



International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 11, Issue 04, April 2022)

A Machine Learning Approach to Predicting Student Performance

Muskan Chakrawarti¹, Prof. Jayesh Jain², Prof. Saurabh Sharma³, Prof. Vishal Paranjape⁴
^{1,2,3,4}Global Nature Care Sangathan Group of Institutions, Jabalpur (M.P), India

Abstract: Machine learning is used in a variety of fields, including education, pattern identification, games, business, social media services, online customer support, and product recommendations. The future of the children is making the educational system more important. There is a great amount of data in higher education because today's students all want to go to college, which raises the need for M.L. procedures in the educational system. For the purpose of examining student performance, many tools are available. The review of student data will be aided by data mining, which is used to unearth buried information. In the realm of education, there is a tonne of data, and all of it is helpful to both teachers and pupils. As the institute grows, the utilisation of M.L. technology in the classroom is taking on more significance. Clustering is one of the core techniques widely used in data analysis. Modified K-means is among the most well-liked and efficient clustering techniques, albeit there are additional ones. There are numerous methods for classifying data, with decision trees being the most popular. Decision trees are commonly used to analyse student performance even though they are less stable than modified K-means. talk about unsupervised algorithms. These use cluster analysis to classify students into groups based on their traits. The cluster size can be calculated using the elbow approach, which will help in determining the optimal solution. There is an elbow method that incorporates the elbow point and looks across the arm in the sum of the squares. It is simple to improve children's performance and future with the M.L. technique. Not only can students raise their performance, but so can institutions and teachers.

Record Terms – Prediction using SVM, Machine Learning.

I. INTRODUCTION

Systems can learn on their own thanks to the learning branch of artificial intelligence (AI machine). They have an automatic mechanism for learning. Additionally, experience can be used to improve the system. Machine learning finds patterns in the data for improved outcomes.

There are several applications for pattern recognition science. Machine learning is related to computational statistics, which emphasises computer-based prediction. The main goal of the machine learning idea known as "data mining" is data analysis.

Machine learning is significant because it enables models to change to accommodate fresh data. They draw lessons from previous calculations to produce decisions and outcomes that are trustworthy and reproducible. The machine learning branch of AI organises a lot of data into modules that people can use. Computer science's machine learning field differs from traditional computational approaches in two ways. A collection of programmes are employed in traditional computing to carry out calculations. Machine learning generates outcomes from data input using a variety of methods, including statistical research.

Any higher organization's main priority is regaining administrative result generation. The evaluation of students' performance in esteemed institutions is one of the pillars for boosting educational standards. Student performance is a significant and essential element in higher education facilities. This is true because the calibre of their knowledge is determined by institutions' impressive record of academic triumphs. There is a tonne of data produced by the educational system that can be utilised in study. Analysis of the data is therefore even more important now. Data mining in education is therefore important and useful today.

Machine learning uses the performance in the past to improve performance in the future. Learning in this context refers to the optimization of the algorithm and subsequent use of the optimised algorithm. There are rules in place, and if something can't learn, it can't be said to be intelligent. Therefore, the most important aspect of an intelligent system is its capacity for learning [1].

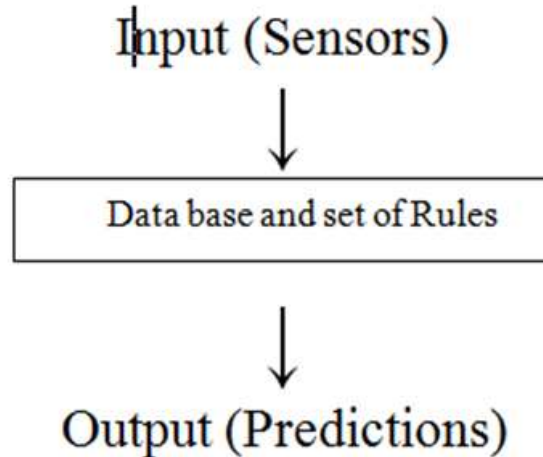


Figure 1.1 Machine Learning

Machine learning has several applications, including fraud detection and product recommendations. There are a lot of e-commerce companies that use this highly important application. For instance, if we buy a phone, we can be persuaded to buy the case as well. Machine learning is a concept that is applied in social networks to suggest friends.

Automatic Predicting student achievement is a big issue because educational databases contain a lot of information.

To complete this goal, educational data mining is being used (EDM). EDM develops methods for finding data generated in educational contexts. Using these methods, one may comprehend students and their learning environment. Educational institutions sometimes question how many students will pass or fail in order to make the required plans. In previous studies, it has been noted that many researchers concentrate on selecting the best algorithm for just classification and neglect to find solutions to problems that emerge during the data mining phases, such as data high dimensionality, class imbalance, and classification error, among others. The model was less accurate as a result of these problems. Although there are many popular categorization techniques utilised in this area, the model for predicting student achievement that this research suggested is based on supervised learning decision trees. Additionally, an ensemble method is applied to improve the performance of the classifier. To solve problems with categorization and prediction, ensemble approaches are used. This study shows the need of gathering data and honing algorithms to deal with issues with data quality. Local to Portugal's Alentejo region, the experimental data set used in this study is from the UCI Machine Learning Repository.

Three supervised learning algorithms—J48, NNge, and MLP—are experimentally used in this study. The results showed that J48 performed better than all other models, with an accuracy rate of 95.78%.

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

In order to improve curriculum design and prepare interventions for academic support and guidance on the curriculum provided to the students, it is frequently important to be able to predict the behaviour of future students. Data mining (DM) is used in this situation. For later usage, DM approaches examine datasets to extract information and restructure it into comprehensible forms. Recommender Systems (RS), Collaborative Filtering (CF), Machine Learning (ML), and Artificial

The primary computer methods used to process this data to forecast kids' performance, grades, or danger of dropping out of school are neural networks (ANN). Predicting students' behaviour is one of several related areas of interest in the field of education that have generated a significant amount of research in recent years. In fact, there are a tonne of studies on this subject that have been presented at conferences and in journals.

Consequently, the primary objective of this study is to provide an in-depth analysis of the various strategies and algorithms that have been presented and used in this field.

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 & CHARACTERISTICS

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

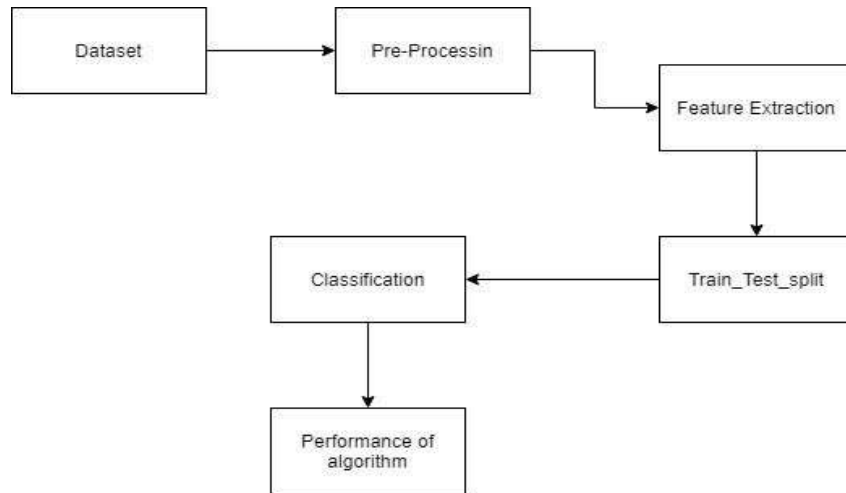


Fig 2: System Architecture

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 usefulas 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

The modified K-Means algorithm is utilised to analyse the performance of the students. These saw the division of n objects into k clusters. By examining the closest mean, the data are positioned. It implies that data of the same type are grouped together, whereas data of different types are found in separate groups. The basis for analysis is a parameter with identical data. Numerous applications, including market research, pattern recognition, etc., use clustering analysis extensively. The elbow approach is used to calculate modified K means for the cluster size. The best answer will be obtained by utilising the elbow approach to obtain the number of clusters, as using a random integer for cluster size may not produce it. Calculating the sum of squares is the idea underlying the elbow method. The sum of squared values for the K value range is plotted in a line-chart-like manner (K may have any number of value). This line graph resembles an arm, and the valve at the elbow of the arm corresponds to the appropriate value of K. The most important thing is to select a small K value with a low sum of squared values. It is simple to execute and produces the best outcome. Python, R, SAS, and other languages are only a few of the many that are frequently used for analytical purposes.

To implement the modified K-means algorithm in the R language, a number of packages must be installed, after which the library of those packages can be used to plot the elbow technique and various graphs between parameters. There are numerous variables that are utilised to analyse student performance, but there are also numerous variables in the dataset that have no bearing on the outcome.

Here, a parameter like result is crucial for the analysis, but a parameter like gender, which cannot be used to determine results, is crucial for the analysis's goal of separating the genders and providing accurate data. There are many options in R Studio, including a help option that allows anyone to get information about any library. As a result, R Studio offers a wide range of options, and the elbow approach is helpful for determining cluster size.

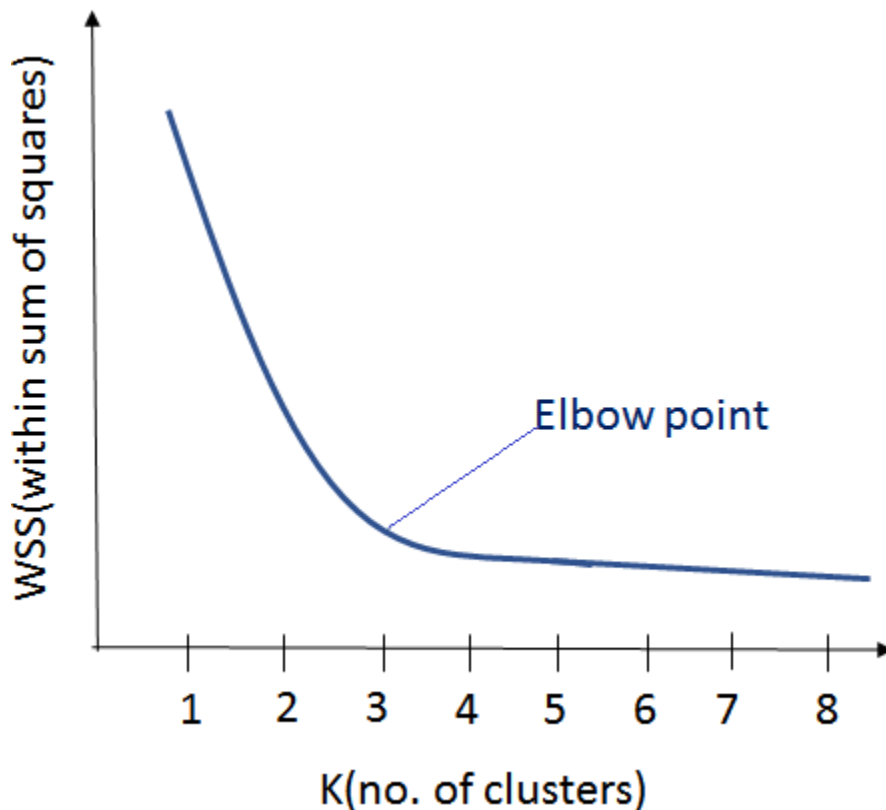


Figure 4.1 Elbow Method

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 graphically 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. 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.

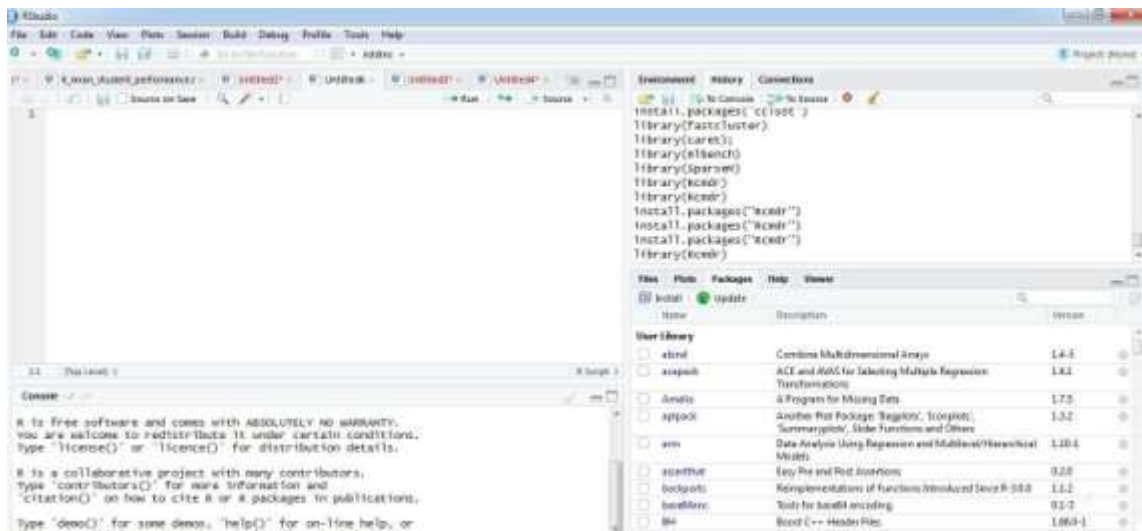


Figure 4.2 R studio

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.

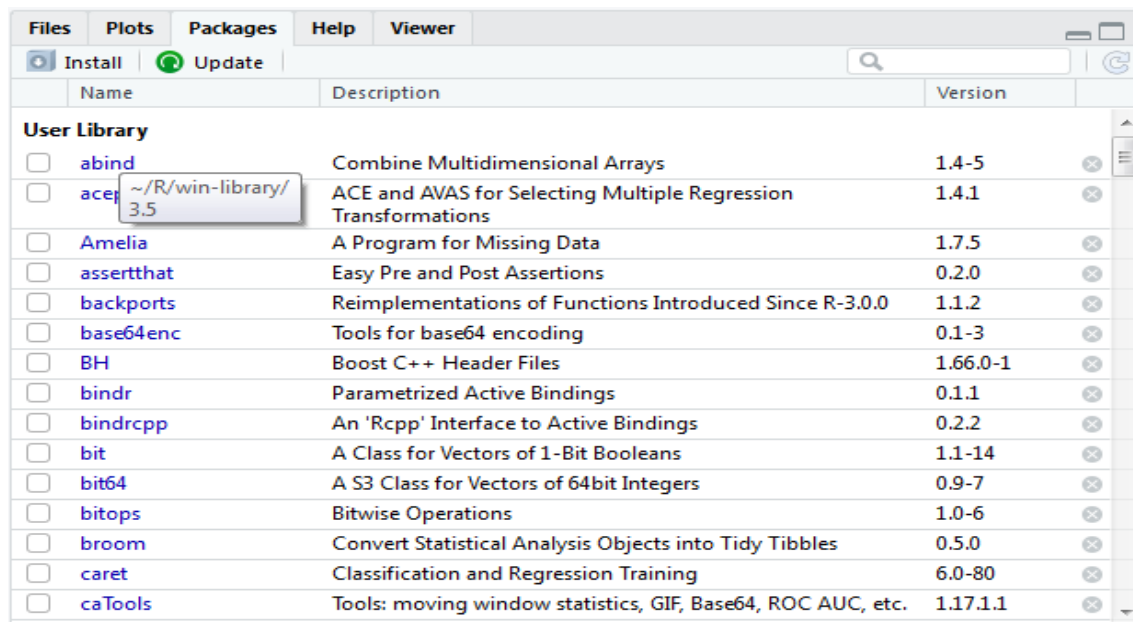


Figure 4.3 Package in R studio

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	Discussion	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	76	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

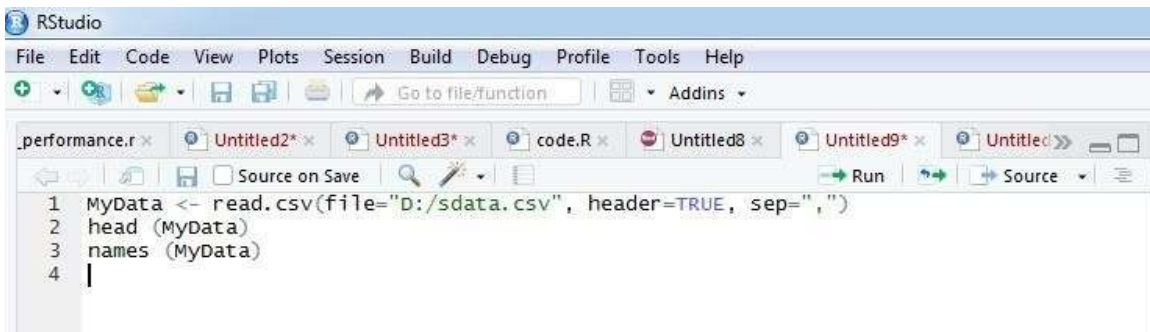
Figure 4.4 Student dataset

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.



```

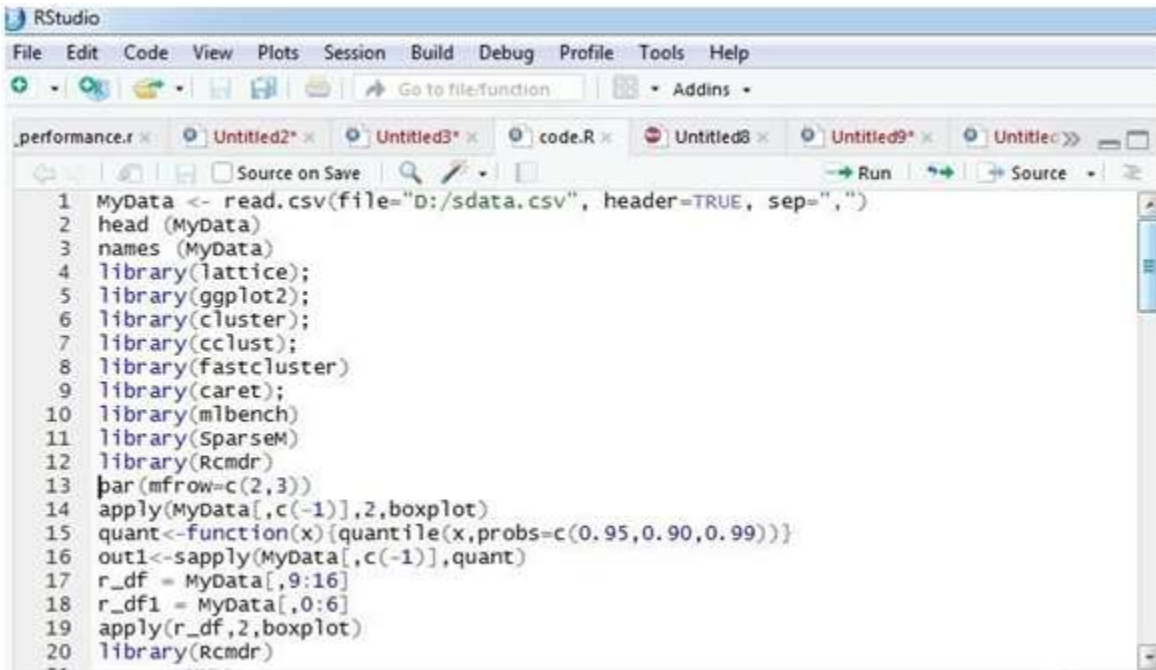
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function Addins
performance.r x Untitled2 x Untitled3 x code.R x Untitled8 x Untitled9 x Untitled10 x
Source on Save Run Source
1 MyData <- read.csv(file="D:/sdata.csv", header=TRUE, sep=",")
2 head(MyData)
3 names(MyData)
4 |
  
```

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.

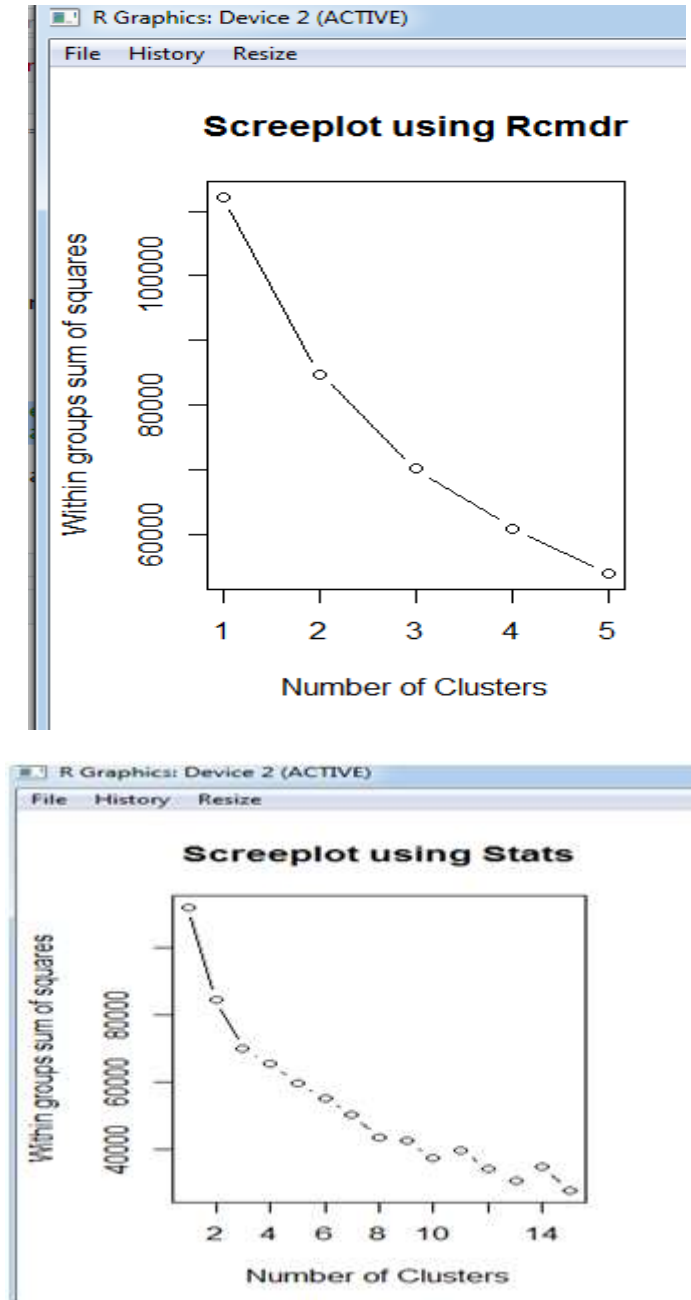


Figure 4.7 Elbow methods Using Rcmdr and Stats

Plotting Graph

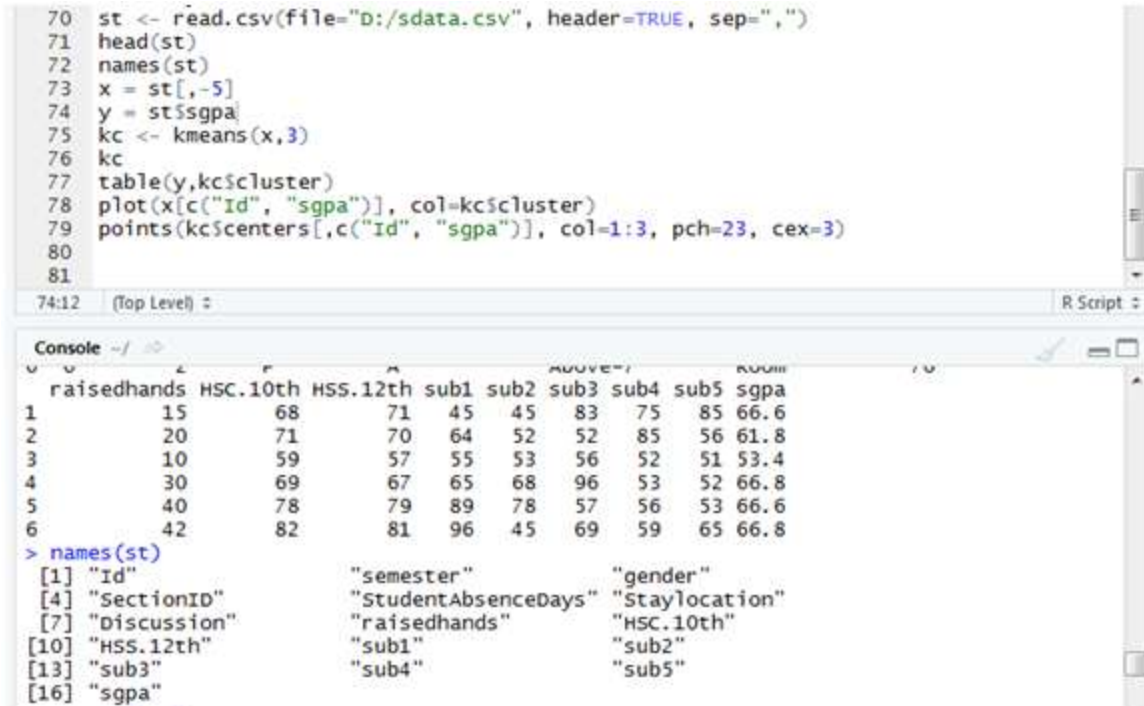
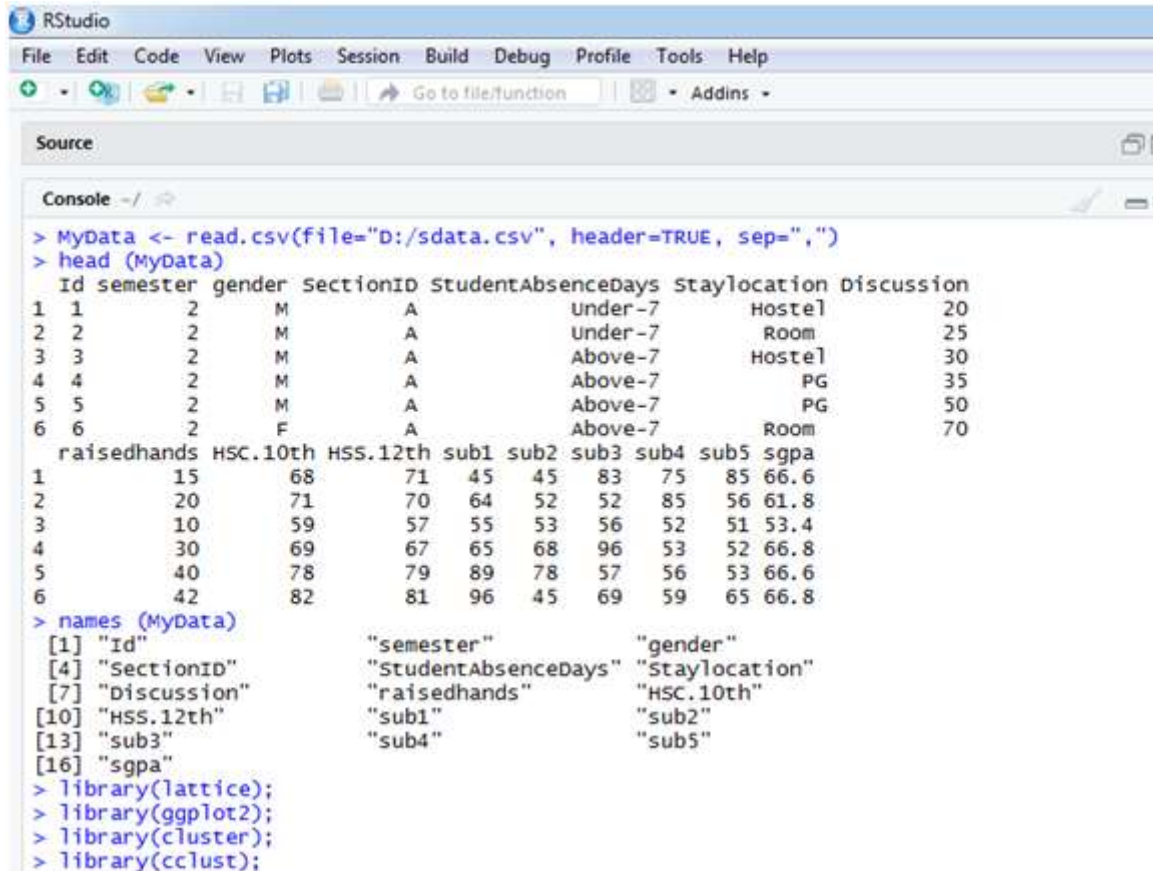


Figure 4.8 Graph Plotting

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"           "semester"     "gender"
 [4] "sectionID"   "StudentAbsenceDays" "Staylocation"
 [7] "Discussion"  "raisedhands"   "HSC.10th"
[10] "HSS.12th"   "sub1"         "sub2"
[13] "sub3"       "sub4"         "sub5"
[16] "sgpa"
> 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 $K=3$

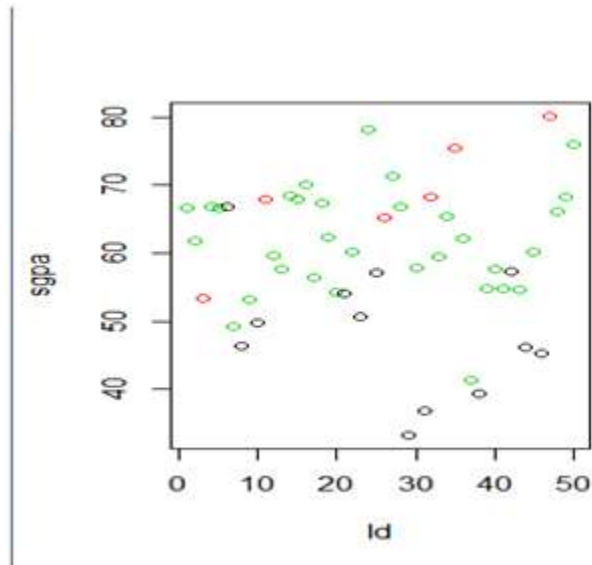
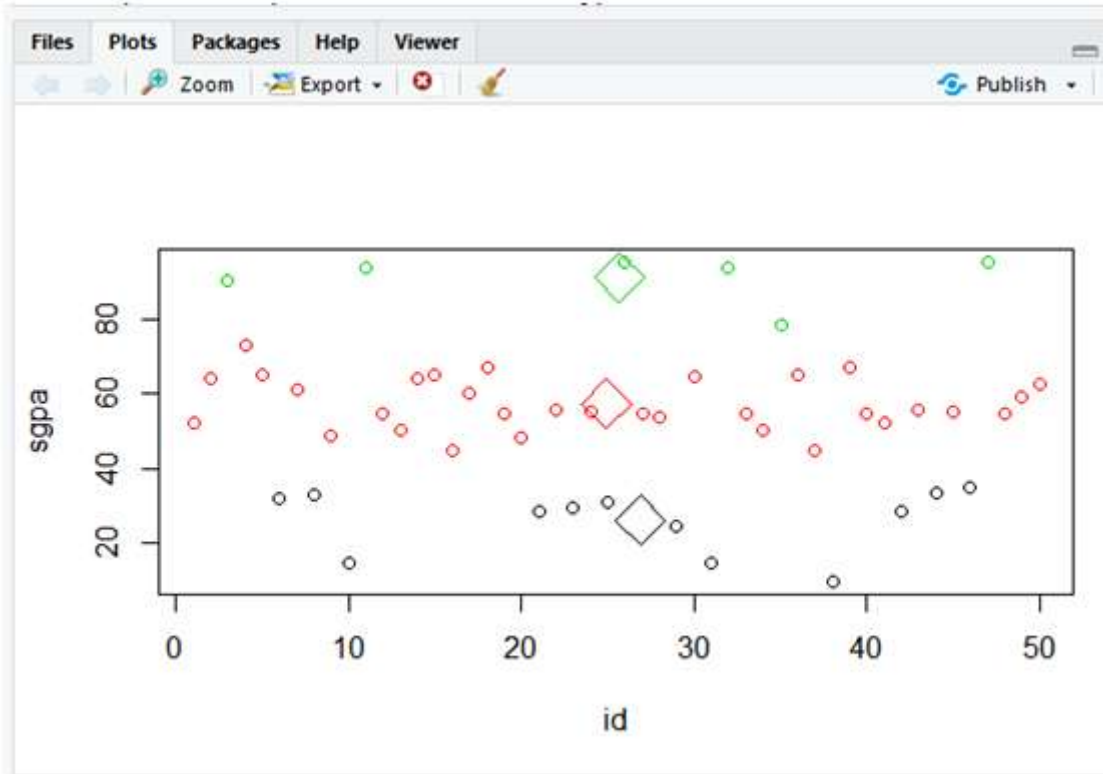


Figure 4.10 Graph between Id and sgpa

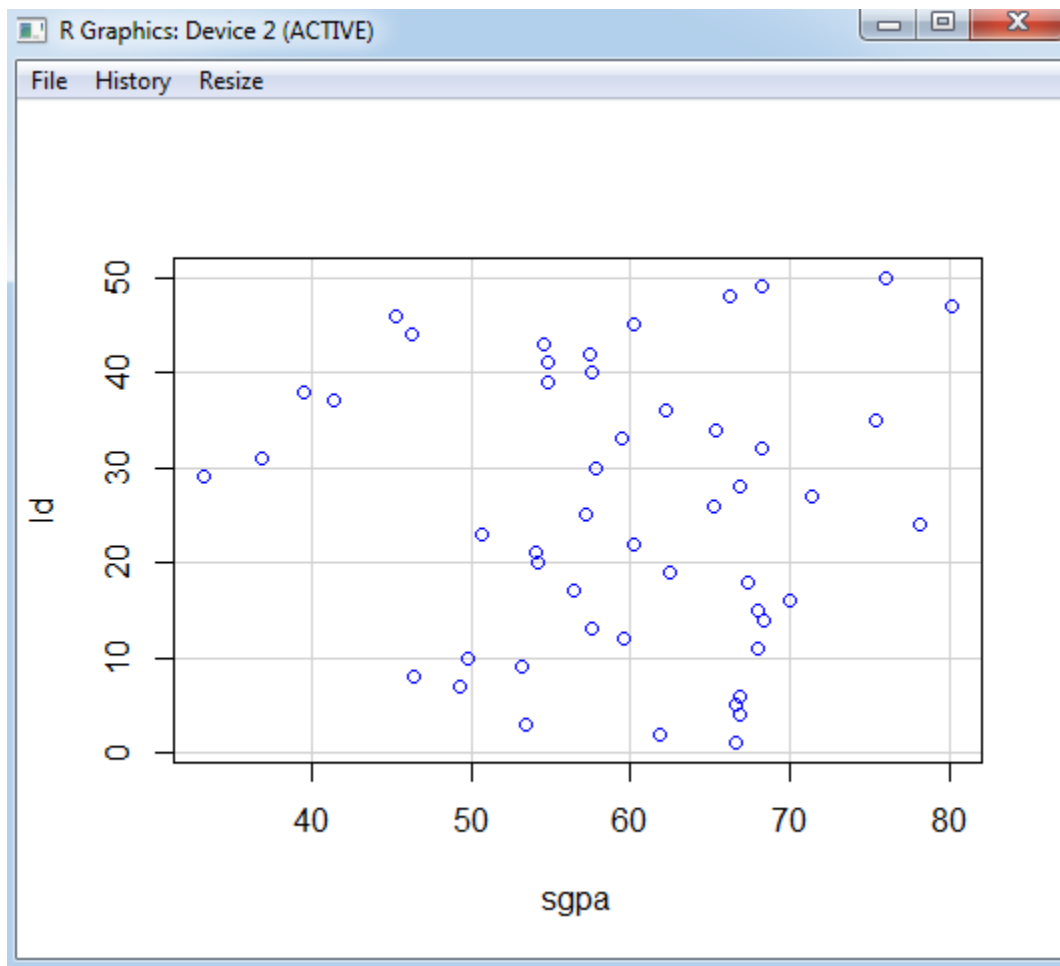


Figure 4.11 Scatter graph of sgpa

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.