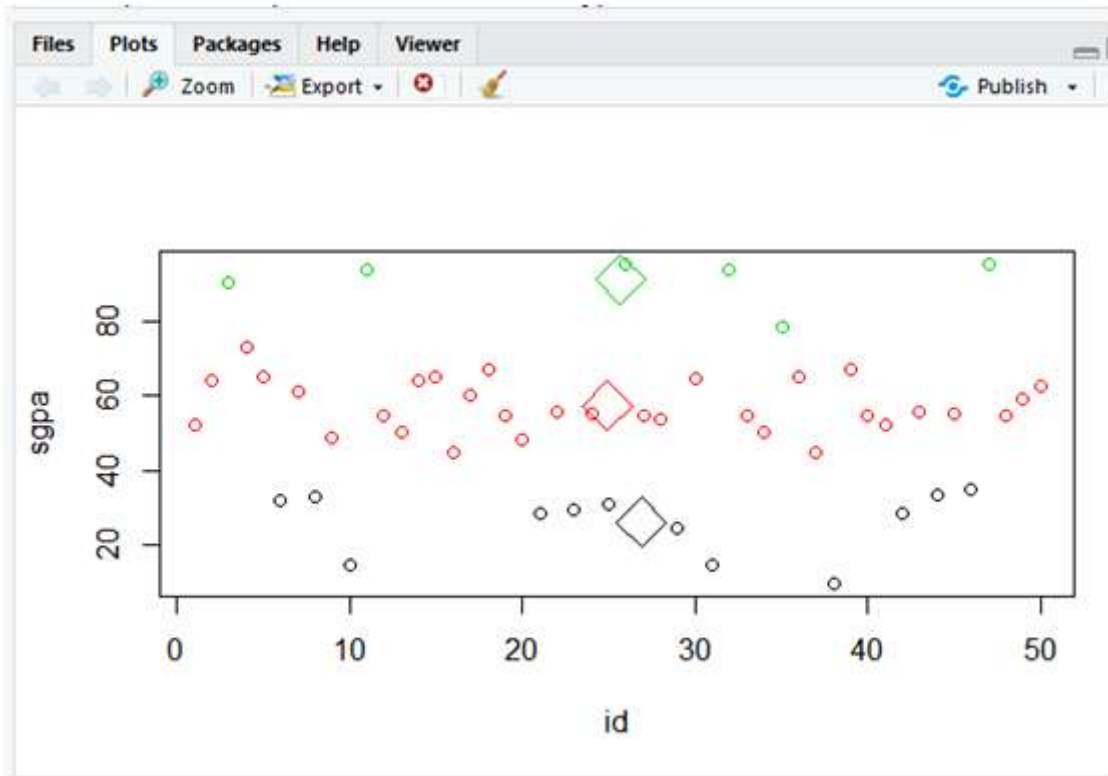


Figure 4.10 Graph between Id and sgpa

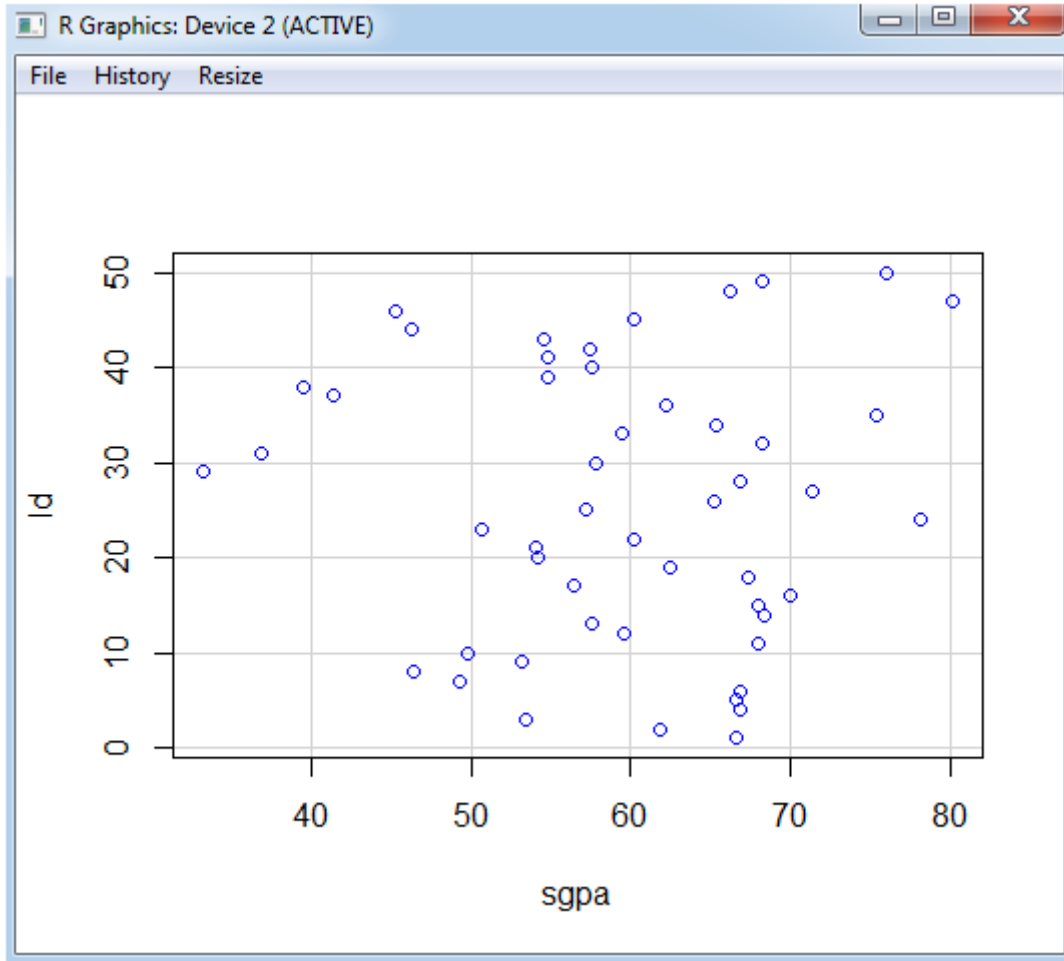


Figure 4.11 Scatter graph of sgpa

Chapter Summary

Modified k-means algorithm is used; the cluster size is determined by elbow method. There is a graph between id and sgpa, The R studio is used in which packages are installed then the library is taken and data is imported. The operation is done by taking cluster size which is taken from elbow point, the graph is plotted.

X. CONCLUSION

Present studies shows that academic performances of the students are primarily dependent on their past performances. Our investigation confirms that past performances have indeed got a significant influence over students' performance. Further, we confirmed that the performance of neural networks increases with increase in dataset size.

Machine learning has come far from its nascent stages, and can prove to be a powerful tool in academia. In the future, applications similar to the one developed, as well as any improvements thereof may become an integrated part of every academic institution.

REFERENCES

- [1] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Preventing student dropout in distance learning systems using machine learning techniques," AI Techniques in Web-Based Educational Systems at Seventh International Conference on Knowledge-Based Intelligent Information & Engineering Systems, pp. 3-5, September 2003.
- [2] B.K. Bharadwaj and S. Pal. "Data Mining: A prediction for performance improvement using classification", International Journal of Computer Science and Information Security (IJCSIS), Vol. 9, No. 4, pp. 136-140, 2011.