A STUDENT PERFORMANCE PREDICTION APPROACH USING MACHINE LEARNING

**Abstract**: Education, pattern recognition, games, business, social media services, online customer support, and product recommendations are just a few of the industries that use machine learning. The importance of the educational system is increasing due to the children's future. Higher education generates a lot of data because all of today's students desire to attend college, which increases the demand for M.L. operations in the educational system. There are numerous instruments available for the purpose of evaluating student performance. Data mining, which is used to find hidden information, will help with the review of student data. There is a tonne of data in the field of education, and all of it is beneficial huge both teachers and students. The use of M.L. technology in the classroom is becoming more important as the institute expands. One of the fundamental methods frequently employed in data analysis is clustering. Among the most popular and effective clustering methods is modified K-means, though there are others as well. Data classification techniques abound, with decision trees being the most widely used. Though less stable than modified K-means, decision trees are frequently employed to analyse student performance. Discussion about unsupervised algorithms These divide students into groups according to attributes using cluster analysis. The elbow approach can be used to compute the cluster size, which will aid in finding the best answer. There is an elbow technique that includes the elbow point and scans the length of the arm. The M.L. approach makes it easy to enhance children's performance and future. Institutions and teachers can improve performance in addition to pupils.

**Record Terms** – Prediction using SVM, Machine Learning.

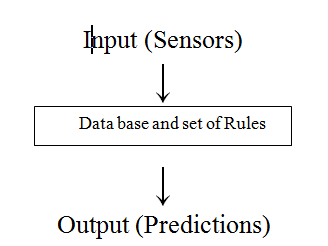
# INTRODUCTION

Thanks to the learning subfield of AI, systems can learn on their own (AImachine). They have an automatic learning system. Experience can also be used to enhance the system. For better results, machine learning identifies patterns in the data. The study of pattern recognition has many applications. Computational statistics, which emphasises computer-based prediction, is connected to machine learning. Data analysis is the primary objective of the machine learning concept known as "data mining."

Machine learning is important because it allows models to adapt to new data. To produce conclusions and results that are reliable and repeatable, they learn from earlier calculations. A lot of data is organised into usable modules by the machine learning area of artificial intelligence. Machine learning in computer science differs from conventional computational methods in two ways. In conventional computing, a number of programmes are used to perform calculations. Data input is used to produce results using a variety of techniques, including statistical analysis.

Regaining administrative result generation is the top priority for any higher organisation. One of the foundations for raising educational standards is the appraisal of students' performance in prestigious institutions. An important and crucial component of higher education facilities is student performance. This is true since institutions' remarkable track record of academic victories determines the calibre of their expertise. The educational system generates a tonne of data that can be used for research. As a result, data analysis is now much more crucial. Therefore, data mining in education is crucial and valuable today.

Machine learning makes use of past performance to enhance future performance. In this sense, learning refers to the process of optimising an algorithm and then using that optimised algorithm. There are laws in place, and something cannot be considered intelligent if it cannot learn. The ability of an intelligent system to learn is therefore its most crucial component [1].ss



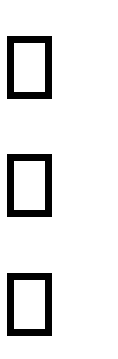
#### Figure 1.1 Machine Learning

Applications for machine learning include fraud detection and product recommendations. Numerous e-commerce businesses make use of this crucial application. We might be convinced to purchase the case if we purchase a phone, for instance. Social networks use the idea of machine learning to suggest friends.

Automatic Because educational databases contain a wealth of data, predicting student progress is a significant challenge. Using educational data mining, this objective will be accomplished (EDM). EDM creates tools for locating data produced in educational settings. These techniques allow one to understand students and their learning environment. In order to make the necessary preparations, educational institutions occasionally wonder how many students will pass or fail. It has been observed in earlier studies that many researchers focus on choosing the best algorithm for merely classification and fail to identify solutions to issues that arise during the data mining phases, such as data high dimensionality, class imbalance, and classification error, among others. These issues led to a decrease in the model's accuracy. The model for predicting student accomplishment that this research presented is based on supervised learning decision trees, despite the fact that there are other widely used classification techniques used in this field. An ensemble method is also used to enhance the classifier's performance. Ensemble techniques are used to address classification and prediction issues. This study demonstrates the necessity for data collection and algorithm development to address problems with data quality. The experimental data set used in this study is from the UCI Machine Learning Repository and is exclusive to Portugal's Alentejo region. In this investigation, three supervised learning algorithms—J48, NNge, and MLP—are experimentally used. With an accuracy rating of 95.78%, the results demonstrated that J48 performed better than all other models.

# 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.

# PROBLEM DEFINATION

It is frequently crucial to be able to forecast the behaviour of future students in order to improve curriculum design and create interventions for academic help and guidance on the curriculum supplied to the students. In this case, data mining (DM) is utilised. DM techniques analyse datasets to extract information and rearrange it into understandable forms for later use. Collaborative filtering (CF), machine learning (ML), recommender systems (RS), and artificial

Neural networks are the main computer algorithms used to forecast children's performance, grades, or risk of dropping out of school (ANN). One of the related topics of interest in the field of education that has generated a lot of research recently is predicting students' behaviour. In reality, there are a tonne of studies on this topic that have been published in journals and presented at conferences. As a result, the main goal of this research is to offer a thorough review of the numerous approaches and algorithms that have been proposed and applied in this area.

# 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

# 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.

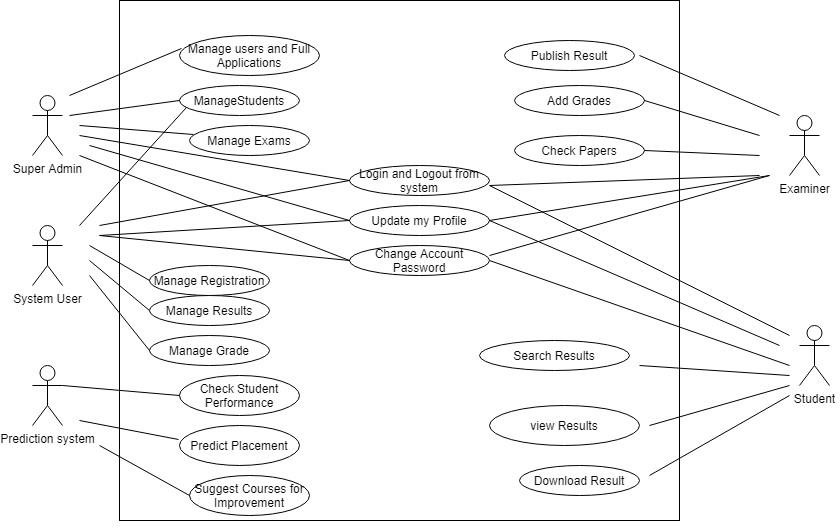


Fig 1: Use Case Diagram

# SYSTEM ARCHITECTURE

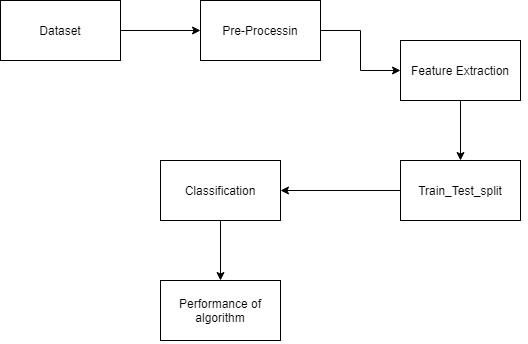


Fig 2: System Architecture

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

# 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.

# LIMITATION

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

# 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 technique is used to evaluate the pupils' performance. In these, n objects were split up into k groups. The data are placed by looking at the closest mean. It suggests that similar sorts of data are grouped together and other types are found in various groups. A parameter with identical data serves as the foundation for analysis. Clustering analysis is widely used in many areas, such as market research, pattern identification, etc. Modified K means are computed using the elbow method for the cluster size. The elbow method will be used to determine the number of clusters, as utilising a random integer for cluster size may not yield the optimum result. The elbow method's principle is to calculate the sum of squares. A line-chart-like depiction of the sum of squared values for the K value range is used to represent the data (K may have any number of value). The valve at the elbow of this line graph, which resembles an arm, corresponds to the right value of K. The most crucial step is choosing a K value with a modest sum of squared values. It is easy to carry out and results in the greatest results. Only a few of the many languages that are often used for analytical purposes include Python, R, SAS, and others. The elbow technique and various graphs between parameters can be plotted using the library of those packages after installing the necessary packages to implement the modified K-means algorithm in the R language. Numerous variables are used to analyse student performance, however the dataset also contains a large number of variables that have no influence on the result. In this case, a parameter such as result is important for the analysis, but a parameter such as gender, which cannot be used to predict outcomes, is important for the analysis's objective of separating the genders and delivering accurate data. R Studio has a wide range of choices, including a help option that enables anyone to learn more about any library. Because of this, R Studio provides many possibilities, and the elbow method is useful for figuring out cluster size.

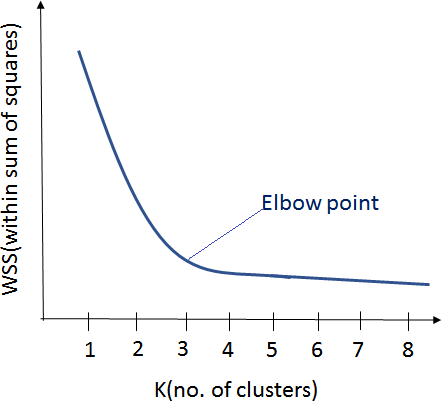


Figure 4.1 Elbow Method

**The R setup**

Programming in R is mostly used for statistical computing. R language mostly uses graphics for representation. R is freely available. It has an environment for analysis. It is accessible under the general public licence and is compatible with Windows, Linux, and Mac. R is software that also supports operations written in other languages, like C, C++, Python, and others. R is a very user-friendly, straightforward programming language that supports all loops, functions, and other programming concepts. Both a data handling facility and a storage facility are available in R. A function like matrix and vector exist in R. It has a wide range of data analysis tools. R was given that name because Ross and Robert created it, and the core development team oversaw its management. R is the statistics language that is utilised the most globally.

**Platform required running R studio**

Unix and Unix like systems

Linux

Windows XP/7/8

R studio server.

**R Studio**

The open source integrated development environment R Studio is free to use. It is used for statically computing and R programming. The studio also has a capability for graphic representation. R studio comes in two versions, R studio Desktop and R studio Web. The programme is operating locally as a regular application and has the appearance of a standard desktop application. Operating systems for the desktop version include Windows, Linux, and Mac.

Server for R studio: A web browser is present, and R Studio is accessed via it.

The qt framework, which is designed in both Java and the C++ computer language, is used by R studio for its graphical user interface. Because of how nicely organised its interface is, a user can view a graph, data tables, R code, and the output all at once. It includes several features that enable users to import different files, like csv, excel, etc. As seen in the diagram below, R Studio has four quadrants, each of which has a unique feature. The script is in the first quadrant, while the console is in the second. The results are generated in the console under all running conditions. The third quadrant, titled environment, contains all the variables and displays how each one interacts with its environment. The final quadrant of the graph is where more options, such as packages, files, and viewers, are available. Every quadrant has a distinct trait and significance. Each quadrant may be resized to suit the user's needs.s

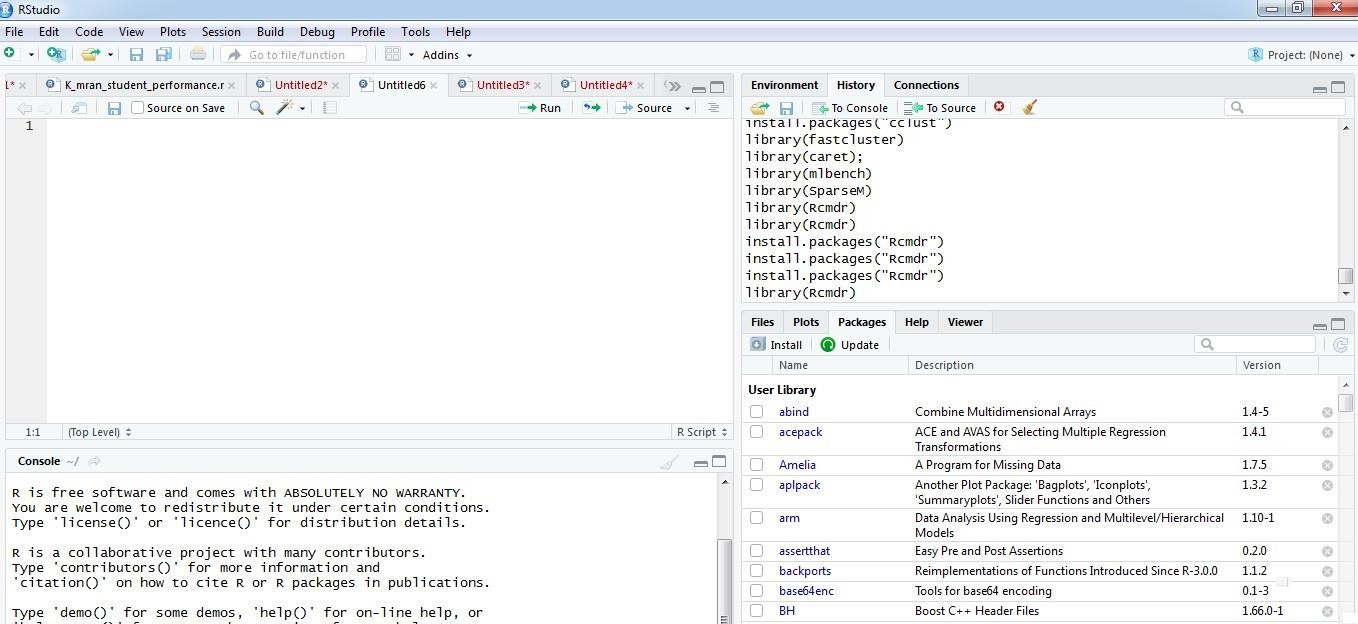


Figure 4.2 R studio

**R Packages**

R has a large number of packages, as well as a R function and a collection of R packages that contain it. R installs a number of packages during installation that are useful at a particular moment. Additional packages can be added at a later time as needed; there is a vast collection of packages that can be added as needed. All of these are kept in a directory called "library." By using it, the already-installed package can be loaded. When the console is started, a default package is automatically loaded. The package's library location is:.libpaths ()

The list of installed packages will be provided when we run library(). By using the command install.packages ("Package name"), we may also add new packages.

Some of the most popular packages in the R library are caret, ggplot2, Rcmdr, and Stats. The package can be installed in R in a variety of ways. Some options are available directly, while others need scripting. Numerous packages are dependent on one another, and before any of them can be utilised, they all need to be installed. When we require more information about a certain package, we can use the help option to view the product's information.

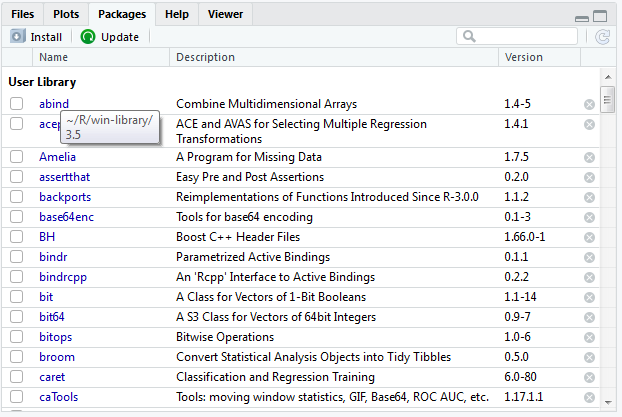


Figure 4.3 Package in R studio

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.



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,

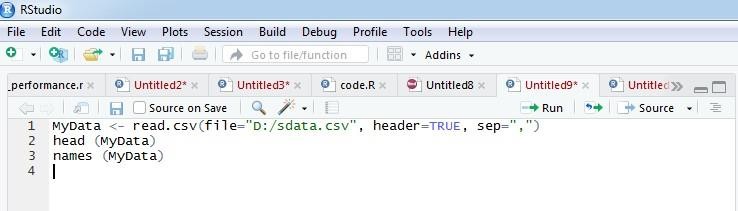
Import of Data

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.

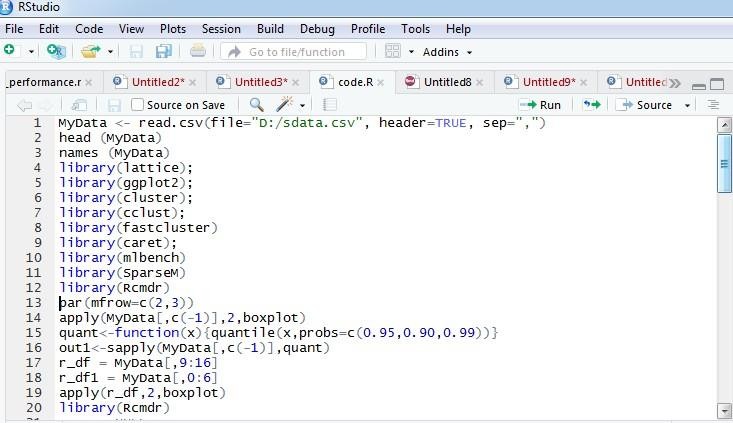
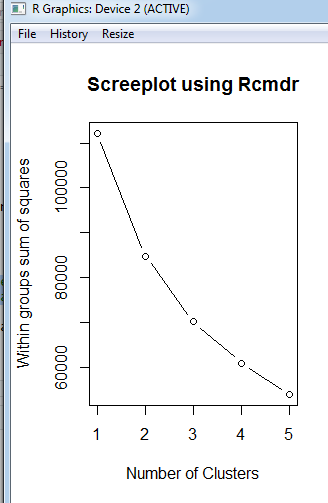
Cluster Size By Elbow Method

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.



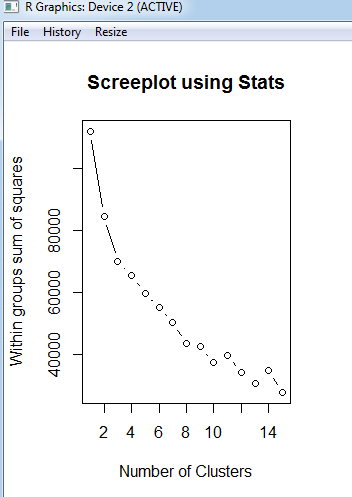


Figure 4.7 Elbow methods Using Rcmdr and Stats

Plotting Graph

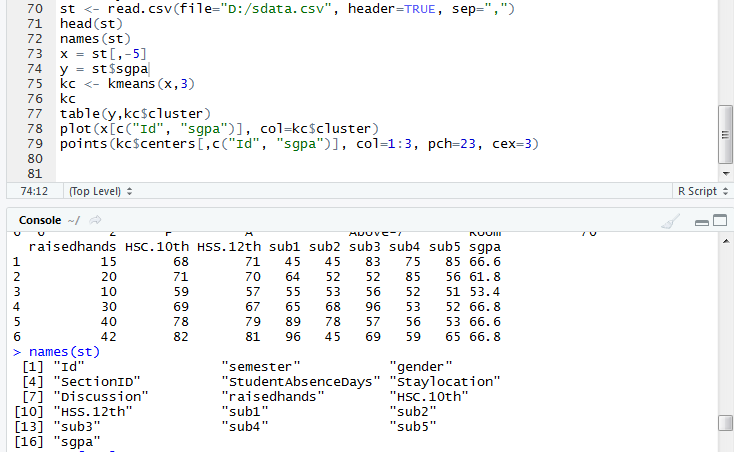


Figure 4.8 Graph Plotting

The above Figure explains that in order to plot a graph, the library must first be launched before any data is imported, the algorithm is then used, points are selected for the graph, and the parameter by which the axis is displayed is taken. The graph between Id and SGP is shown here. By using k=3, it can be modified based on the situation. However, the elbow approach is used here using Rcmdr and Stats, and it is obvious from the line cart that there is an arm close to K=3, meaning that the cluster size is 3. By utilising the elbow method, the best answer will be provided here because the cluster size is 3. The cluster size and the sum of squared errors within the data are shown on the x and y axes, respectively, of the line chart created using the elbow approach. Other packages are also set up and the elbow method's library is imported as necessary. In the clustering process, the elbow method is crucial.

**Console of R studio**

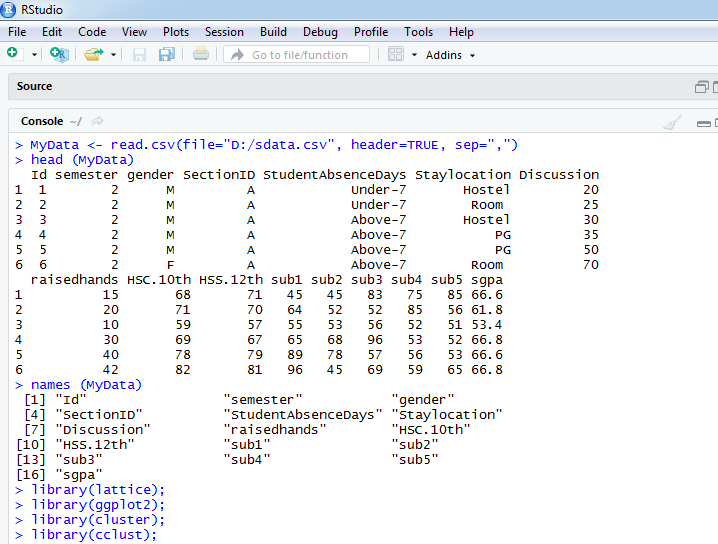
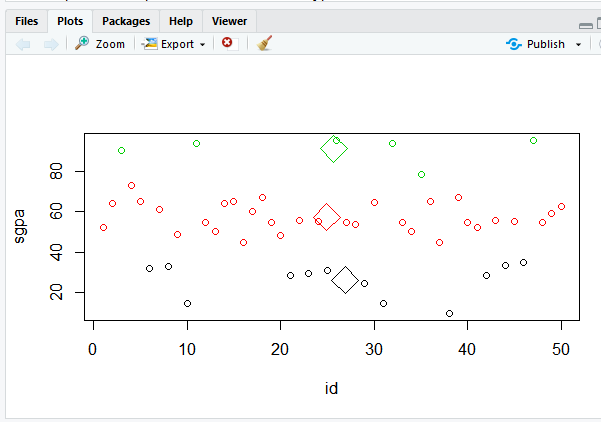
Figure 4.9 Console of R

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 size12, 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



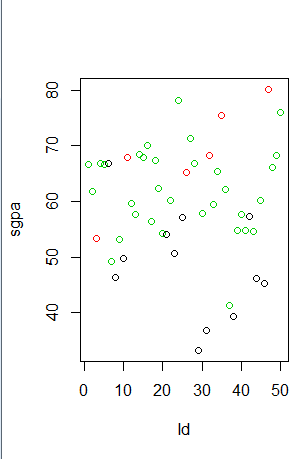


Figure 4.10 Graph between Id and sgpa

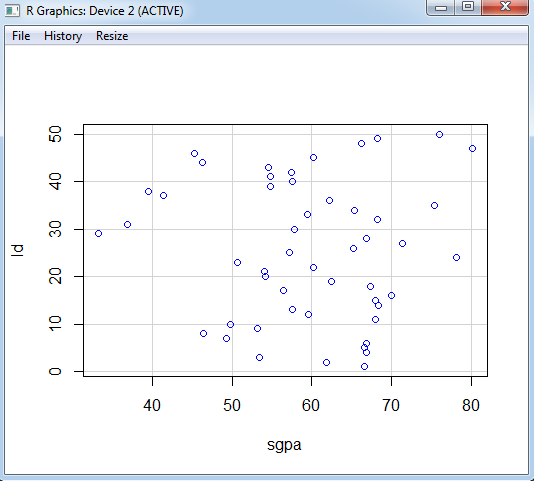


Figure 4.11 Scatter graph of sgpa

**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.

# CONCLUSION

* • According to recent studies, a student's academic success is largely influenced by their prior achievements. Our research demonstrates that past performance has a major impact on pupils' success. Furthermore, we verified that when dataset sizes grow, neural network performance does too. Machine learning has advanced significantly since its early beginnings and has the potential to be an effective tool in academia. Applications like the one created in the future, together with any enhancements thereto, might be integrated into every academic institution.

# REFERENCES

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.

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.