A STUDENT PERFORMANCE PREDICTION APPROACH USING MACHINE LEARNING

Abstract

Machine learning is used in a variety of fields, including education, pattern identification, gaming, business, social media services, online customer care, and product recommendations. The future of the children is a major factor in the importance of the educational system. All of today's kids want to go to college, which raises the need for M.L. operations in the educational system and leads to higher education producing a lot of data. For the objective of assessing student performance, many tools are available. Reviewing student data will benefit from data mining, which is a technique for discovering hidden information. The amount of information available in the subject of education is very helpful to both instructors and pupils. As the institute grows, it is becoming increasingly crucial to integrate M.L. technology in the classroom. Clustering is one of the core techniques widely used in data analysis. Modified K-means is one of the most well-liked and successful clustering techniques, while there are others as well. There are several methods for classifying data, with decision trees being the most popular. Decision trees are commonly used in analyses of student performance even though they are less stable than modified K-means. the topic of unsupervised algorithms is raised They use cluster analysis to classify the students into groups based on characteristics. The cluster size may be calculated using the elbow technique, which will help in determining the optimal solution. There is an elbow method that scans the length of the arm and the elbow point. It is simple to improve children's performance and

Authors

Geetanjali Mourya
Shri Ram group of Institution.

future using the M.L. technique. Together with students, institutions and teachers may boost performance.

Record Terms: Prediction using SVM, Machine Learning.

I. INTRODUCTION

Systems may learn independently thanks to the learning branch of AI (AImachine). They have a mechanism for automated learning. Also, experience may be utilised to improve the system. Machine learning finds patterns in the data for improved outcomes. There are several applications for the study of patterns. 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 use lessons from past computations in order to develop conclusions and findings that are trustworthy and reproducible. The artificial intelligence field of machine learning organises a lot of data into useful modules. Computer science machine learning varies from traditional computational techniques in two ways. Many programmes are used in traditional computing to carry out computations. Results are generated from data supplied using a number of methods, including statistical analysis.

Any higher organization's primary objective is to regain administrative result generating. The evaluation of students' performance at esteemed institutions is one of the pillars for boosting educational standards. Student performance is a critical and significant element of higher education facilities. This is accurate since the calibre of an institution's knowledge is determined by the impressive record of academic successes of the institution. A tonne of data produced by the educational system can be employed in study. Data analysis is becoming much more important as a result. Data mining is therefore essential and helpful in today's schooling.

The process of optimising an algorithm and then employing that algorithm is referred to as Machine learning uses the performance of the past to improve the performance of the present. learning in this sense. There are rules in place, and an object cannot be said to be intelligent if it is incapable of learning. An intelligent system's most important feature is consequently its capacity for learning.

Machine learning has a variety of uses, including fraud detection and product suggestions. Many e-commerce companies employ this essential application. If we buy a phone, for example, we could be persuaded to buy the case. Machine learning is a concept that social networks employ to propose friends.

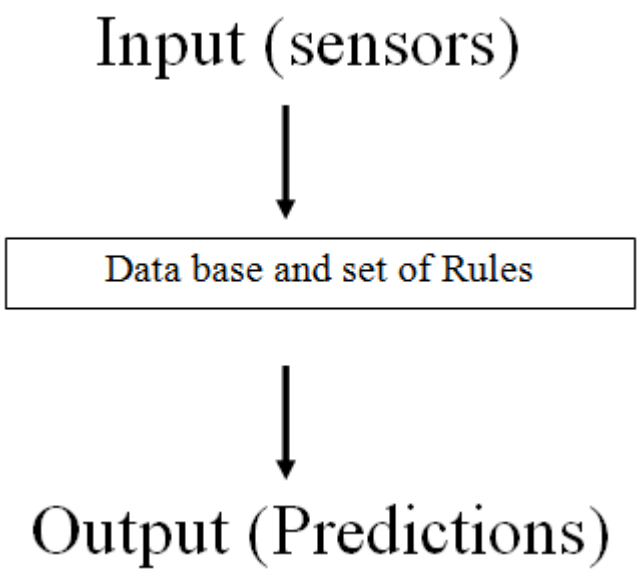


Figure 1: Machine Learning

Automatic the abundance of data in educational databases makes it difficult to forecast students' development. This goal will be attained through educational data mining (EDM). EDM develops resources for finding data generated in educational environments. One is able to comprehend students and their learning environment using these strategies. Educational institutions periodically question how many students will pass or fail in order to make the required preparations. In earlier studies, it was noted that many researchers concentrated on selecting the best method for just classification and neglected to find solutions to problems that arose during the data mining stages, such as data high dimensionality, class imbalance, and classification mistake, among others. The model's accuracy declined as a result of these problems. Despite the fact that this study's model for predicting student achievement is based on supervised learning decision trees, Despite the fact that this field uses other common categorization approaches, The performance of the classifier is additionally improved by using an ensemble technique. Issues with categorization and prediction are dealt with using ensemble approaches. This study highlights the need for algorithm development and data collecting to solve issues with data quality. The Alentejo area of Portugal is the only one represented in the experimental data set utilised in this work, which comes from the UCI Machine Learning Repository. Three supervised learning algorithms—J48, NNge, and MLP—are experimentally employed in this study. The findings showed that J48 performed better than all other models, with an accuracy rating of 95.78%.

II. MOTIVATION

To forecast a student's likelihood of being hired by a firm or their requirement for lessons.
The prospects for their futures are simply understood by students. in order to educate pupils about their future.
Improvement in doing tasks within the allotted time.

III. PROBLEM DEFINATION

The ability to predict future student behaviour is frequently essential for improving curriculum design and developing interventions for academic support and advice on the curriculum provided to the students. Data mining (DM) is applied in this scenario. To extract information and reorganise it into forms that are intelligible for subsequent use, DM approaches analyse datasets. Artificial intelligence (AI), recommender systems, collaborative filtering, and machine learning

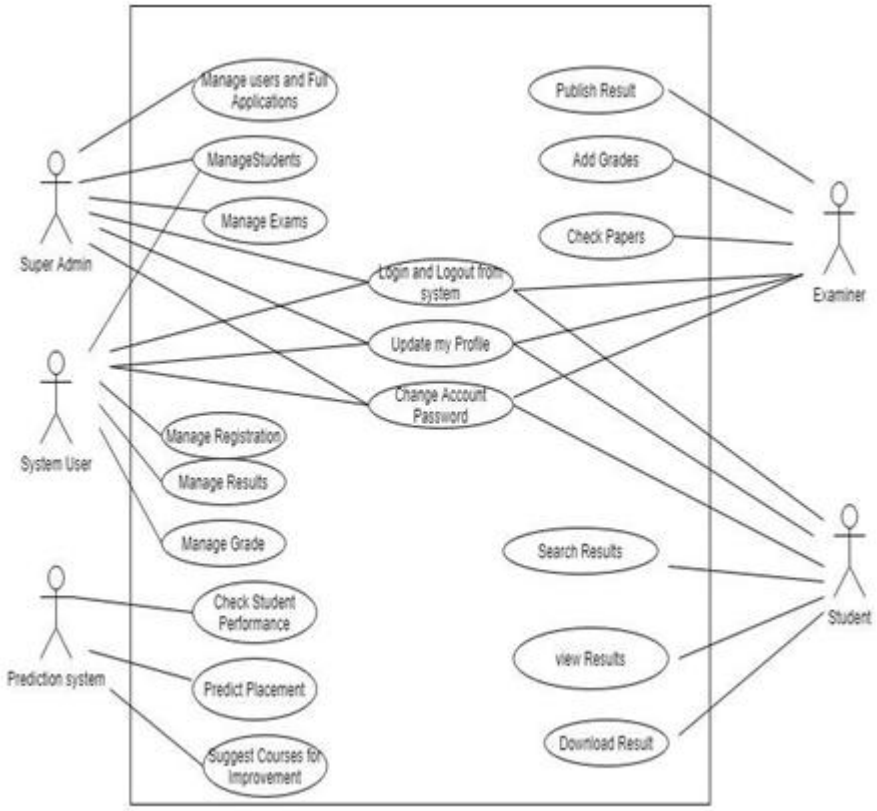
In order to predict a child's performance, grades, or danger of dropping out of school, neural networks are typically utilised in computer algorithms (ANN). Predicting students' behaviour is one of the closely connected subjects of interest in the field of education that has lately produced a lot of study. In fact, a tonne of research have been conducted on this subject and have been presented at conferences and in publications. Because of this, the major objective of this research is to provide an in-depth analysis of the many strategies and algorithms that have been put out and used in this field.

IV. PROJECT SCOPE

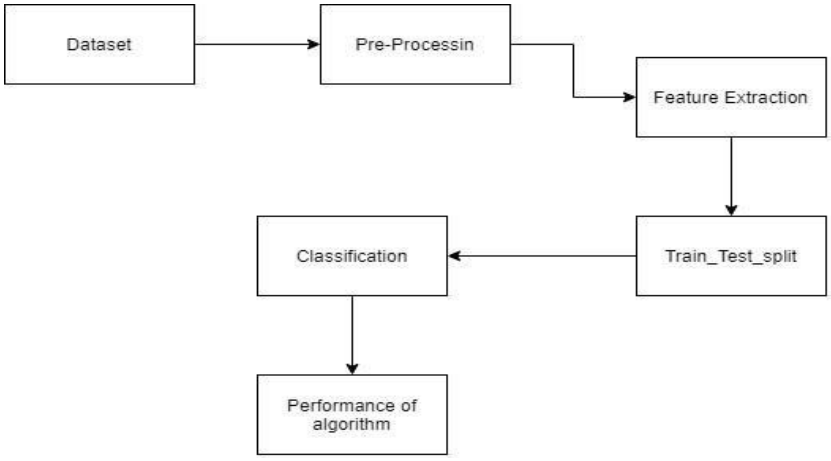
- To introduce a real-time system for monitoring student performance.
- To check student performance, conduct a variety of operations on the student record.
- To determine future prospects for students.
- To experience various effects quickly

V. USER CLASSES & CHARACTERSTICS

1. **Registration:** Initially, students must register themselves on the site (Registration).
2. **Upload Marks:** Students should upload their marks for the second phase in accordance with their academics.
3. **Prediction:** Students will receive their career prediction information after submitting their grades and other information.



SYSTEM ARCHITECTURE



The system's abstract perspective is shown in the figure above 2. 3 actors make up the system

A. Advantages

- By examining his hobbies and academic achievements, a student might receive counselling through which he will learn about the field in which he has potential.
- To comprehend the student advancement rate, student performance prediction is crucial.

- Global accessibility through the use of this technology alone. It is helpful in this epidemic condition since the instructor won't perform any physical analyses.
- Helpful for the instructor since it allows her to save time from having to evaluate each and every student.

B. Limitation

We may conclude that physical analysis will be superior to digital analysis in this regard because the latter merely makes predictions about students' academic achievement.

Application

- Both students and teachers can utilise Student Performance Prediction in a variety of ways.
- The student may receive advice for his upcoming actions from this person. For instance, if an engineering student uses this, he will receive company recommendations based on his performance and interests.
- The instructor may utilise this if she has a large number of students, which might result in time savings.

In this chapter, we explore into depth regarding R and how the modified K means method was implemented.

Implementation Information

The modified K-Means method is employed to assess the students' performance. They were divided into k groups from n items. The nearest mean is used to position the data. It implies that comparable types of data are grouped together, whilst other types are found in different groups. The study is based on a parameter with identical data. Market research, pattern recognition, and other fields frequently employ clustering analysis. The elbow approach is used to calculate modified K means for the cluster size. The number of clusters will be determined using the elbow approach because using a random integer for the cluster size could not produce the best outcome. The fundamental idea behind the elbow technique is to compute the sum of squares. The sum of squared values for the K value range is shown in a line-chart-like manner.

Depict the data (K may have any number of value). This line graph's valve at the elbow, which resembles an arm, corresponds to the correct value of K. The most important step is to select a K value that has a low sum of squared values. It is simple to do and produces the best outcomes. Python, R, SAS, and other programming languages are only a handful of the many that are frequently used for analytical reasons. After installing the required packages to construct the modified K-means algorithm, the elbow technique and different graphs between parameters may be plotted using the library of those packages. in the language R. While many factors are utilised to analyse student performance, the dataset also includes many other characteristics that have no bearing on the outcome. In this instance, a parameter like result is crucial for the analysis, but a parameter like gender, which can't be used to forecast results, is crucial for the study's goal of separating the genders and providing correct data. There are several options in R Studio, including a help option that

allows anyone to learn more about any library. As a result, R Studio offers a wide range of options, and determining cluster size using the elbow approach is helpful.

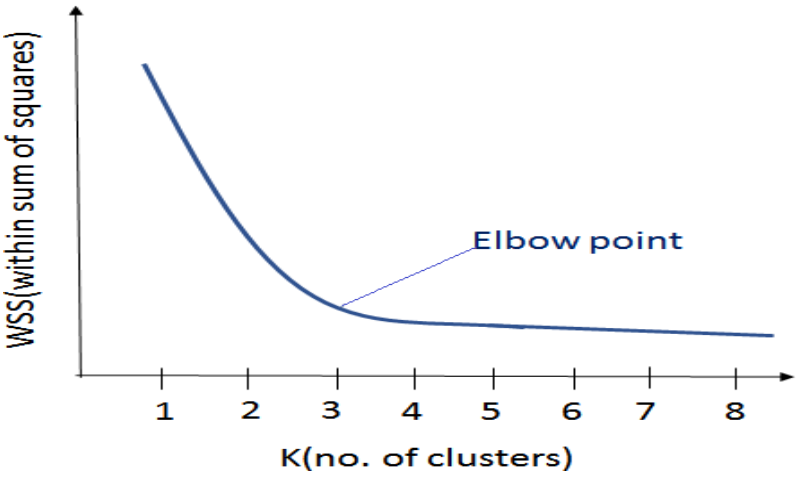


Figure 3: Elbow Method

The R set-up

For statistical computation, R programming is mostly utilised. Most representation in R is done through visuals. R may be downloaded for free. A space for analysis is present. It may be accessed under a licence that is open to everyone and is compatible with Windows, Linux, and Mac. R is software that can also do operations created in other programming languages, including C, C++, Python, and others. All loops, functions, and other programming concepts are supported by the extremely user-friendly, uncomplicated programming language R. In R, there is a capability for handling data as well as one for storing it. R has a function similar to a matrix and a vector. It contains several different data analysis tools. Ross and Robert named the programme R, and the core development team managed its administration. The most popular statistics language used worldwide is R.

Platform required running R studio

Unix and Unix like systems

Linux

Windows XP/7/8

R studio server.

Studio R

It's free to utilise the open source integrated development environment R Studio. It is utilised for R programming and statically computing. The studio also offers graphic representation capabilities. There are two versions of R studio: R studio Desktop and R studio Web. The software looks like a typical desktop application and runs locally as a regular application. The desktop version supports Windows, Linux, and Mac as operating systems.

R Studio server: A web browser is available, and it is used to access R Studio.

R studio's graphical user interface is built using the qt framework, which was created in both Java and C++. A user may examine a graph, data tables, R code, and the result all at once because to how thoughtfully organised the interface is. It has a number of capabilities that let users import various files, including csv, excel, etc. R Studio is divided into four quadrants, each of which having a distinct characteristic, as seen in the picture below. The console is in the second quadrant, while the script is in the first. In all running circumstances, the results are generated on the console. The variables are all shown in the third quadrant, named environment, which also shows how each variable interacts with its surroundings. There are additional possibilities, including packages, files, and viewers, in the graph's last quadrant. Each quadrant has a unique quality and importance. The user may adjust the size of each quadrant to suit their needs.

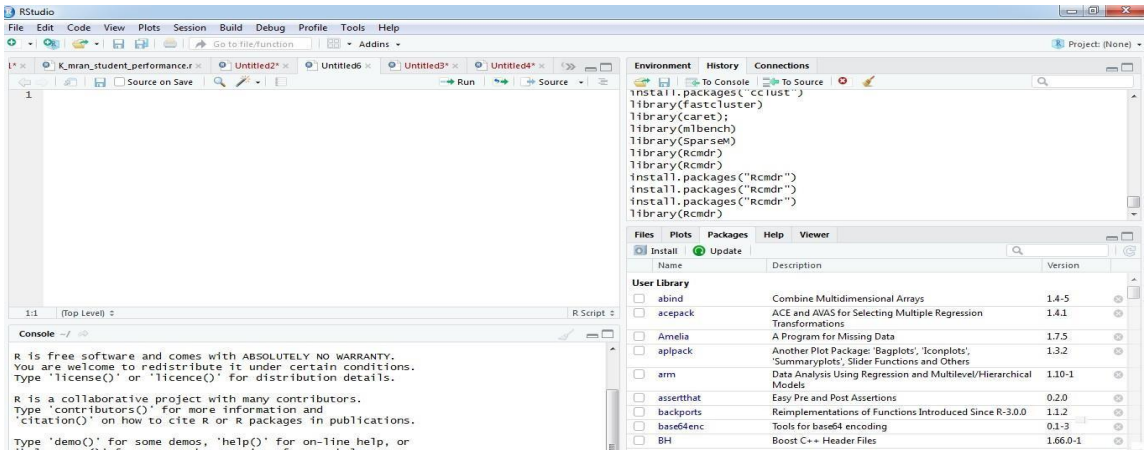


Figure 4: R studio

R Packages: As well as a R function and a number of R packages that contain it, R has a sizable number of packages. R installs a number of packages that are useful at a certain time throughout the installation process. There is a large collection of packages that may be added as needed, and more packages can be added at a later time. They are all stored in the "library" directory. It may be used to load the package that has previously been installed. A default package is loaded automatically when the console is launched. The library location for the package is:.libpaths ()

When we run library, a list of installed packages will be shown (). We may also add additional packages by using the install.packages ("Package name") command.

Caret, ggplot2, Rcmdr, and Stats are a few of the R library's most well-liked programmes. There are several ways to install the package in R. While some choices may be selected directly, some need programming. Before any package can be used, all of the other packages must be installed because they depend on one another. Use the help feature to examine the product's details if you need additional information about a certain bundle.

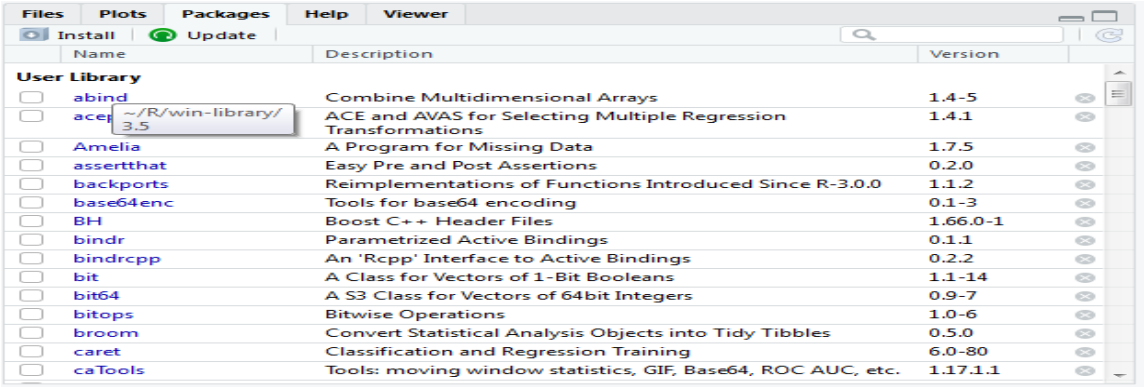


Figure 5: R studio package

The above illustration shows that there are many options in the fourth column of R Studio, including a package option that will provide assistance if needed with any packages as well as instructions on how to use them. This feature is extremely helpful because it will offer immediate assistance if we need any package-related information.

Students' data can be analyzed using a variety of parameters, but some of these parameters are extremely helpful, while others are not as crucial or, in other words, do not have an impact on the outcome. For instance, while Id and sgpa are crucial and must be used in the dataset, gender is not crucial and will not have an impact on the outcome. The sample dataset figure that was taken is shown below. Each student has a unique ID in this dataset, which includes 50 IDs and results such as HSC 10th, 12th, and sgpa. There are additional factors, such as uplifted hands. It refers to how many times pupils raised their hands to indicate a concern. The dataset contains the grades for each of the five subjects, along with the sgpa for each. This variable will be useful in analysing the performance of the pupils. As they are all a part of the dataset, some parameters are valuable and some are not.

Id	semester	gender	SectionID	StudentAl	Staylocati	Discussior	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 6: Dataset from students

The dataset, which has 50 ids and uses a variety of parameters, is depicted in the figure. A result is shown in relation to the dataset. The number of days that students are

present and absent is implied by the parameter student absence days. There is a stay location that indicates the student's present residence. There are therefore a total of 50 IDs, and varied outcomes and sgpa are related to each ID.

Modified K-means algorithm in action

K stands for algorithm, which is executed over the R studio using the aforementioned dataset. A library is installed in the R studio for showing the graph.

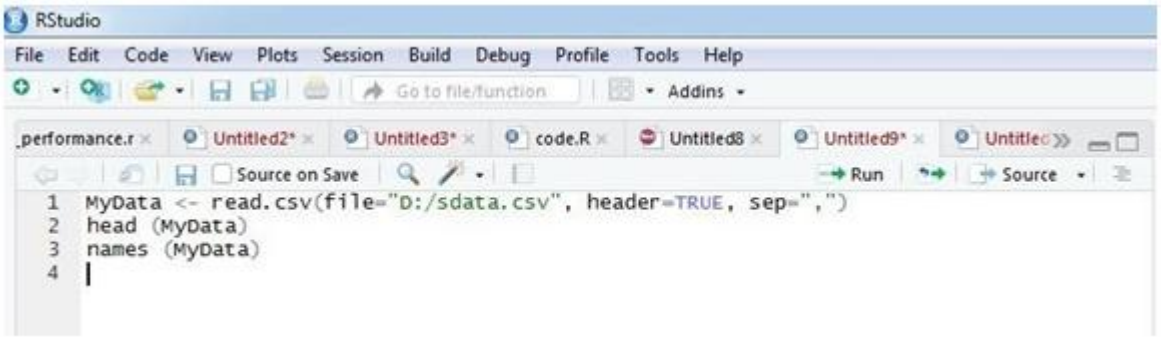


Figure 7: Import of data

The graphic above shows that in order to implement R studio, data must first be read from the location where it is kept and then placed in a data frame called "MyData". This data frame then serves as the basis for all R console operations involving the dataset.

Cluster Size by Elbow Method

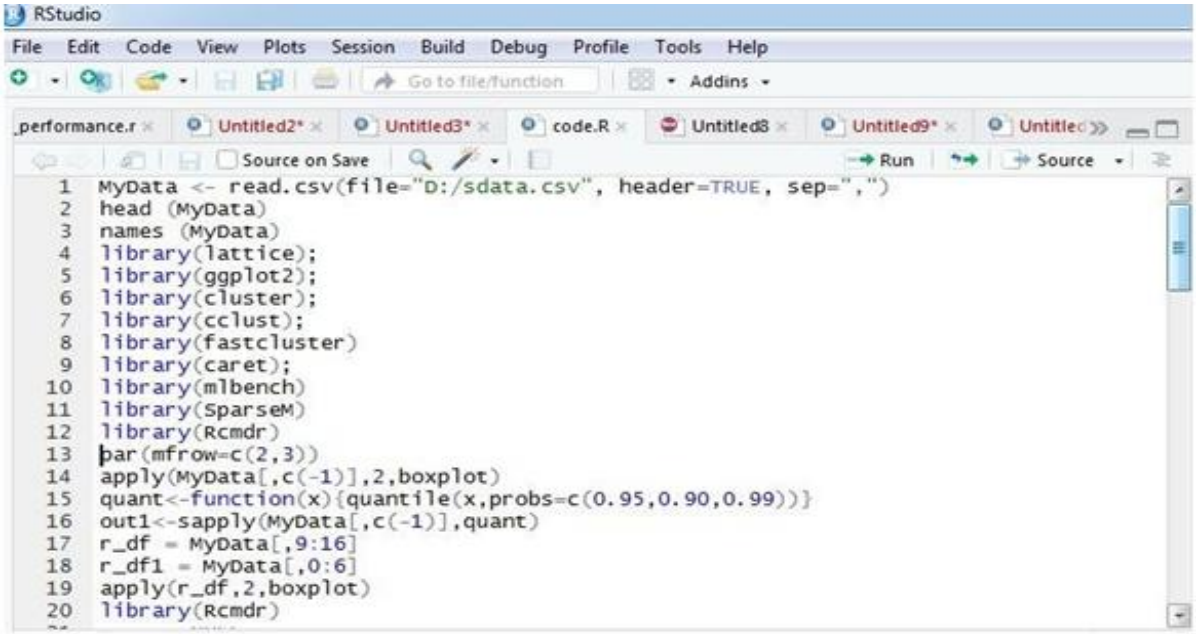


Figure 8: Cluster size,

The above illustration shows how several packages must be installed before the modified K-means clustering can be used. Next, the elbow approach requires the entire

package library. The ggplot2, Caret, and Rcmdr libraries are crucial in this case since they are used to plot the graph, load R Commander, and utilise Rcmdr. Also, the stats library has a line chart, the elbow technique is used to create it, and the elbow point is identified by glancing at the arm. The elbow method is also used to implement the cluster size. Data preprocessing, normalisation, and raw data are all present, and the operation then determines which data is valuable. The data are depicted in a boxplot. The sum of squares and a graph for the elbow approach are available for the cluster size via this method.

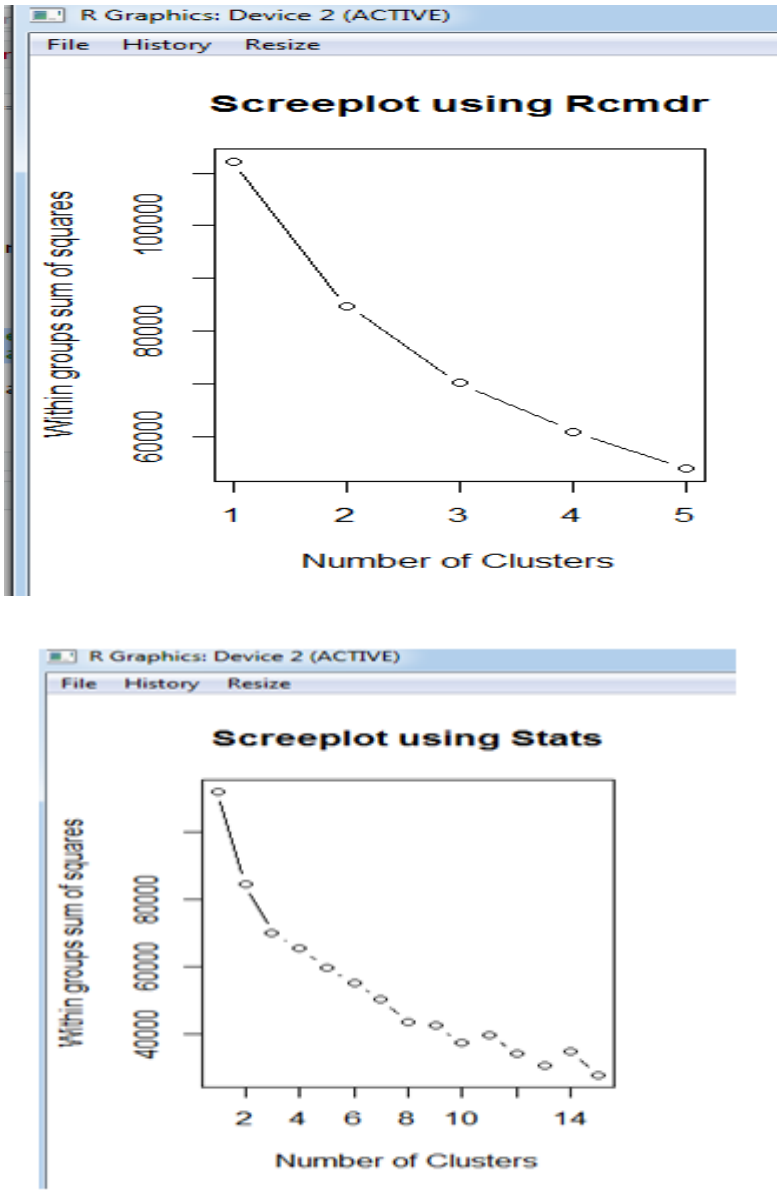


Figure 9: Elbow methods Using Rcmdr and Stats

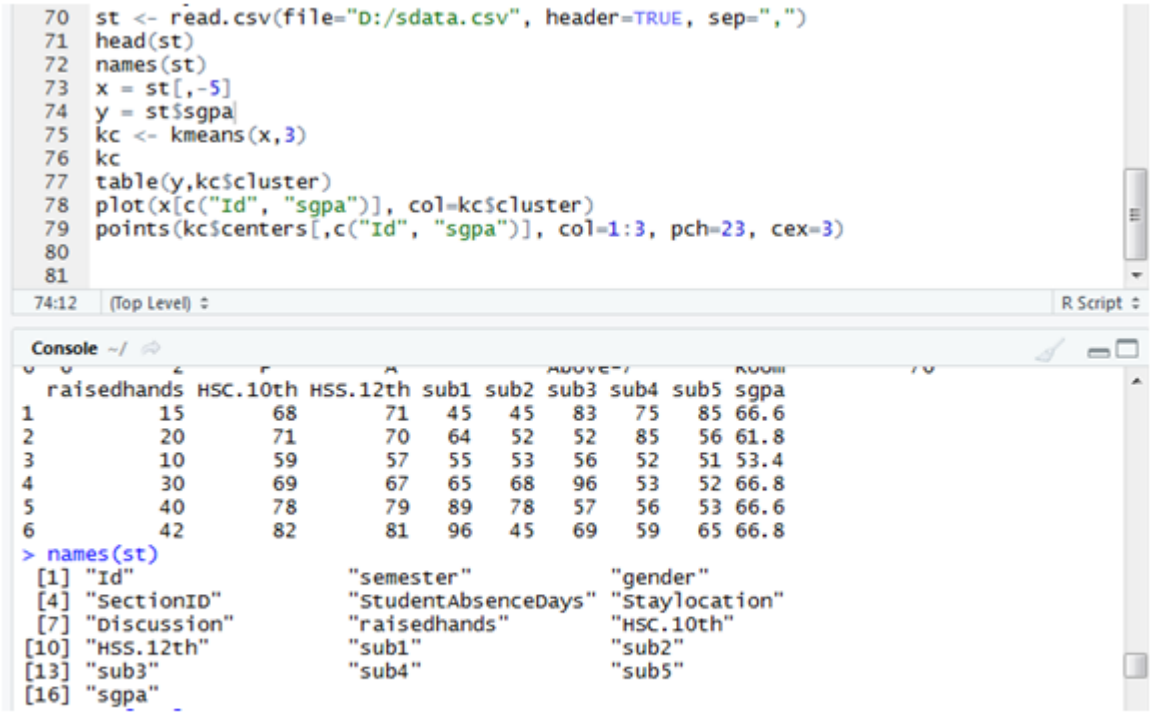


Figure 10: Graph Plotting

In order to plot a graph, the library must first be opened before any data is imported, the algorithm is then applied, the graph's points are chosen, and the parameter by which the axis is presented is taken. This is demonstrated in the above Figure. Below is the graph between Id and SGP. It may be altered dependent on the circumstance by using k=3. Though Rcmdr and Stats are utilised here, the elbow approach, and it is evident from the line cart that there is an arm near to K=3, indicating that the cluster size is 3. Because the cluster size is 3, the elbow approach will get the best result in this situation. The x and y axes of the line chart made using the elbow method, respectively, display the cluster size and the sum of squared errors within the data. Moreover, additional packages are installed, and the elbow method library is imported as required. The elbow approach is essential for clustering.

Console of R studio

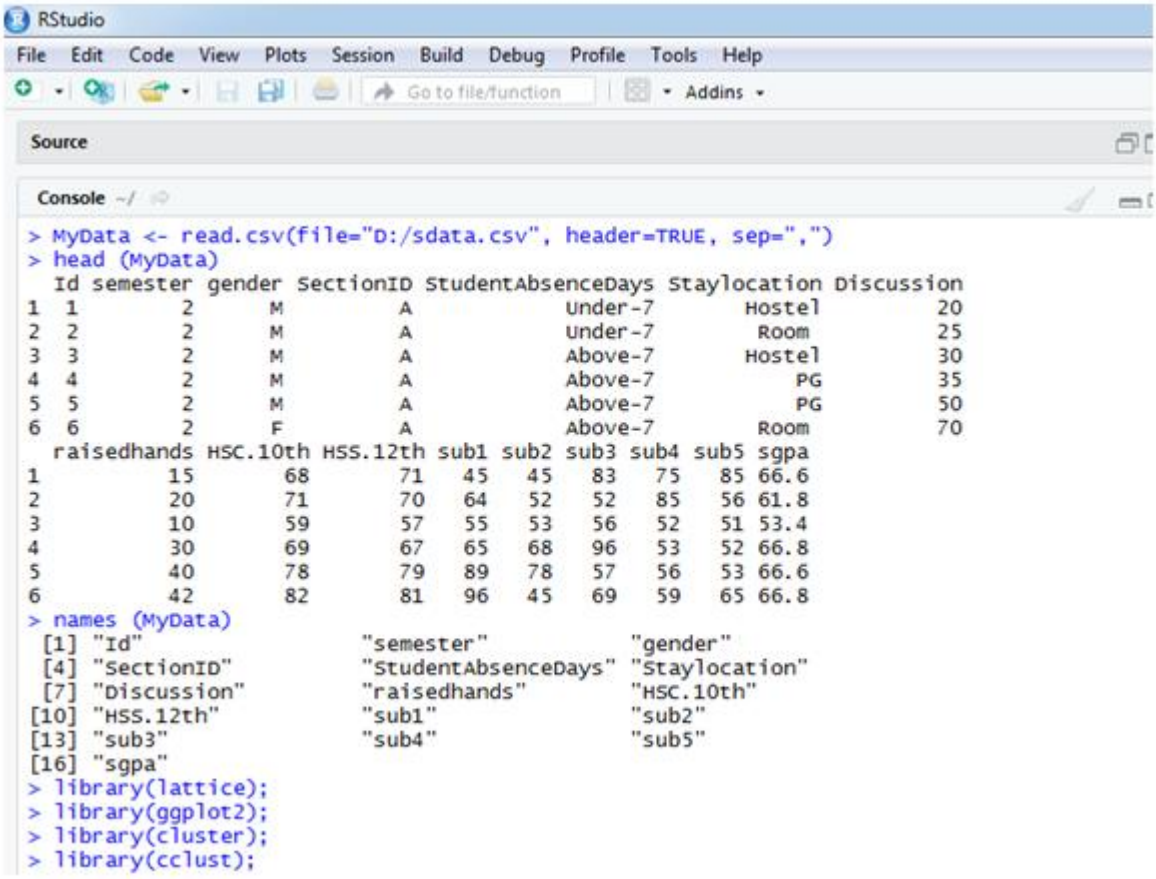
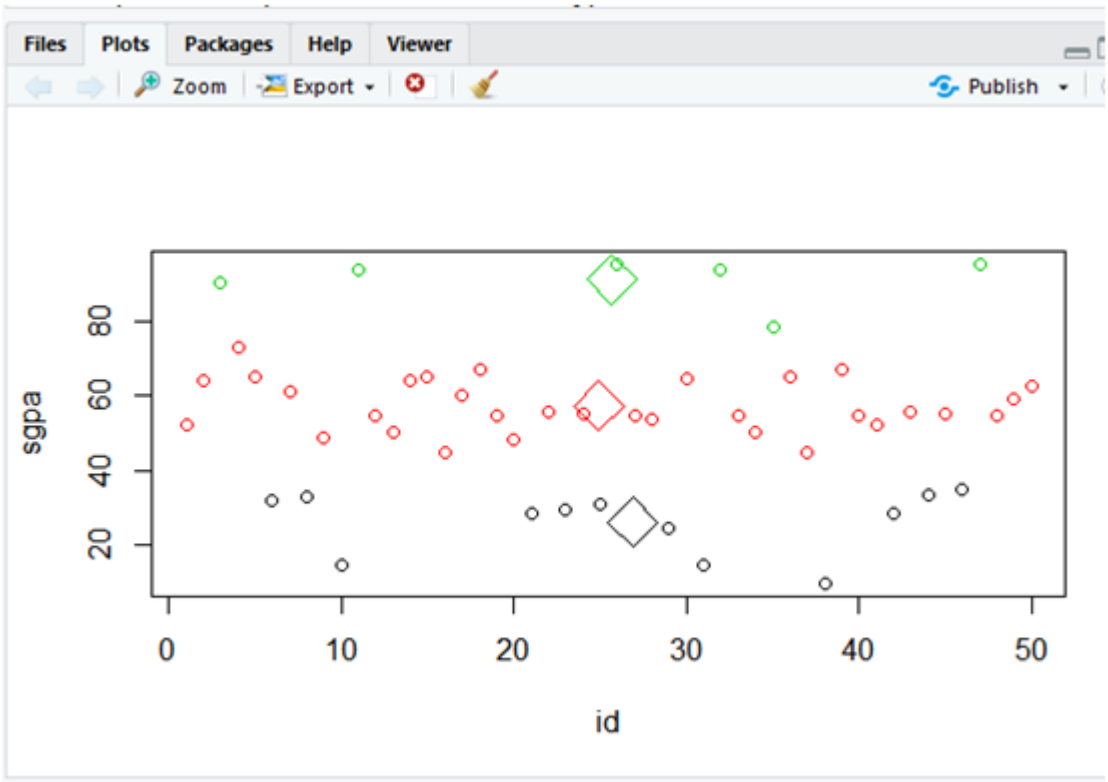


Figure 11: R's Console

Figure depicts the second quadrant of the R studio. When any script is run in R, the results are displayed in the console. In this case, when K means is run and data is imported, the results are displayed in the console. This imports the data and creates the cluster and vector, both of which are crucial. The R studio is divided into four quadrants, each with a unique characteristic. Now that the algorithm has been applied, a cluster of sizes 12, 6, and 32 has been created, and the graph will be shown in the R studio's fourth quadrant. The elbow method determines the cluster size, and all of the parameter names are taken by the names function. Inside the cluster, the sum of squares and the clustering vector are present. The R console is significant since this quadrant alone contains all of the active operations.
Graph between Id and Sgpa When K=3



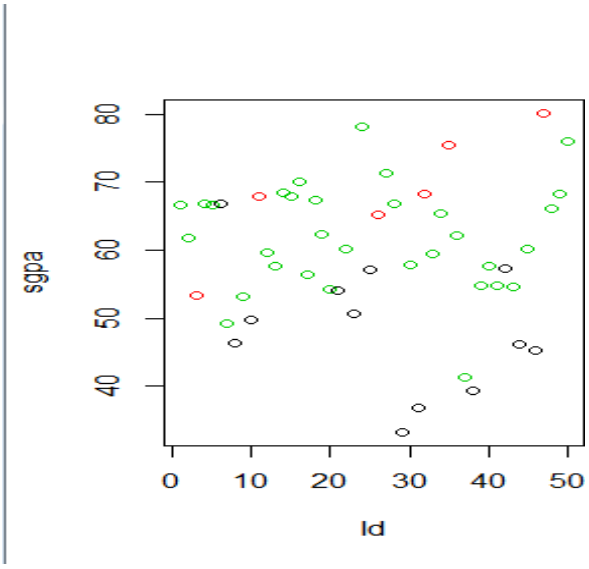


Figure 12: Graph between Id and sgpa

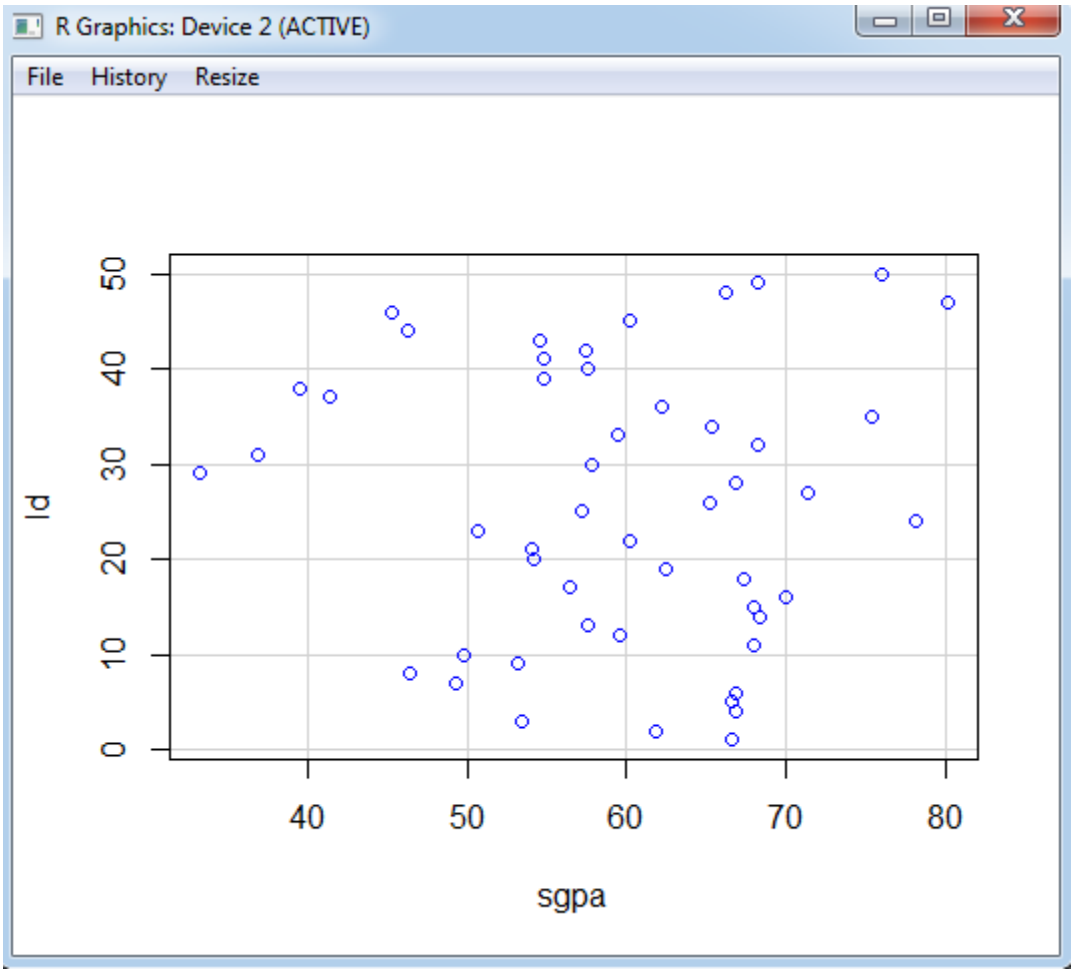


Figure 13: Scatter graph of sgpa

Chapter Summary

The elbow approach is utilised to calculate the cluster size, and the modified k-means algorithm is employed. The R studio is used, in which packages are installed, the library is taken, and data is imported. There is a graph between the variables id and sgpa. By taking the cluster size at the elbow point for the operation, a graph is then shown.

VI. CONCLUSION

Recent research indicates that a student's earlier accomplishments have a significant impact on their academic progress. Our research shows that a student's achievement is significantly influenced by their prior performance. In addition, we demonstrated that neural network performance scales with dataset size. From its early days, machine learning has made considerable strides and has the potential to be a useful tool in academics. Any academic institution may incorporate future applications like this one, along with any improvements made to them.

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.
[3] 9, No. 4, pp. 136-140, 2011.

