CHAPTER 1

INTRODUCTION

In data mining and data analytics, tools and techniques once confined to research laboratories are being adopted by forward-looking industries to generate business intelligence for improving decision making. Higher education institutions are beginning to use analytics for improving the services they provide and for increasing student grades and retention.

With analytics and data mining experiments in education starting to proliferate, sorting out fact from fiction and identifying research possibilities and practical applications are not easy. As more of our commerce, entertainment, communication, and learning are occurring over the Web, the amount of data online activities generate is skyrocketing. Commercial entities have led the way in developing techniques for harvesting insights from this mass of data for use in identifying likely consumers of their products, in refining their products to better fit consumer needs, and in tailoring their marketing and user experiences to the preferences of the individual. More recently, researchers and developers of online learning systems have begun to explore analogous techniques for gaining insights from learners' activities online.

Using data for making decisions is not new; companies use complex computations on customer data for business intelligence or analytics. Business intelligence techniques can discern historical patterns and trends from data and can create models that predict future trends and patterns. Analytics, broadly defined, comprises applied techniques from computer science, mathematics, and statistics for extracting usable information from very large datasets.

1.1 PROJECT OVERVIEW

Our project is divided into 3 phases namely:

1 Data warehouse

2 Data cleaning

3 Data mining algorithm implementation.

DATA WAREHOUSE:

In general, building any data warehouse requires the following steps:

1. Extracting data from data sources into a staging area
2. Transforming the data
3. Loading the data into a dimensional database
4. Building pre-calculated summary values to speed up report generation
5. Building a front end reporting tool

Data warehouse was created by us by collecting data through a survey which included student academic details. The database was then divided into different subsets of databases according to the respective information for example, campus wise detail, student wise detail etc. The correlation between the databases was studied and was connected to each other through a foreign key. Data cube was made using the software SQL Server management studio and business intelligence studio. The dimension model consists of facts and dimension table. The fact table consist of foreign key to each dimension table, as well as measures. The measures are factual representation of how well the business is doing. Dimensions on the other hand are what your business users expect in the report. Like any other stage dimensional modelling is also very important and can be achieved successfully by discussing the business requirements with the users over and over again.

The major challenge in building a consistent data warehouse was the accumulation of data from different sources and converting it into a standard readable format.

DATA CLEANING:

Every dataset contains some errors, and every analyst experiences a rite of passage in wasting days drawing wrong conclusions because the errors have not been first rooted out. Up to half of the time needed for analysis is typically spent in "cleaning" the data. This time is also,

typically, underestimated. Often, once a clean dataset is achieved, the analysis itself is quite straightforward.

Unless the dataset is small (i.e., less than 100 cases and 10 variables), cleaning is done in several stages. To begin with, the key variables are examined and corrected. In a large dataset, variables that become of interest as analysis proceeds need to be cleaned as you go along.

- Once processed and organized, the data may be incomplete, contain duplicates, or contain errors.
- The need for data cleaning will arise from problems in the way that data is entered and stored.
- Data cleaning is the process of preventing and correcting these errors. Common tasks include record matching, reduplication, and column segmentation.
- Such data problems can also be identified through a variety of analytical techniques.
- For example, with institutional information, the totals for particular variables may be compared against separately published numbers believed to be reliable.
- Unusual amounts above or below pre-determined thresholds may also be reviewed. There are several types of data cleaning that depend on the type of data.
- Quantitative data methods for outlier detection can be used to get rid of likely incorrectly entered data.

DATA MINING ALGORITHM IMPLEMENTATON:

The following methods have been implanted till now in our project. However a detailed explanation of each of the methods is given in the subsequent sections:

- K- means
- Dendogram Method
- Density based clustering
- Fuzzy C Means
- Multi Gaussian method
- Linear Regression Method
- Multiple Regression

1.2 HARDWARE SPECIFICATION

- 512 MB ram
- Dual core processor
- Internal storage- Minimum 40 MB
- Computer system

1.3 SOFTWARE SPECIFICATION

- SQL Management server 2008
- Business intelligence studio
- R Studio
- Microsoft Excel
- Shinny Tool (R)

CHAPTER 2

LITERATURE REVIEW

Paper1: Importance of Data Mining in Higher Education System

By: Bhise R.B., Thorat S.S. and Supekar A.K., Pune, India.

Education is an essential element for the betterment and progress of a country. Mining in educational environment is called Educational Data mining, concern with developing new methods to discover knowledge from educational database in order to analyze student's trends and behaviors towards education.

K-Means Clustering

Data mining techniques are used to operate on large volume of data to discover hidden pattern and relationship helpful in decision making. Cluster analysis used to segment a large set of data into subsets. This study makes use of cluster analysis to segment students into groups according to their characteristics.

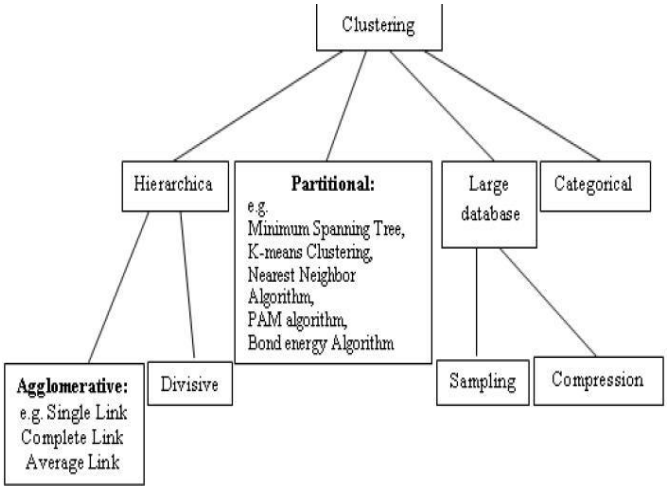


Fig 2.1 Clustering types

Algorithm: Basic K-means Algorithm:

1. Select K points as the initial centroids.
2. Repeat.
3. From K- cluster by assigning all points to the closest centroids.
4. Recomputed the centroid of each cluster.
5. Until the centroids don't change.

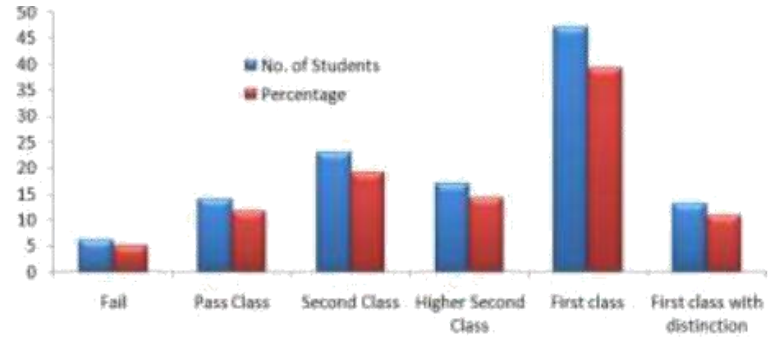


Fig 2.2 Graphical Plot

Results

The students were grouped regarding their final grades in three ways

1. Assign possible labels that are the same as number of possible grades

2. Group the students in three classes “High”, “Medium”, “Low”.

3. Categorized the students with one of two class labels “Passed” for marks greater than or equal to 40 and “Failed” for marks less than 40.

The understanding: In this paper the author made the use of data mining process in a student’s database using K-means clustering algorithm to predict students result. The information generated after the implementation of data mining technique will be helpful for instructors and students.

Paper 2: Relevance of Data Mining Techniques in Education Sector

INTRODUCTION

Data mining, also called Knowledge Discovery in Databases (KDD), is the field of discovering novel and potentially useful information from large amounts of data. The process of converting raw data into useful information consists of a series of steps from data pre-processing to post processing of data mining results.

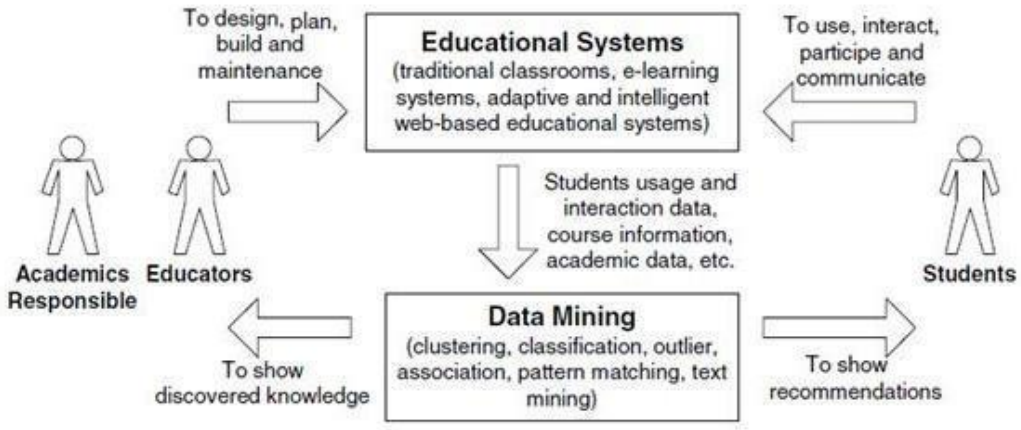


Fig 2.3 EDM

Educational data mining (EDM)

EDM is the process of transforming raw data compiled by education systems in useful information that could be used to take informed decisions and answer research questions Objectives of EDM:

1) Pedagogic objectives – To help the students to improve in academics, designing the content of the course in a better way etc

2) Management objectives - To optimize the organization and maintenance of educational infrastructures, areas of interest, more requested courses

3) Commercial objectives - It allows to make market segmentation and facilitates students recruitment in schools, high schools, colleges and universities

VARIOUS APPROACHES IN EDM:

Data mining tasks are divided into two major categories.

- Predictive task
- Descriptive task.

CONCLUSION

The various data mining algorithms which are applied in business can also be effectively used in the field of education for the betterment of student performance and the institution.

Paper3: Mining Educational Data to Improve Students' Performance: A Case Study

Mohammed M. Abu Tair and Alaa M. El-Halees, Palestine

In this paper through a case study they try to extract useful knowledge from graduate students data collected from the college of Science and Technology – Khanyounis. The data include fifteen years period [1993-2007]. After preprocessing the data, they applied data mining techniques to discover association, classification, clustering and outlier detection rules. In each of these four tasks, they present the extracted knowledge and describe its importance in educational domain.

As part of the data preparation and preprocessing of the data set and to get better input data for data mining techniques, preprocessing for the collected data was done before loading the data set to the data mining software, so that irrelevant attributes should be removed. The attributes such as the Student_Name or Student_ID, etc. are not selected to be part of the mining process; this is because they do not provide any knowledge for the data set processing and they present personal information of the students, also they have very large variances or duplicates information which make them irrelevant for data mining. After applying the preprocessing and preparation methods, they tried to analyze the data visually and figure out

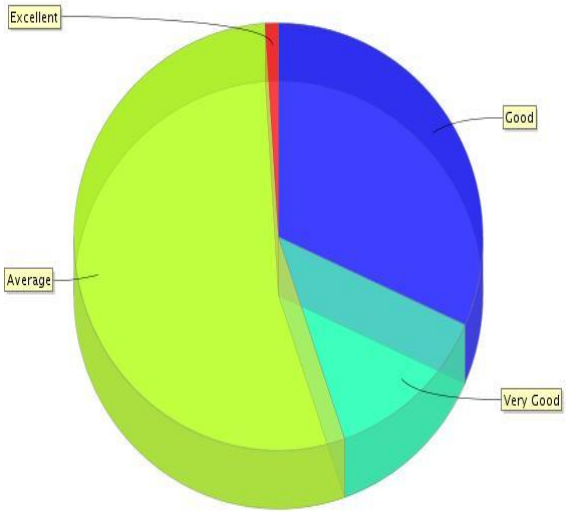


Fig 2.4 Pie chart Representation

the distribution of values, specifically the grade of students. The Figure depicts the distribution of graduate students.

Application Process:

- **Association Rules:** Mining association rules searches for interesting relationships among items in a given data set. It allows finding rules of the form 'If' antecedent 'then' (likely) consequent where antecedent and consequent are item sets.
- **Classification**: In this paper, the classification approaches are used to predict the Grade of the graduate student and there are four grades (Excellent, Very good, Good, and Average) and how other attributes affect them. Two classification methods are used which are Rule Induction and Naïve Bayesian classifier. The benefit of these two methods is that it can predict low grades on time.
- **Outlier Detection:** Outlier detection discovers data points that are significantly different than the rest of the data. In educational data mining outlier analysis can be used to detect students with learning problems
    - **Distance-based Approach:** It Identifies the number of outliers in the given data set based on the distance to their k nearest neighbors, and the result of applying this method is to flag the records either to be outlier or not, with true or false.
    - **Density-Based Approach:** It Computes local densities of particular regions and declare instances in low density regions as potential outliers.

CONCLUSION

The two outlier methods are used which are Distance-based Approach and Density-Based Approach and each one of these tasks can be used to improve the performance of graduate student.

Paper 4: Improving the Student's Performance Using Educational Data Mining

K.Shanmuga Priya A.V.Senthil Kumar

In this paper, the data classification and decision tree is used to help improve the student's performance in a better way. This method helps to identify the students who need special advising or counseling by the teacher which gives high quality of education.

A Decision tree is a classification scheme which generates a tree and set rules, representing the model of different classes, from a given data set. A decision tree can be used to clarify and find an answer to a complex problem.

For this research the data set is obtained from, Hindustan College of Arts and Science, Coimbatore. The attributes considered are Overall Semester Marks, Class Test, Seminar Performance, Communication Skill, Paper Presentations, Attendance, Practical Lab, End semester Marks

The data table which consist of more than several classes are known as heterogeneous or impurity of table. There are several ways to measure impurity of table, the best method is entropy, gini index and classification error.

Entropy is used to calculate the quantitative impurity. Entropy of pure table becomes zero when the probability becomes one and it reaches maximum values when all classes in the dataset have equal probability.

Information gain is used to determine the best attribute for the particular node in the tree. Selecting the new attribute and partitioning the given values will be repeated for each non terminal node. If any attribute has been incorporated higher in the tree, that attribute will be excluded.

ID3 Algorithm

Step 1: Create a tree with root node.

Step 2: Return the single tree root node with label +, if all the given values are positive.

Step 3: Return the single tree root node with label -, if all the given values are negative.

Step 4: Return the single tree root node with label = most common values of target attributes in the given value. It can be performed when predicting attribute is empty.

Step 5: else begin

(i) A specifies best attribute in the given value.

(ii) In decision tree the root for an attribute is A

(iii) For A, each possible values Vi as,

(a) If A = Vi add corresponding branch below root.

(b) Let given value Vi which is subset of the given value Vi for A.

(c) If the given value Vi is empty

(i) Add a new leaf to the branch node which is equal to most common target value in the given value.

(ii) Add the sub tree ID3 to this new branch node (values given Vi, Target Attribute, Attribute).

Step 6: End the process.

Step 7: Return the root node.

The root node can be deducted by calculating gain information from the given student data set. Here by we have to calculate the entropy value first.

$$H(S) = -\sum_{x \in X} p(x) \log_2 p(x)$$

$p(x)$ - The proportion of the number of elements in class x to the number of elements in set S

$$IG(A, S) = H(S) - \sum_{t \in T} p(t) H(t)$$

The attribute with maximum gain value is selected as the root node of the decision tree. These calculations will be continued until all the data classification has been done or else till all the given attributes get over.

CONCLUSION

The concept of Decision Tree which integrates with the classification technique helps to achieve the goal by extracting the discovery of knowledge from the end semester mark. The inclusion of

extracurricular activities makes to gain more knowledge along with the End semester Mark. This technique is one of the ways to improve performance of the students in education.

Paper 5: Use of Data Mining in Education Sector

This paper discusses the problems faced by higher education institutes.

One of the biggest challenges that higher education faces today is predicting the paths of students. Institutions would like to know, which students will enrol in which course, and which students will need more assistance in particular subject. Data Mining will help the institution to take decision more accurately. This paper discusses about data mining, their different phase's, advantages and also the classification of data using weak data mining tool. This paper makes use of the J48 algorithm to predict the result of the student

Data mining tools and algorithms highlighted:

• Machine Learning	• Artificial Intelligence
• Computer science	• Emulating human intelligence
• Heuristics	• Neural Networks
• Induction algorithms	• Biological models, psychology and engineering

Phases of Data Mining

Data mining is an iterative process that typically involves the following phases:

- Problem definition
- Data exploration
- Data preparation
- Modelling Evaluation

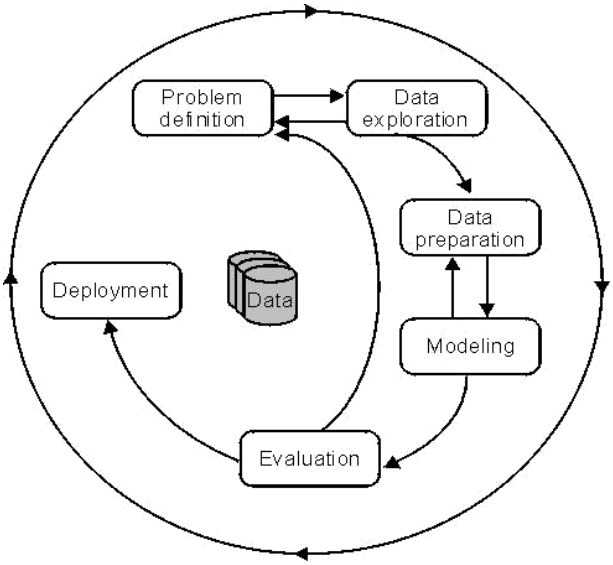


Fig 2.5: The process

Tools of Data Collection & Analysis

Various tools are needed for that project some for analyzing data, some for designing, implementation and some developing software tool these are:

- MYSQL DATBASE
- EXCEL
- MS ACCESS
- SPSS
- MATLAB TOOL
- WEKA DATA MINING TOOL
- TANGARA DATA MINING TOOL
- WEB MINER
- V.B 6.0

Advantage of Data mining in Academics:

Data mining gives the answers of questions like:

- Who is the weak student?
- Who are the students taking most credit hours?
- The interesting subject of the students?

- What type of course can we offer to attract more students?
- How can help weak student?
- How improve the result of college?
- Predicting the result of students?

CONCLUSION

The Proposed system shows data graphically according to the need of the organization, which helps them to take important decisions. For future work the author wishes to use clustering, as this will help in obtaining an in depth analysis of the various domains associated with the students.

Paper 6: Data Mining Techniques

Data mining is a process of extraction of useful information and patterns from huge data. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data /pattern analysis.

Knowledge Discovery Process

Knowledge discovery is a process that extracts implicit, potentially useful or previously unknown information from the data.

1. Selection
2. Pre-processing
3. Transformation
4. Data mining
5. Interpretation and evaluation

Data Mining Techniques

There are several major data mining techniques have been developed and used in data mining projects recently including:

1. Association

Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship of a particular item on other items in the same transaction.

2. Classification

Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. For Example, Teachers classify students' grades as A, B, C, D, or F. E.g. regression, decision trees, etc.

3. Clustering

Clustering is "the process of organizing objects into groups whose members are similar in some way". A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.

4. Prediction

The prediction as it name implied is one of a data mining techniques that discovers relationship between independent variables and relationship between dependent and independent variables. In data mining independent variables are attributes already known and response variables are what we want to predict unfortunately, many real-world problems are not simply prediction.

5. Sequential Patterns

Sequential patterns analysis in one of data mining technique that seeks to discover similar patterns in data transaction over a business period. The uncover patterns are used for further business analysis to recognize relationships among data.

CONCLUSION

Data mining has wide application field almost in every industry where the data is generated that's why data mining is considered one of the most important frontiers in database and information systems and one of the most promising interdisciplinary developments in Information Technology also.

CHAPTER 3
REQUIREMENTS AND ANALYSIS

The Analysis and Design phases of our project were the primary stages in which the primary planning and overview was performed.

Analysis:

We envisioned the data which we will finally need and primarily enlisted the attributes which we could potentially acquire. We each came up with a few and then started planning and seeing the overview. We initially started by building a data ware-house subsequently bringing together the diverse fields into a single data.

Following this we made a data cube which comprised of all the fields in one location hence by using the software and database querying in SQL Management server 2008 and Business intelligence studio. By collaborating the data in Business intelligence studio and connecting the data base using commands and queries data cube was framed through which multiple graphs and pivot tables were developed to make a comparative study.

After this we formulated a potential analysis plan in which we tried to pair the attributed together and discern what potential information and inferences could be predicted and obtained from the data. Additionally, we read some research papers to gain an idea about the various algorithms so that we come to know what those papers have contributed and give us further ideas about which algorithms to use for our analysis. A glimpse of the same has been illustrated in the future scope of the report.

Furthermore we did an extensive research about the various soft wares available for data mining and weighed them against each other. We were looking for a good balance between presentation and ease of implementation along with comprehensive outputs. After the discussion with the mentor the R Language was finalised for the processing.

The following steps were performed as a primary focus during this stage:

- A thorough description of the data mining activity, its goals, and the target dates for the deployment of the data mining activity.
- A thorough description of the technology that will be used.

- A thorough description of the data sources that being or will be used.
- An assessment of the efficacy of the data mining activity in providing accurate information consistent with and valuable to the stated goals and plans for the use or development of the process.
- Ensuring that only accurate and complete information is collected, reviewed, gathered, analysed, or used, and guard against any harmful consequences of potential inaccuracies.

CHAPTER 4

PREPROCESSING

Data pre-processing is an important step in the data mining process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values (e.g., Income: −100), impossible data combinations (e.g., Sex: Male, Pregnant: Yes), missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis.

Data Cleaning

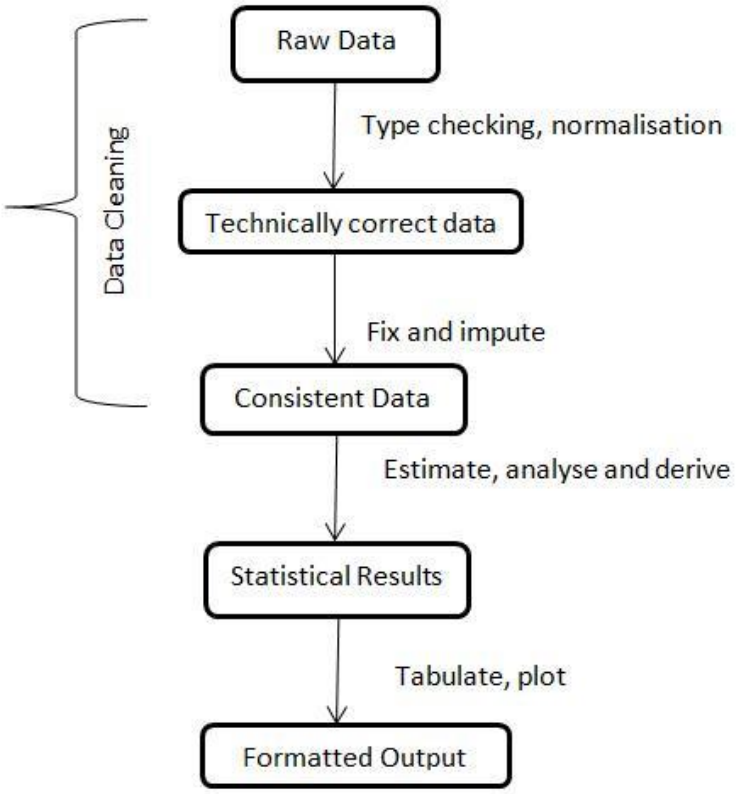


Figure4.1 The process of cleaning

The process involved the following implementation:

From raw data to technically correct data

- Raw data is collected from the resources and then processed such that for each unit, a text variable should be stored as text, a numeric variable as a number, and so on, and

all this in a format that is consistent across the data set (type matching and normalisation of data is performed)

From technically correct data to consistent data

- Consistent data are technically correct data that are fit for statistical analysis. They are data in which missing values, special values, (obvious) errors and outliers are either removed, corrected or imputed. The data are consistent with constraints based on real-world knowledge about the subject that the data describe. Consistency can be understood to include in-record consistency, meaning that no contradictory information is stored in a single record, and cross-record consistency, meaning that statistical summaries of different variables do not conflict with each other.

  The process is carried out in three steps:

  - *Detection* of an inconsistency
  - *Selection* of the field or fields causing the inconsistency
  - *Correction* of the fields that are deemed erroneous by the selection method

From consistent data to Statistical Results

- To attain statistical results from the given consistent data we use the various clustering algorithms such as:

  K-means, Fuzzy C method, Density Based Method, Mixed Gaussian Method**.**

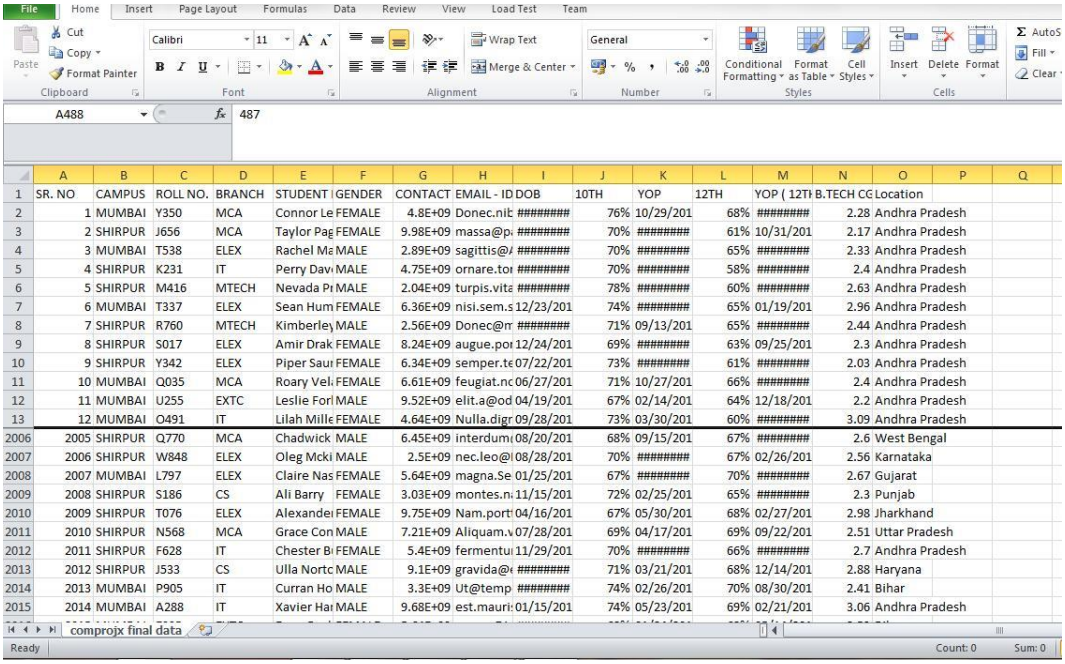
From Statistical Results to Formatted Output

- In order to analyse the data and understand the outputs we need a formatted output with help of which we derive sense and inferences from it. This is usually carried out using the help of graphs, plots, charts like pie charts, bar graphs, tabular formats.

CHAPTER 5

DATA WAREHOUSE IMPLEMENTATION

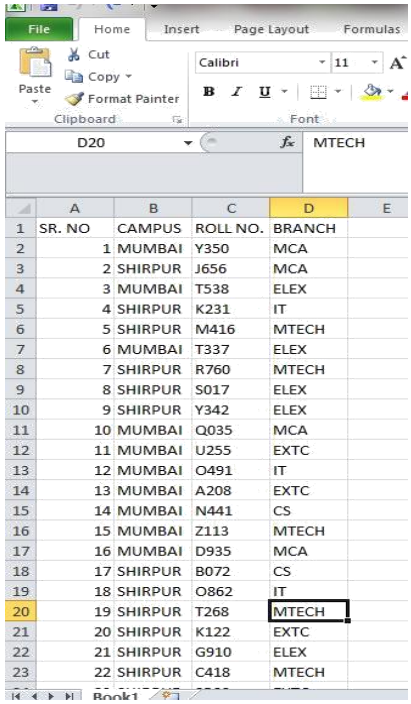
The project's backbone is the data ware-house created. The structure and architecture of the data mart is as follows:



SR. NO	CAMPUS	ROLL NO.	BRANCH	STUDENT	GENDER	CONTACT	EMAIL - ID	DOB	10TH	YOP	12TH	YOP (12TH	B.TECH CG	Location
1	MUMBAI	Y350	MCA	Connor Le	FEMALE	4.8E+09	Donec.nib	########	76%	10/29/201	68%	#########	2.28	Andhra Pradesh
2	SHIRPUR	J656	MCA	Taylor Pag	FEMALE	9.98E+09	massa@pa	########	70%	#########	61%	10/31/201	2.17	Andhra Pradesh
3	MUMBAI	T538	ELEX	Rachel Ma	MALE	2.89E+09	sagittis@/	########	70%	#########	65%	#########	2.33	Andhra Pradesh
4	SHIRPUR	K231	IT	Perry Dav	MALE	4.75E+09	ornare.tor	########	70%	#########	58%	#########	2.4	Andhra Pradesh
5	SHIRPUR	M416	MTECH	Nevada Pr	MALE	2.04E+09	turpis.vita	########	78%	#########	60%	#########	2.63	Andhra Pradesh
6	MUMBAI	T337	ELEX	Sean Hum	FEMALE	6.36E+09	nisi.sem.s	12/23/201	74%	#########	65%	01/19/201	2.96	Andhra Pradesh
7	SHIRPUR	R760	MTECH	Kimberley	MALE	2.56E+09	Donec@m	########	71%	09/13/201	65%	#########	2.44	Andhra Pradesh
8	SHIRPUR	S017	ELEX	Amir Drak	FEMALE	8.24E+09	augue.por	12/24/201	69%	#########	63%	09/25/201	2.3	Andhra Pradesh
9	SHIRPUR	Y342	ELEX	Piper Saut	FEMALE	6.34E+09	semper.te	07/22/201	73%	#########	61%	#########	2.03	Andhra Pradesh
10	MUMBAI	Q035	MCA	Roary Vel	FEMALE	6.61E+09	feugiat.nc	06/27/201	71%	10/27/201	66%	#########	2.4	Andhra Pradesh
11	MUMBAI	U255	EXTC	Leslie For	MALE	9.52E+09	elit.a@od	04/19/201	67%	02/14/201	64%	12/18/201	2.2	Andhra Pradesh
12	MUMBAI	O491	IT	Lilah Mille	FEMALE	4.64E+09	Nulla.digr	09/28/201	73%	03/30/201	60%	#########	3.09	Andhra Pradesh
2005	SHIRPUR	Q770	MCA	Chadwick	MALE	6.45E+09	interdum(	08/20/201	68%	09/15/201	67%	#########	2.6	West Bengal
2006	SHIRPUR	W848	ELEX	Oleg Mcki	MALE	2.5E+09	nec.leo@i	08/28/201	70%	#########	67%	02/26/201	2.56	Karnataka
2007	MUMBAI	L797	ELEX	Claire Nas	FEMALE	5.64E+09	magna.Se	01/25/201	67%	#########	70%	#########	2.67	Gujarat
2008	SHIRPUR	S186	CS	Ali Barry	FEMALE	3.03E+09	montes.n	11/15/201	72%	02/25/201	65%	#########	2.3	Punjab
2009	SHIRPUR	T076	ELEX	Alexander	FEMALE	9.75E+09	Nam.port	04/16/201	67%	05/30/201	68%	02/27/201	2.98	Jharkhand
2010	SHIRPUR	N568	MCA	Grace Con	MALE	7.21E+09	Aliquam.v	07/28/201	69%	04/17/201	69%	09/22/201	2.51	Uttar Pradesh
2011	SHIRPUR	F628	IT	Chester Bi	FEMALE	5.4E+09	fermentu	11/29/201	70%	#########	66%	#########	2.7	Andhra Pradesh
2012	SHIRPUR	J533	CS	Ulla Nortc	MALE	9.1E+09	gravida@e	########	71%	03/21/201	68%	12/14/201	2.88	Haryana
2013	MUMBAI	P905	IT	Curran Ho	MALE	3.3E+09	Ut@temp	########	74%	02/26/201	70%	08/30/201	2.41	Bihar
2014	MUMBAI	A288	IT	Xavier Har	MALE	9.68E+09	est.mauri	01/15/201	74%	05/23/201	69%	02/21/201	3.06	Andhra Pradesh

Fig 5.1 Main database

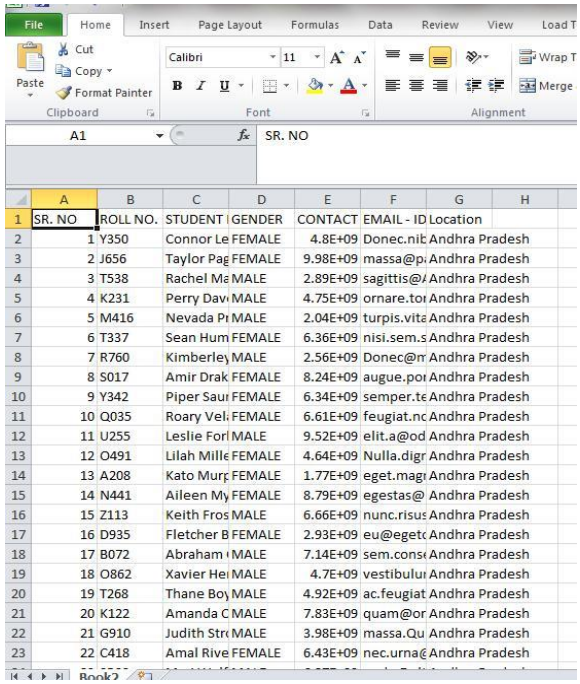
This figure shows the student database of about 2000 students which was collected through a survey.



	A	B	C	D	E
1	SR. NO	CAMPUS	ROLL NO.	BRANCH	
2	1	MUMBAI	Y350	MCA	
3	2	SHIRPUR	J656	MCA	
4	3	MUMBAI	T538	ELEX	
5	4	SHIRPUR	K231	IT	
6	5	SHIRPUR	M416	MTECH	
7	6	MUMBAI	T337	ELEX	
8	7	SHIRPUR	R760	MTECH	
9	8	SHIRPUR	S017	ELEX	
10	9	SHIRPUR	Y342	ELEX	
11	10	MUMBAI	Q035	MCA	
12	11	MUMBAI	U255	EXTC	
13	12	MUMBAI	O491	IT	
14	13	MUMBAI	A208	EXTC	
15	14	MUMBAI	N441	CS	
16	15	MUMBAI	Z113	MTECH	
17	16	MUMBAI	D935	MCA	
18	17	SHIRPUR	B072	CS	
19	18	SHIRPUR	O862	IT	
20	19	SHIRPUR	T268	MTECH	
21	20	SHIRPUR	K122	EXTC	
22	21	SHIRPUR	G910	ELEX	
23	22	SHIRPUR	C418	MTECH	

Figure5.2 Database separated based on branch details

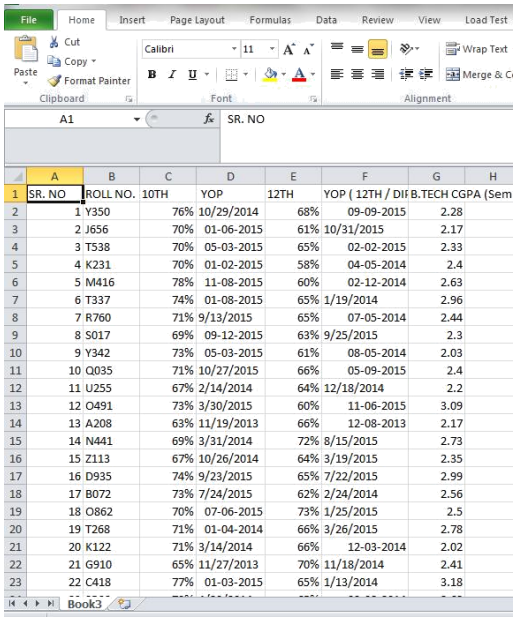
The data warehouse was converted into a data mart comprising of branch and campus detail.



	A	B	C	D	E	F	G	H
1	SR. NO	ROLL NO.	STUDENT	GENDER	CONTACT	EMAIL - ID	Location	
2	1	Y350	Connor Le	FEMALE	4.8E+09	Donec.nit	Andhra Pradesh	
3	2	J656	Taylor Pag	FEMALE	9.98E+09	massa@p	Andhra Pradesh	
4	3	T538	Rachel Ma	MALE	2.89E+09	sagittis@	Andhra Pradesh	
5	4	K231	Perry Dav	MALE	4.75E+09	ornare.to	Andhra Pradesh	
6	5	M416	Nevada Pi	MALE	2.04E+09	turpis.vita	Andhra Pradesh	
7	6	T337	Sean Hum	FEMALE	6.36E+09	nisi.sem.s	Andhra Pradesh	
8	7	R760	Kimberley	MALE	2.56E+09	Donec@m	Andhra Pradesh	
9	8	S017	Amir Drak	FEMALE	8.24E+09	augue.po	Andhra Pradesh	
10	9	Y342	Piper Saur	FEMALE	6.34E+09	semper.te	Andhra Pradesh	
11	10	Q035	Roary Vel	FEMALE	6.61E+09	feugiat.nc	Andhra Pradesh	
12	11	U255	Leslie For	MALE	9.52E+09	elit.a@od	Andhra Pradesh	
13	12	O491	Lilah Mille	FEMALE	4.64E+09	Nulla.digr	Andhra Pradesh	
14	13	A208	Kato Murp	FEMALE	1.77E+09	eget.magr	Andhra Pradesh	
15	14	N441	Aileen My	FEMALE	8.79E+09	egestas@	Andhra Pradesh	
16	15	Z113	Keith Fros	MALE	6.66E+09	nunc.risus	Andhra Pradesh	
17	16	D935	Fletcher B	FEMALE	2.93E+09	eu@eget	Andhra Pradesh	
18	17	B072	Abraham (	MALE	7.14E+09	sem.cons	Andhra Pradesh	
19	18	O862	Xavier Hei	MALE	4.7E+09	vestibulu	Andhra Pradesh	
20	19	T268	Thane Boy	MALE	4.92E+09	ac.feugiat	Andhra Pradesh	
21	20	K122	Amanda C	MALE	7.83E+09	quam@or	Andhra Pradesh	
22	21	G910	Judith Str	MALE	3.98E+09	massa.Qu	Andhra Pradesh	
23	22	C418	Amal Rive	FEMALE	6.43E+09	nec.urna@	Andhra Pradesh	

Figure5.3 Separation based on student details

The data warehouse was converted into a data mart comprising of student personal details.



SR. NO	ROLL NO.	10TH	YOP	12TH	YOP (12TH / DIP)	B.TECH CGPA (Sem
1	Y350	76%	10/29/2014	68%	09-09-2015	2.28
2	J656	70%	01-06-2015	61%	10/31/2015	2.17
3	T538	70%	05-03-2015	65%	02-02-2015	2.33
4	K231	70%	01-02-2015	58%	04-05-2014	2.4
5	M416	78%	11-08-2015	60%	02-12-2014	2.63
6	T337	74%	01-08-2015	65%	1/19/2014	2.96
7	R760	71%	9/13/2015	65%	07-05-2014	2.44
8	S017	69%	09-12-2015	63%	9/25/2015	2.3
9	Y342	73%	05-03-2015	61%	08-05-2014	2.03
10	Q035	71%	10/27/2015	66%	05-09-2015	2.4
11	U255	67%	2/14/2014	64%	12/18/2014	2.2
12	O491	73%	3/30/2015	60%	11-06-2015	3.09
13	A208	63%	11/19/2013	66%	12-08-2013	2.17
14	N441	69%	3/31/2014	72%	8/15/2015	2.73
15	Z113	67%	10/26/2014	64%	3/19/2015	2.35
16	D935	74%	9/23/2015	65%	7/22/2015	2.99
17	B072	73%	7/24/2015	62%	2/24/2014	2.56
18	O862	70%	07-06-2015	73%	1/25/2015	2.5
19	T268	71%	01-04-2014	66%	3/26/2015	2.78
20	K122	71%	3/14/2014	66%	12-03-2014	2.02
21	G910	65%	11/27/2013	70%	11/18/2014	2.41
22	C418	77%	01-03-2015	65%	1/13/2014	3.18

Figure5.4 Separation based on marks

The data warehouse was converted into a data mart comprising of academic details of the student.

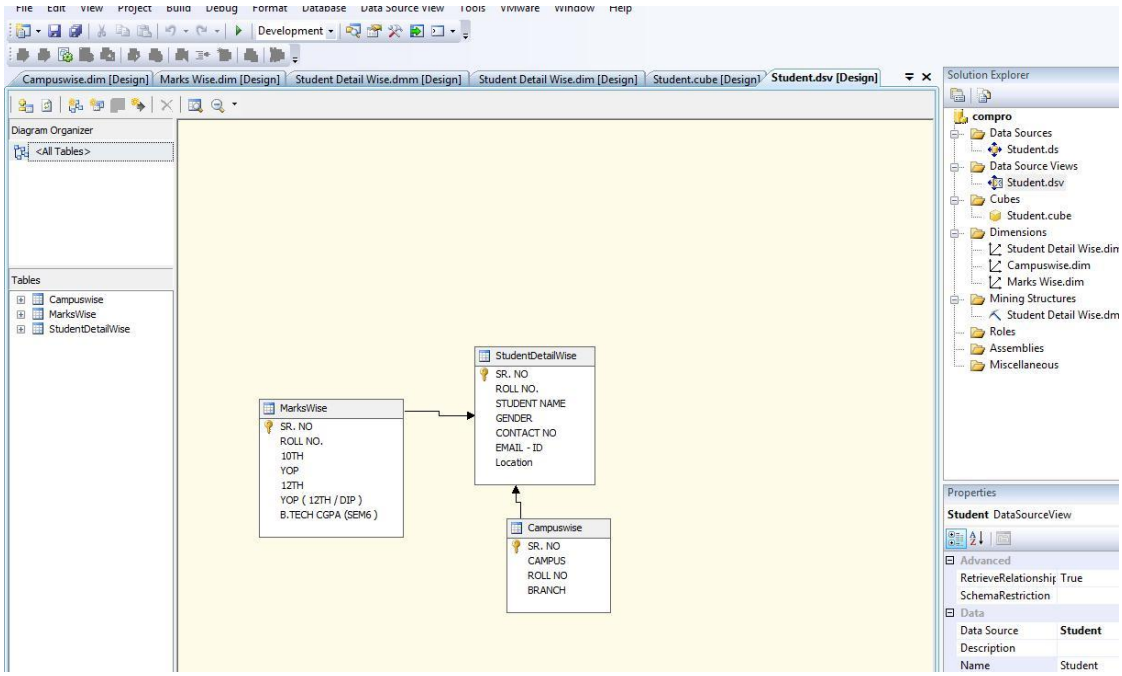


Figure5.5 Data Cube connected classes with SR.NO. as primary key

A connection was made to the software and relation between data marts was established with Sr. No as the foreign key.

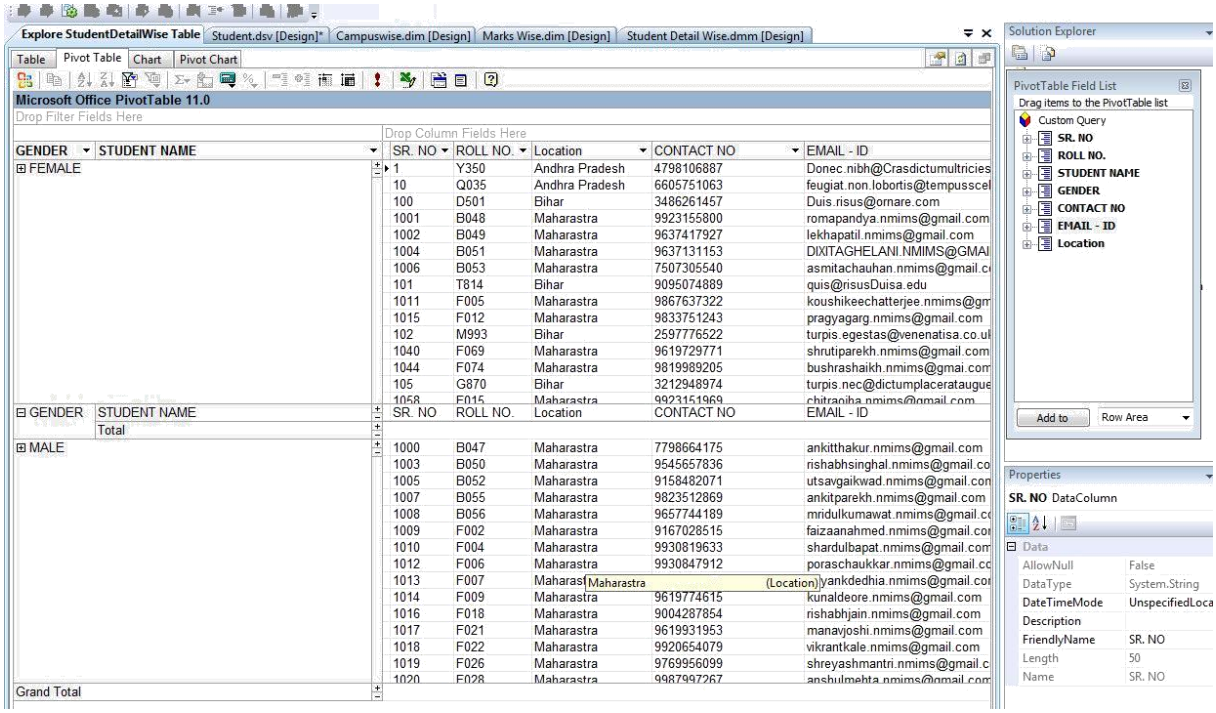


Figure5.6 Pivot Table

A pivot table was made with separating MALE and FEMALE students with ease of view

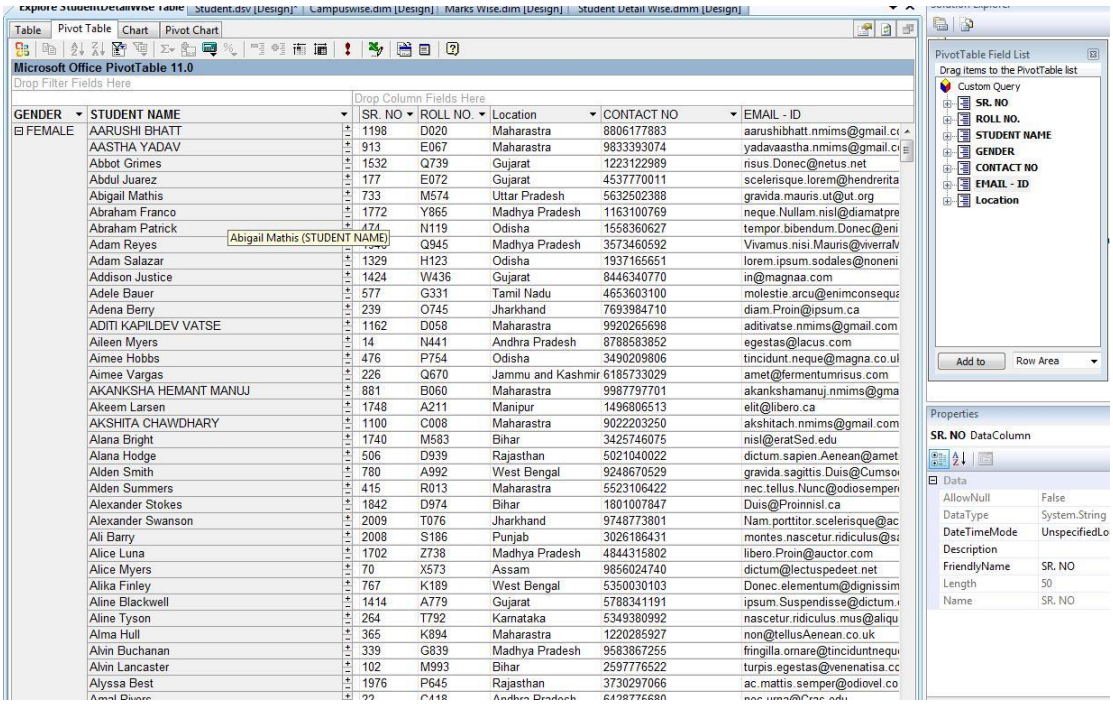


Figure5.7 'Female' category (similar can be done for Male)

Extending View of the Female category in the pivot table for convenience.

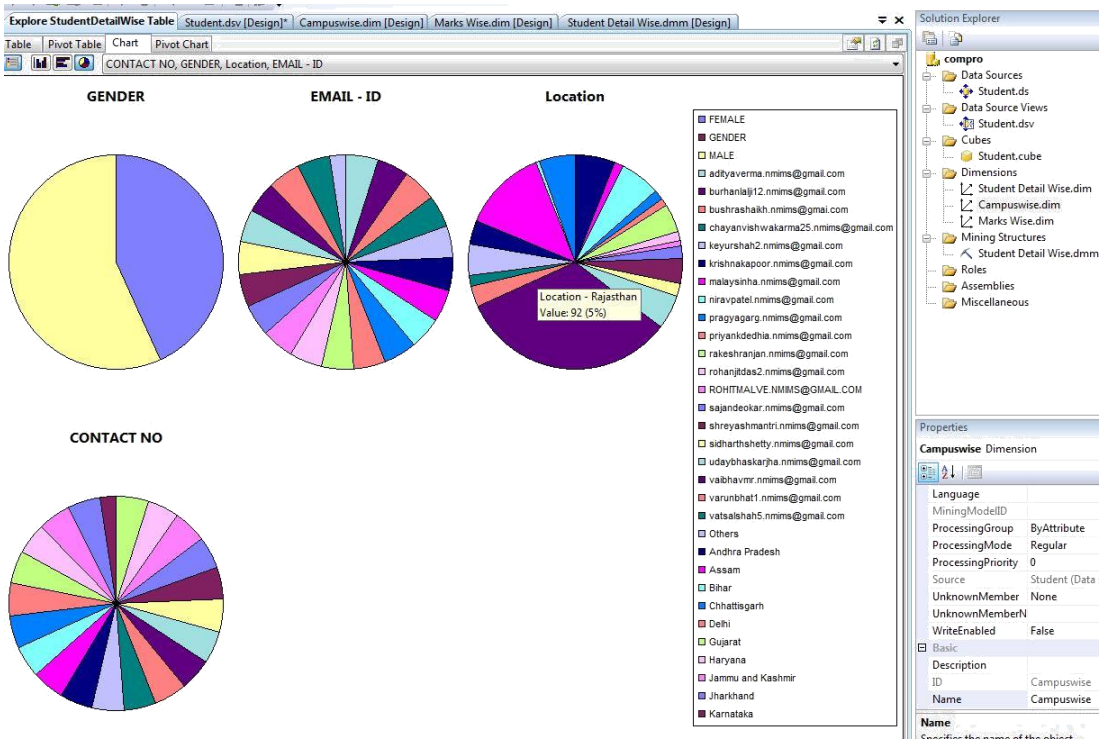


Figure5.8 Pie charts

Comparative Pie Charts and Legend for various combinations and analysis about the percentage of population

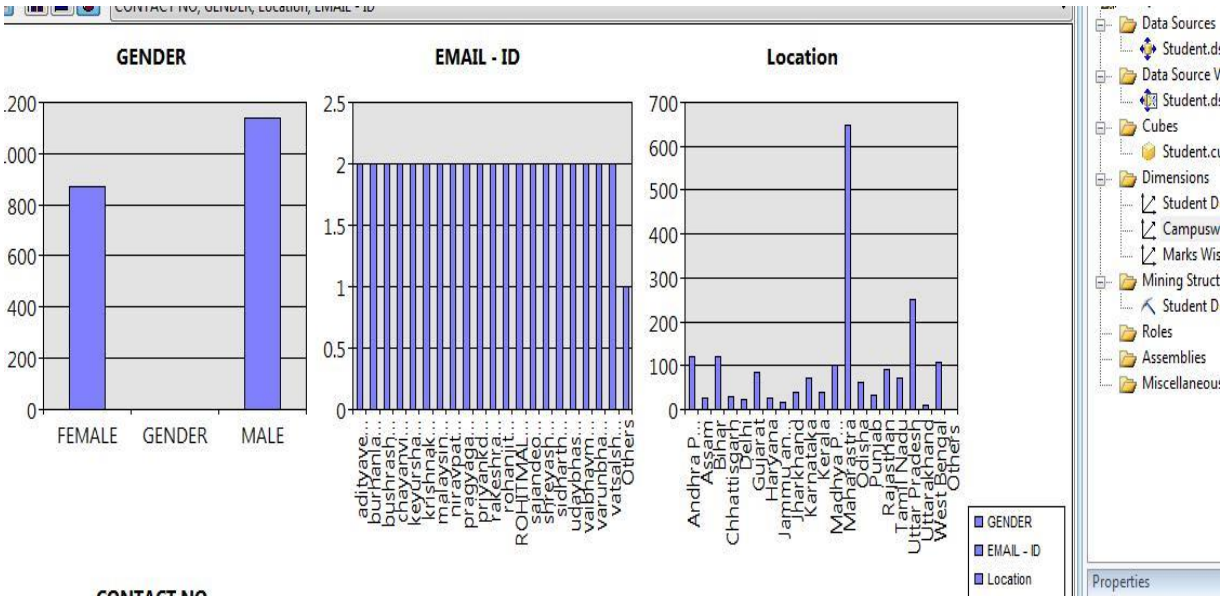


Figure5.9 Bar graphs

Bar graphs representing number of students based on gender and per state admissions

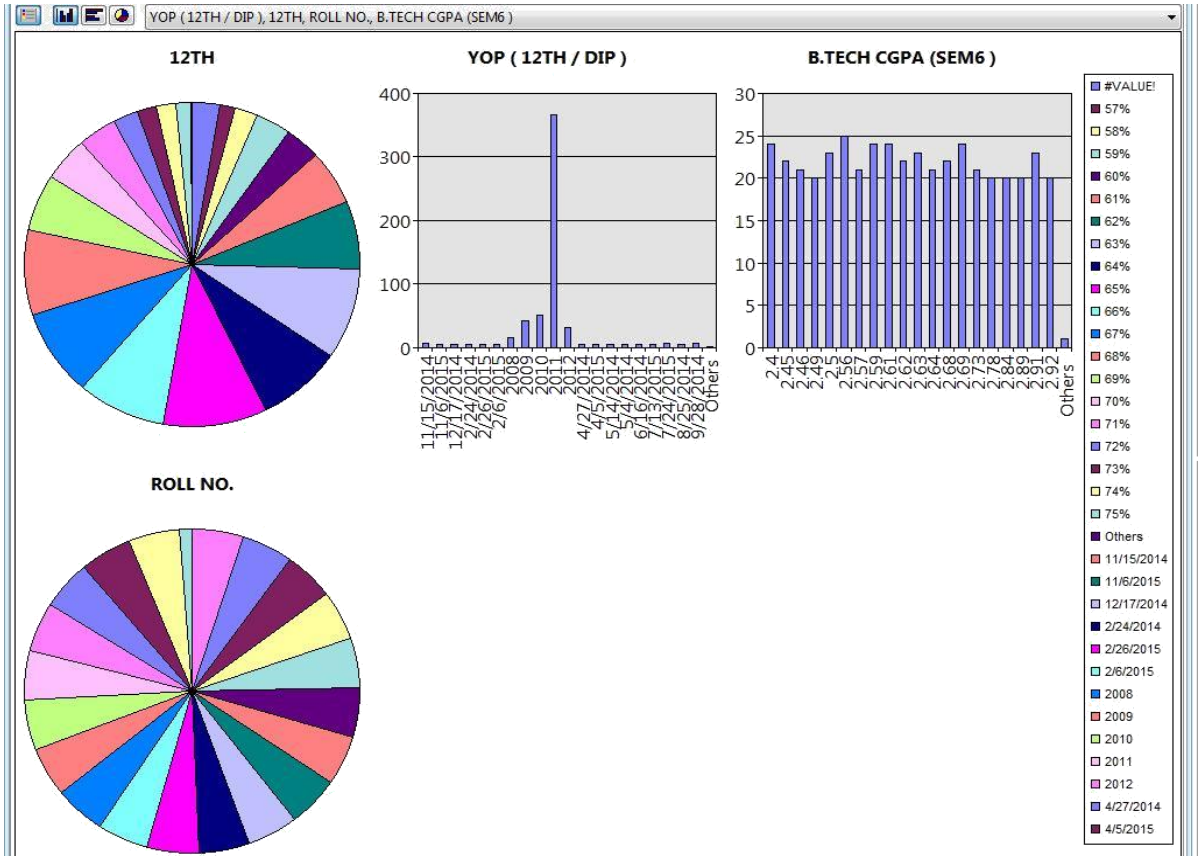


Figure5.10 Pie charts and graphs

Charts representing CGPA and 12th marks segregation depicting number of students falling under each category

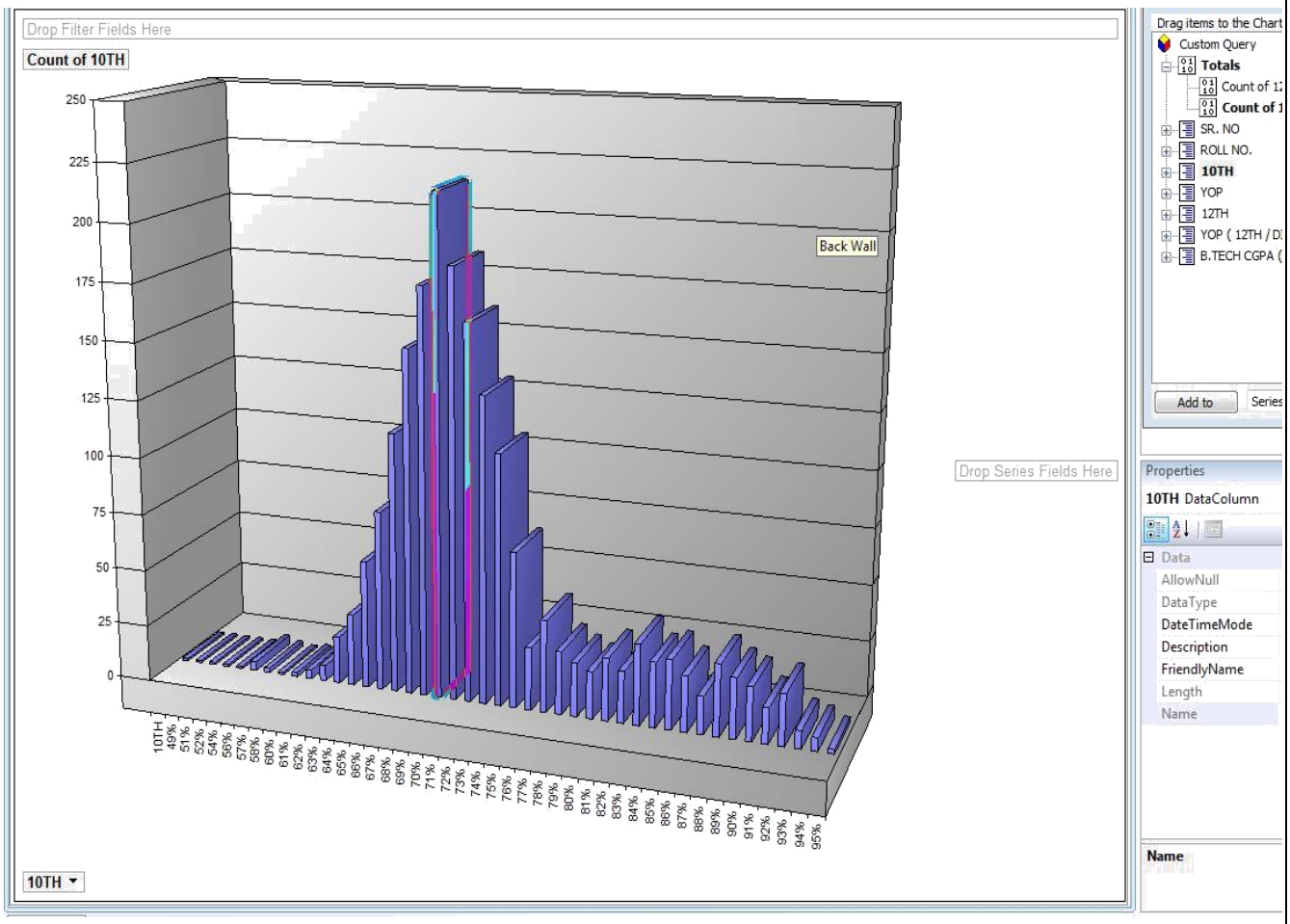


Figure5.11 Number of students in the campus 10th percentage wise

We combined 2 attributes 10th standard percentage and the number of students giving us a 3 dimensional view stating maximum students of that year was of 71%

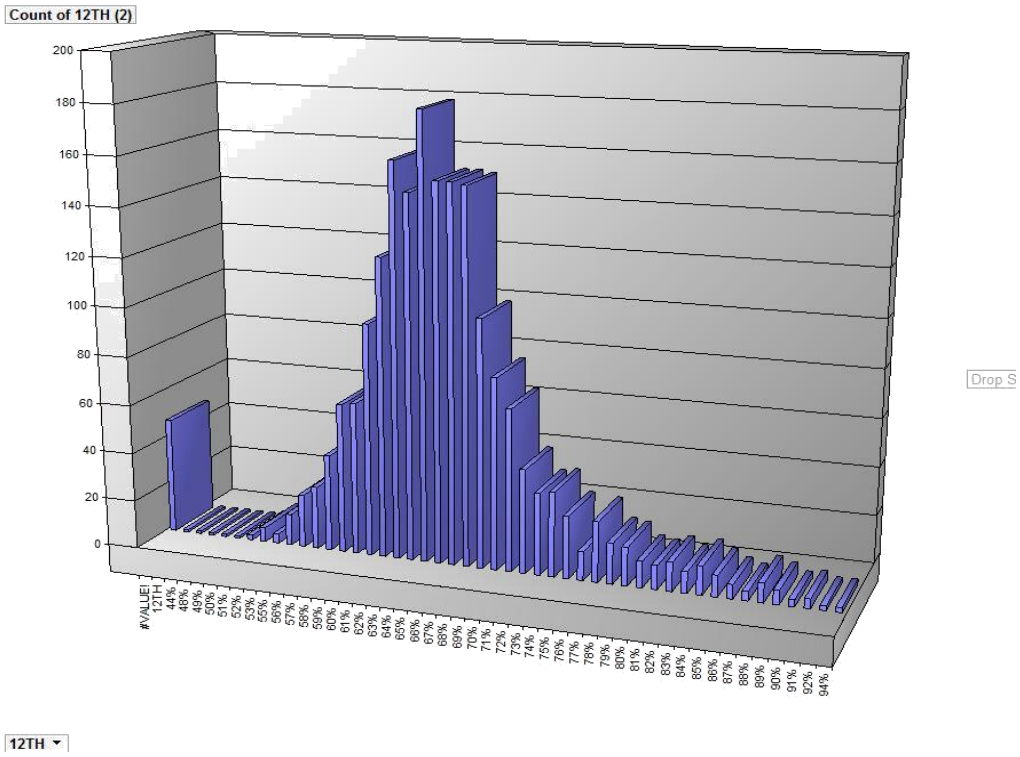


Figure5.12 Number of students in campus plotted against the 12th percentages

We combined 2 attributes 12th standard percentage and the number of students giving us a 3 dimensional view stating maximum students of that year was of 66%

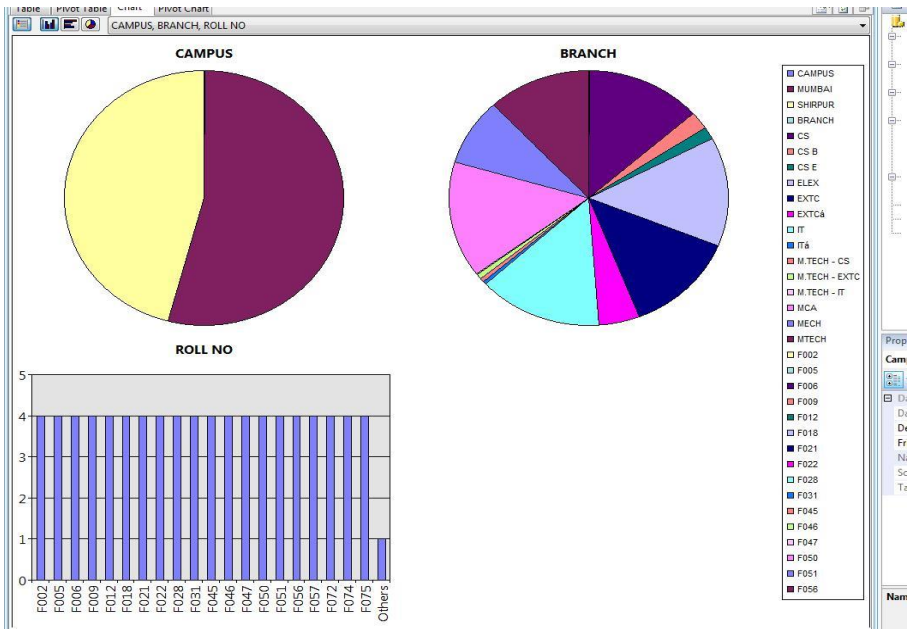


Figure5.13 Campus and Branch wise distribution in a pie representation

This pie chart shows the division of the students according to different branches and campus

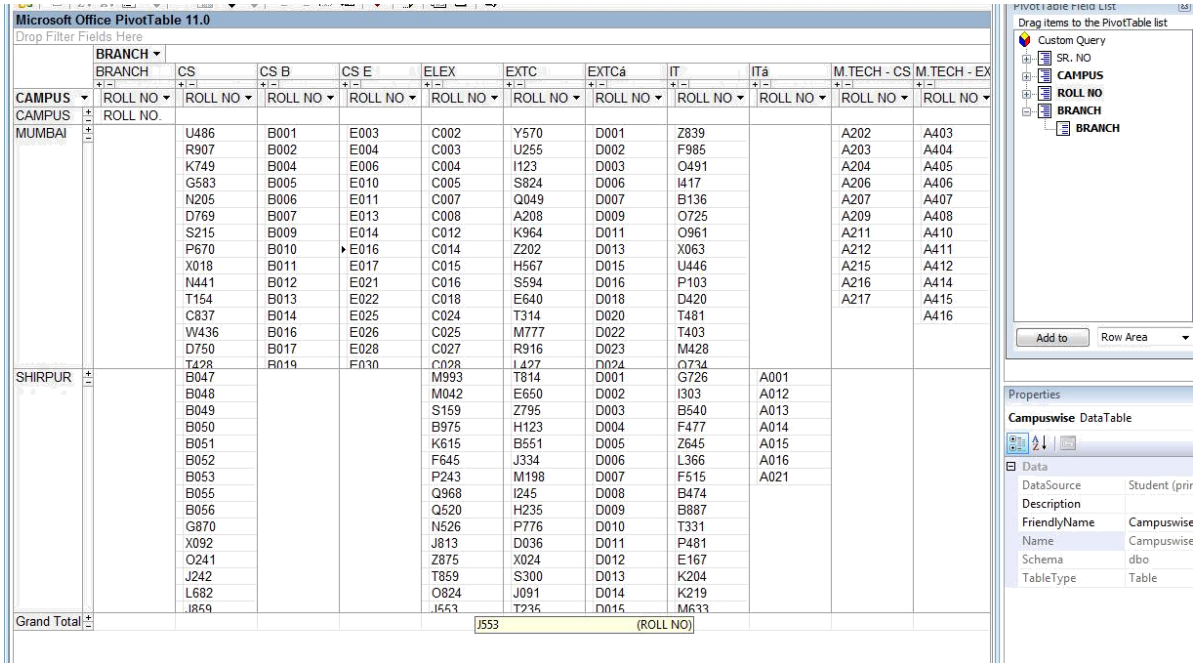


Figure5.14 Pivot table to segregate each student by campus allotted

A pivot table was made with separating the students based on the campus

CHAPTER 6

DATA MINING TECHNIQUES

Many different data mining, query model, processing model, and data collection techniques are available. We Examined different data mining and analytics techniques and solutions, and learn how to build them using existing software and installations. We explored the different data mining tools that are available, and learnt how to determine whether the size and complexity of your information might result in processing and storage complexities.

6.1 K-Means

- Construct a partition of a database D of n objects into a set of k clusters
- Given a k, find a partition of k clusters that optimizes the chosen partitioning criterion
- Basic version works with numeric data only

    1) Pick a number (K) of cluster centers - centroids (at random)

    2) Assign every item to its nearest cluster center (e.g. using Euclidean distance)

    3) Move each cluster center to the mean of its assigned items

    4) Repeat steps 2, 3 until convergence (change in cluster assignments less than a threshold)

Analysis:

- Result can vary significantly depending on initial choice of seeds
- Can get trapped in local minimum
- To increase chance of finding global optimum: restart with different random seeds

Advantages

• Simple, understandable

• Items automatically assigned to clusters

Disadvantages

• Must pick number of clusters before hand

• Often terminates at a local optimum.

• All items forced into a cluster

• Too sensitive to outliers (Since an object with an extremely large value may substantially distort the distribution of the data)

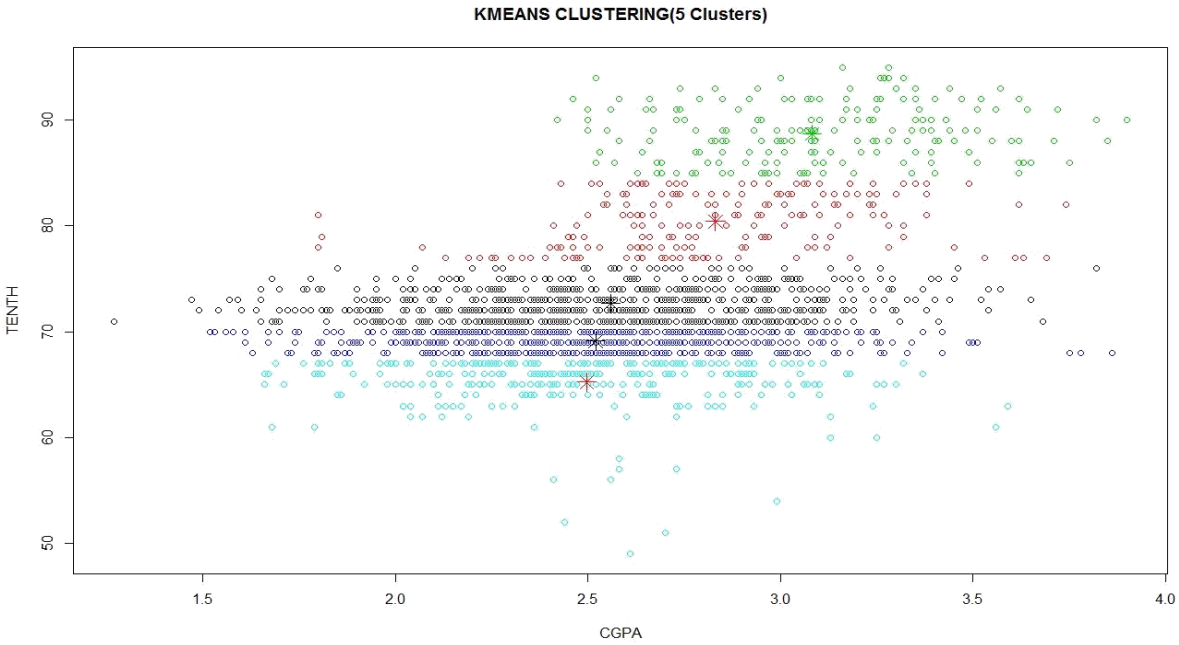


Figure6.1 K-means clustering with 5 clusters for CGPA and Tenth

This figure makes use of K-means algorithm and finds 5 clusters based on their centroids. And since the results cluster only the tenth marks and not the cgpa this result is not suitable for our purpose.

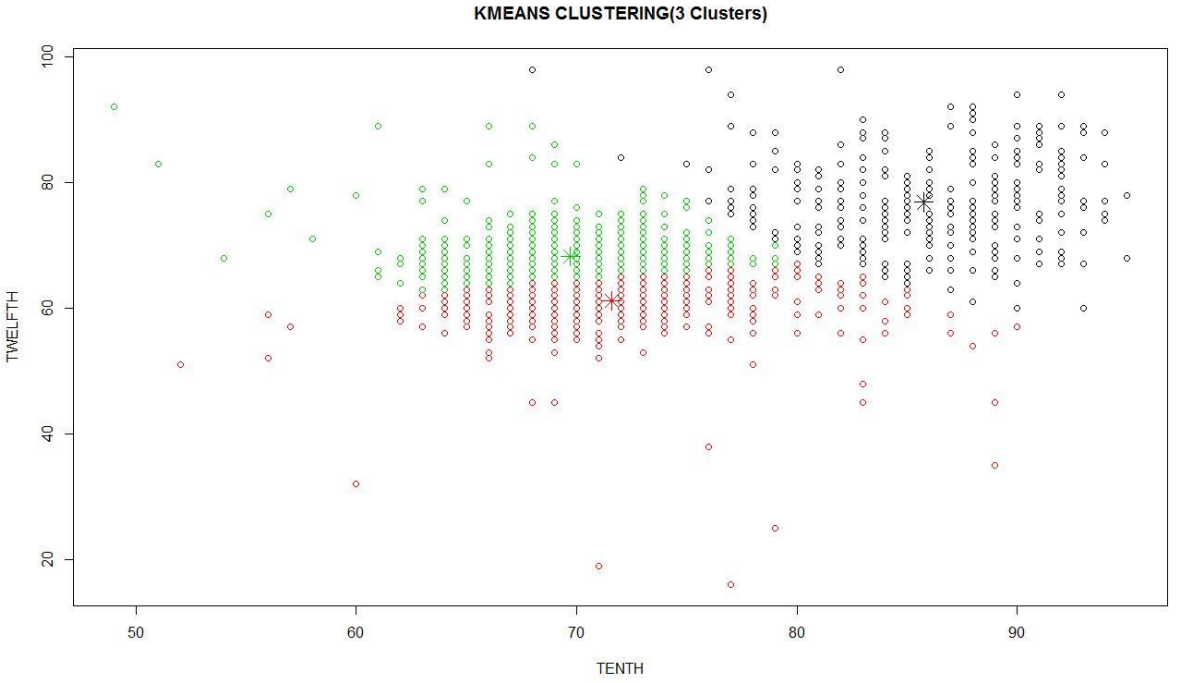


Figure6.2 K-means clustering with 3 clusters for Twelfth and Tenth

This figure makes use of K-means algorithm and finds 3 clusters based on their centroids. In this figure the outliers are not detected and are merged in the 3 clusters, hence this method is not suitable for accurate clustering.

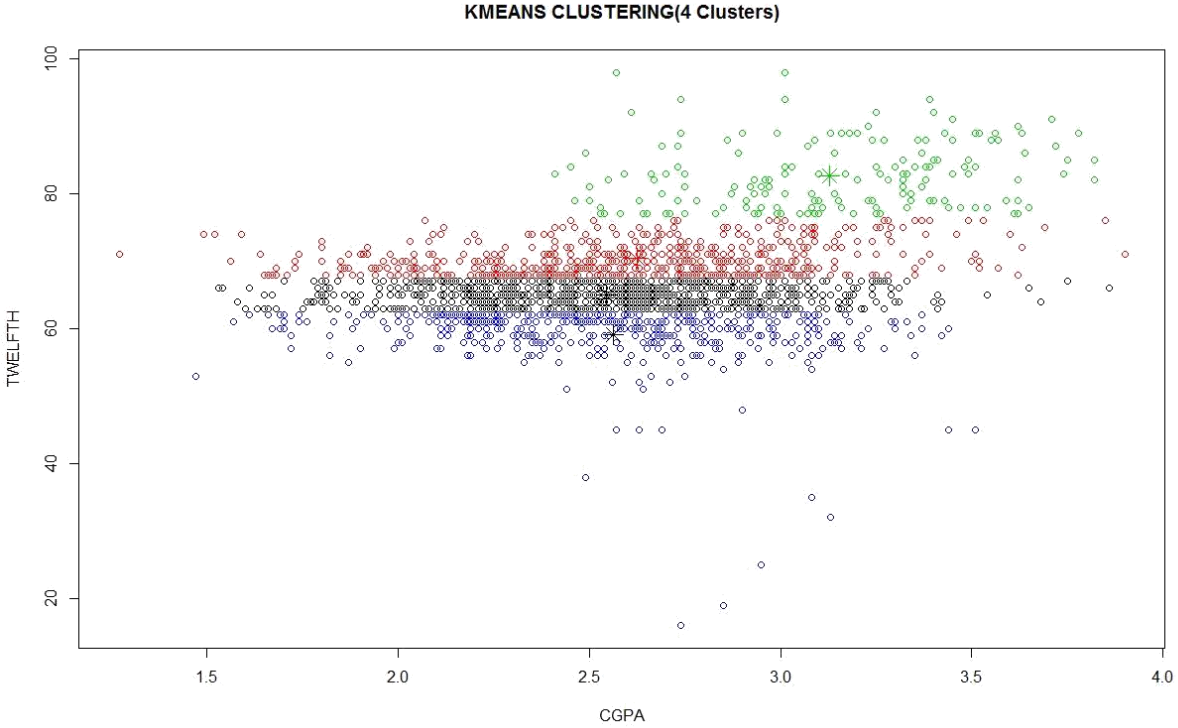


Figure 6.3 K-means clustering with 4 clusters for CGPA and Twelfth

In this figure we compare the twelfth and CGPA grades. We can see from the figure that the algorithm does not cluster the cgpa as per need when compared to twelfth marks.

6.2 Dendogram

It means to build a tree-based hierarchical taxonomy (dendrogram) from a set of unlabelled examples. It implies that a Recursive application of a standard clustering algorithm can produce a hierarchical clustering. A clustering of the data objects is obtained by cutting the dendrogram at the desired level, and then each connected component forms a cluster. The types are:

Bottom up (agglomerative)

- Start with single-instance clusters
- At each step, join the two closest clusters
- Design decision: distance between clusters
- Example: two closest instances in clusters vs. distance between means

Top down (divisive approach / deglomerative)

- Start with one universal cluster
- Find two clusters
- Proceed recursively on each subset

• Can be very fast

Algorithm:

- Start with all instances in their own cluster.
- Until there is only one cluster:
- Among the current clusters, determine the two clusters, ci and cj, which are most similar.
- Replace ci and cj with a single cluster ci ∪ cj

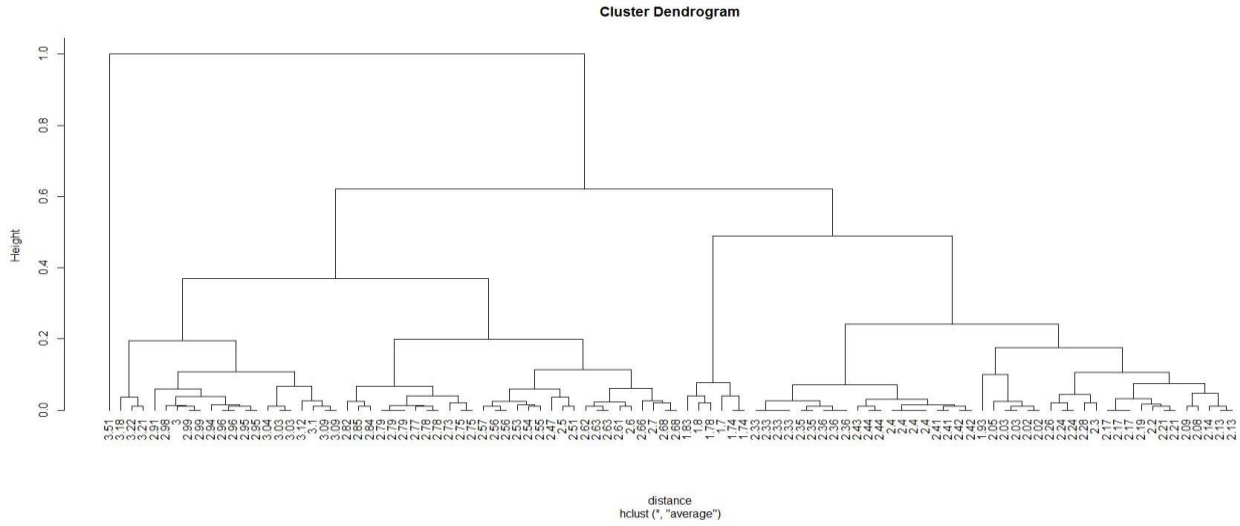


Figure6.4 Hierarchical Cluster dendrogram for CGPA

This figure represents a dendogram based on the cgpa and we can see that the outliers are very clearly visible in this. Hence we can do cluster analysis easily in this.

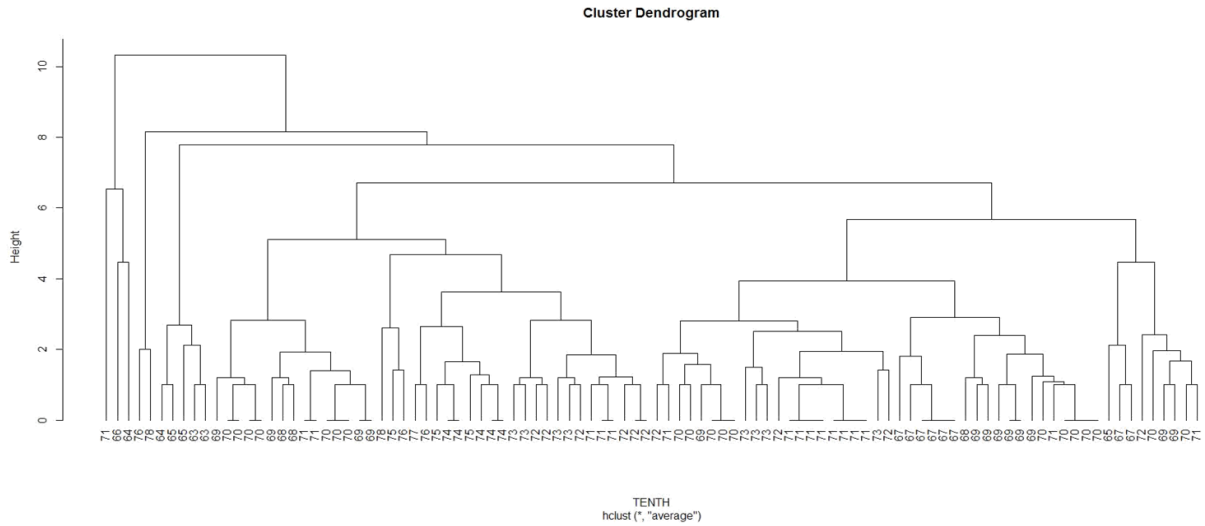


Figure6.5 Hierarchical Cluster dendrogram for 10th Marks

This figure represents a dendogram based on the tenth grades and we can see that the outliers are very clearly visible in this. Hence we can do cluster analysis easily in this.

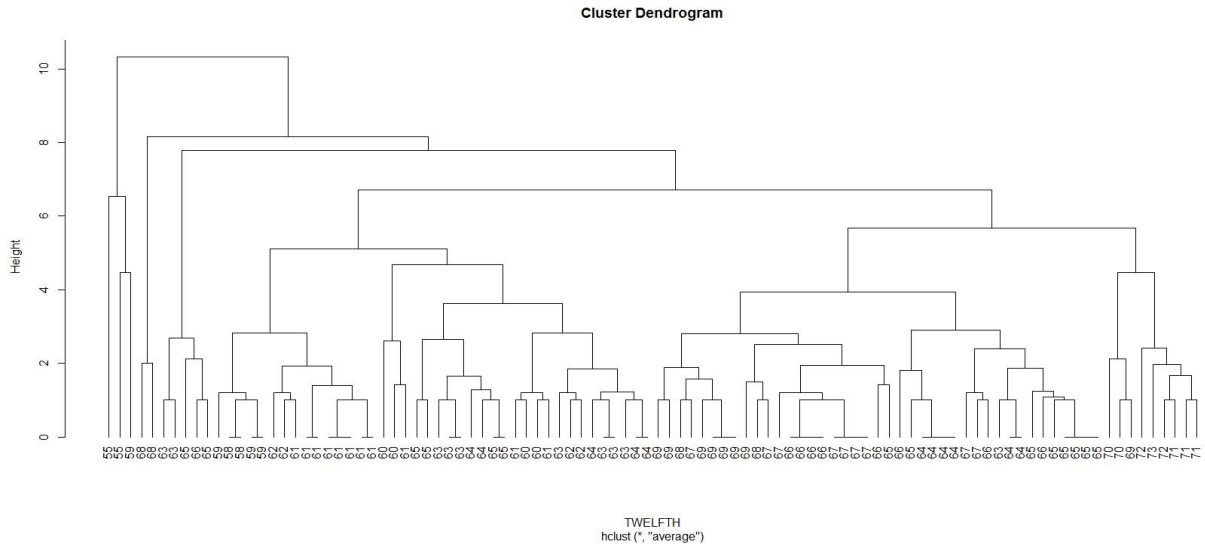


Figure6.6 Hierarchical Cluster dendrogram for 12th Marks

This figure represents a dendogram based on the twelfth and we can see that the outliers are very clearly visible in this. Hence we can do cluster analysis easily in this.

Cluster dendrograms were excellent for outlier analysis and we could choose the best height at which we want to cluster the data and divide them into groups. Using this dendrogram we can pinpoint the data entry that we need and further process it according to algorithm.

6.3 Density Based method

In density-based clustering, clusters are defined as areas of higher density than the remainder of the data set. Objects in these sparse areas - that are required to separate clusters - are usually considered to be noise and border points. It is based on connecting points within certain distance thresholds. It only connects points that satisfy a density criterion, in the original variant defined as a minimum number of other objects within this radius.

The key drawback of DBSCAN and OPTICS is that they expect some kind of density drop to detect cluster borders. Moreover, they cannot detect intrinsic cluster structures which are prevalent in the majority of real life data

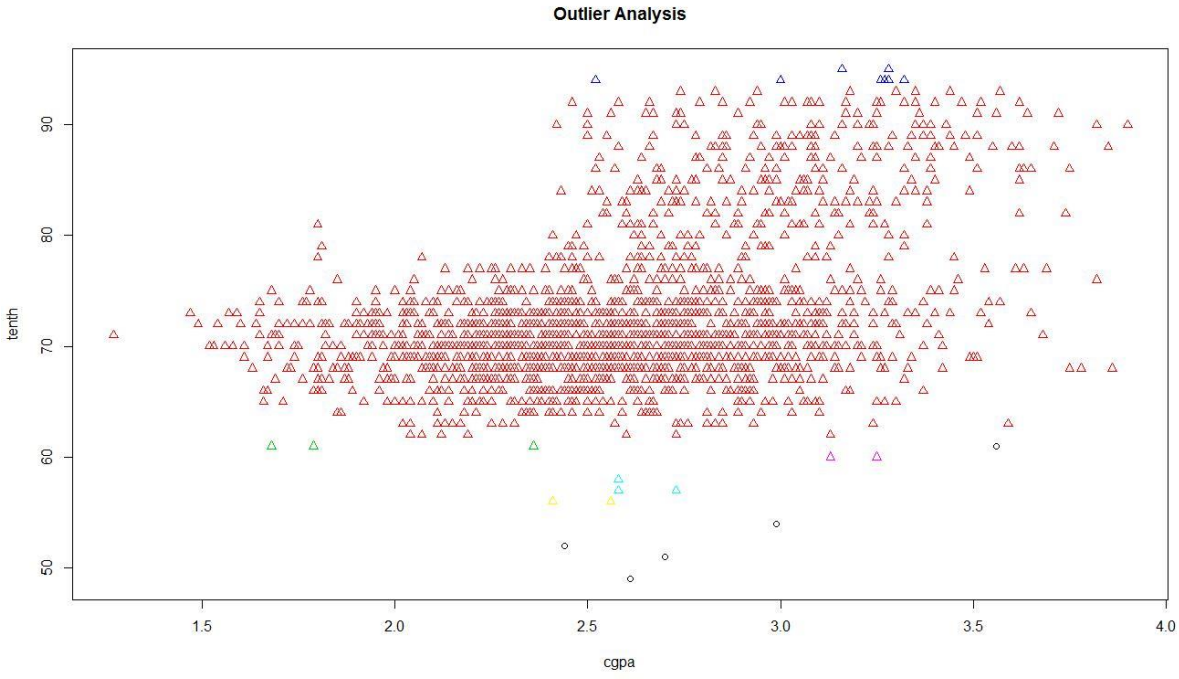


Figure6.7 Density based Analysis for CGPA and 10th for outlier analysis

DB Scan method is the best method for analysis of the outliers. The blue nodes depict the outliers.

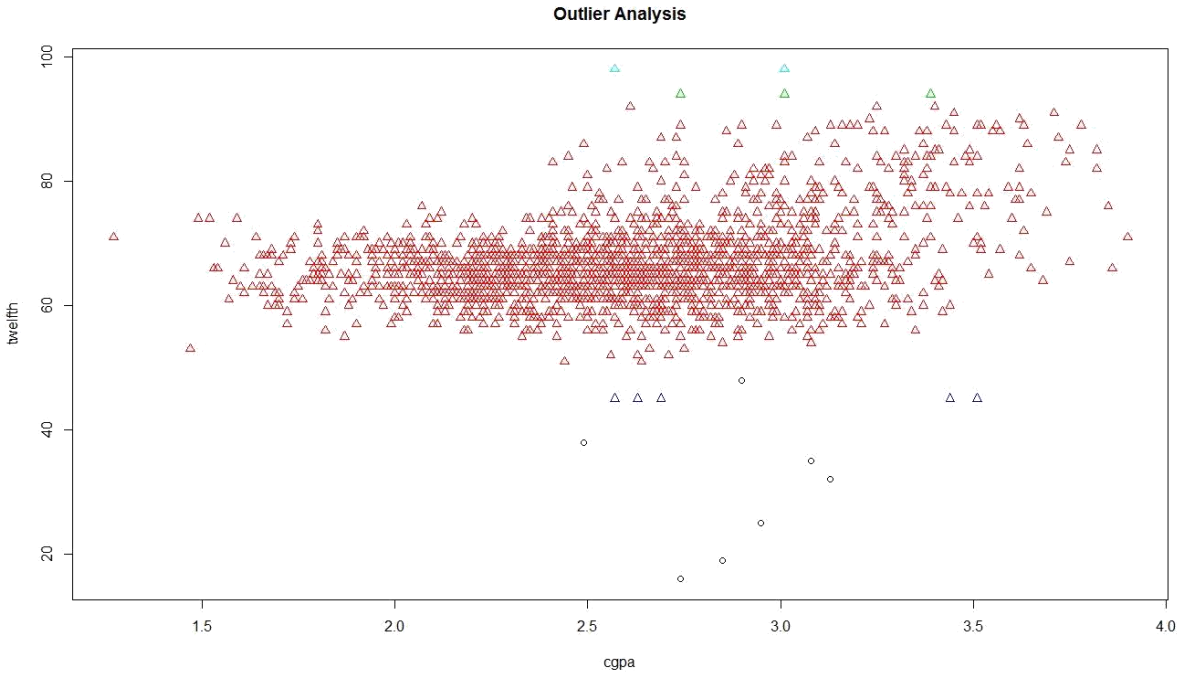


Figure6.8 Density based Analysis for CGPA and 12th for outlier analysis

DB Scan method is the best method for analysis of the outliers. The blue nodes depict the outliers.

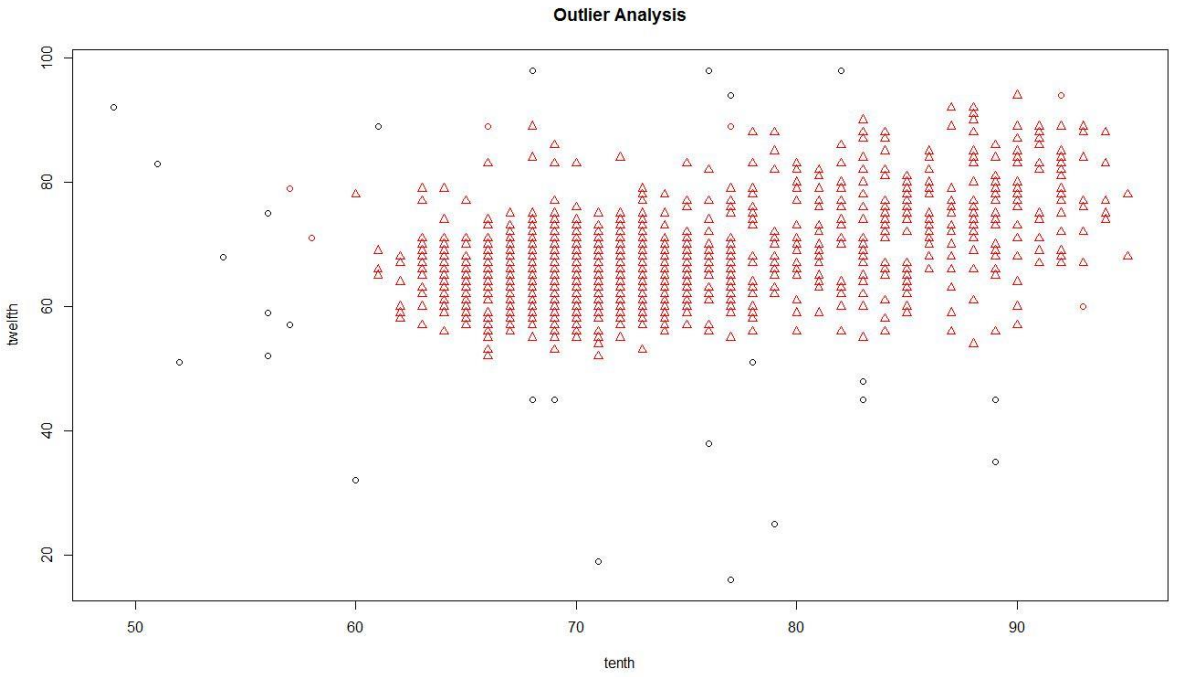


Figure6.9 Density based Analysis for 10th and 12th for outlier analysis

DB Scan method is the best method for analysis of the outliers. The blue nodes depict the outliers.

Density based clustering clearly shows outliers and the clusters can be formed in any shape as they are formed based on density. We can identify outliers, extract them and do a further analysis on them with separate algorithm

6.4 Fuzzy C means method

Fuzzy c method is also known as 'soft clustering'. It gives probabilities that an instance belongs to each of a set of clusters. Each instance is assigned a probability distribution across a set of discovered categories (probabilities of all categories must sum to 1).

In fuzzy clustering, data elements can belong to more than one cluster, and associated with each element is a set of membership levels. These indicate the strength of the association between that data element and a particular cluster. Fuzzy clustering is a process of assigning these membership levels, and then using them to assign data elements to one or more clusters The FCM algorithm attempts to partition a finite collection of n elements X = { x_1, . . , x_n} into a collection of c fuzzy clusters with respect to some given criterion:

Given a finite set of data, the algorithm returns a list of c cluster centres C={C1,…,Cc} and a partition matrix, W=w where each element w_{ij} tells the degree to which element x_i belongs to cluster c_j

The fuzzy c-means algorithm is very similar to the k-means algorithm: Choose a number of clusters.

- Assign randomly to each point coefficients for being in the clusters.
- Repeat until the algorithm has converged (that is, the coefficients' change between two iterations is no more than ε, the given sensitivity threshold)
- Compute the centroid for each cluster, using the formula above.
- For each point, compute its coefficients of being in the clusters, using the formula above.

The algorithm minimizes intra-cluster variance as well, but has the same problems as k-means; the minimum is a local minimum, and the results depend on the initial choice of weights.

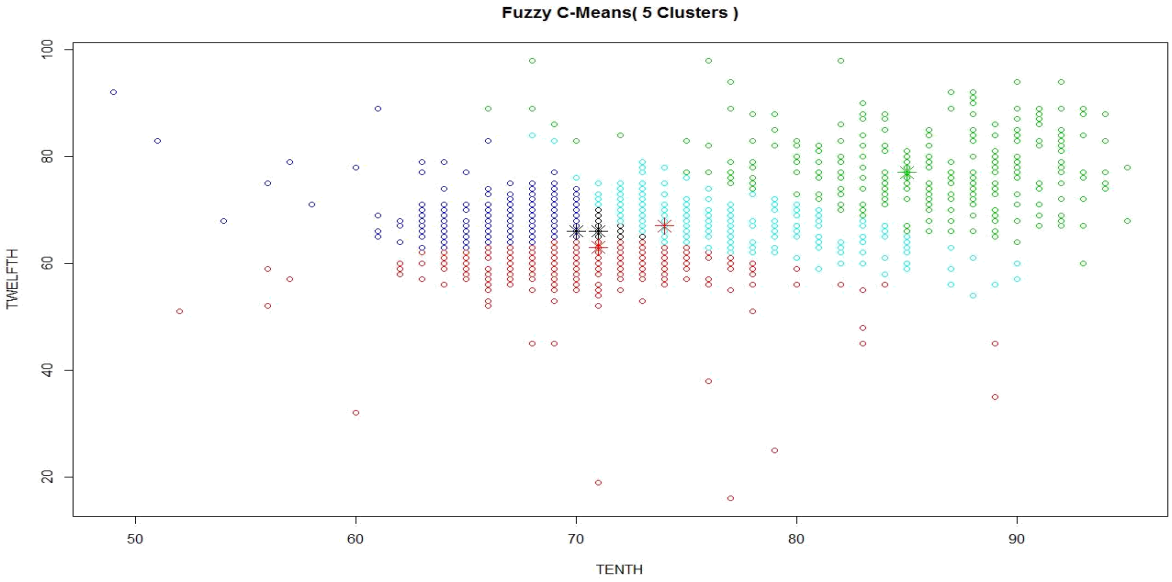


Figure6.10 Fuzzy C-means analysis of 10th and 12th

The result obtained is almost like k means, the centroids are assigned according to the probability which is determined by fuzzy logic. The numbers of clusters formed are 5 which are depicted by different colours.

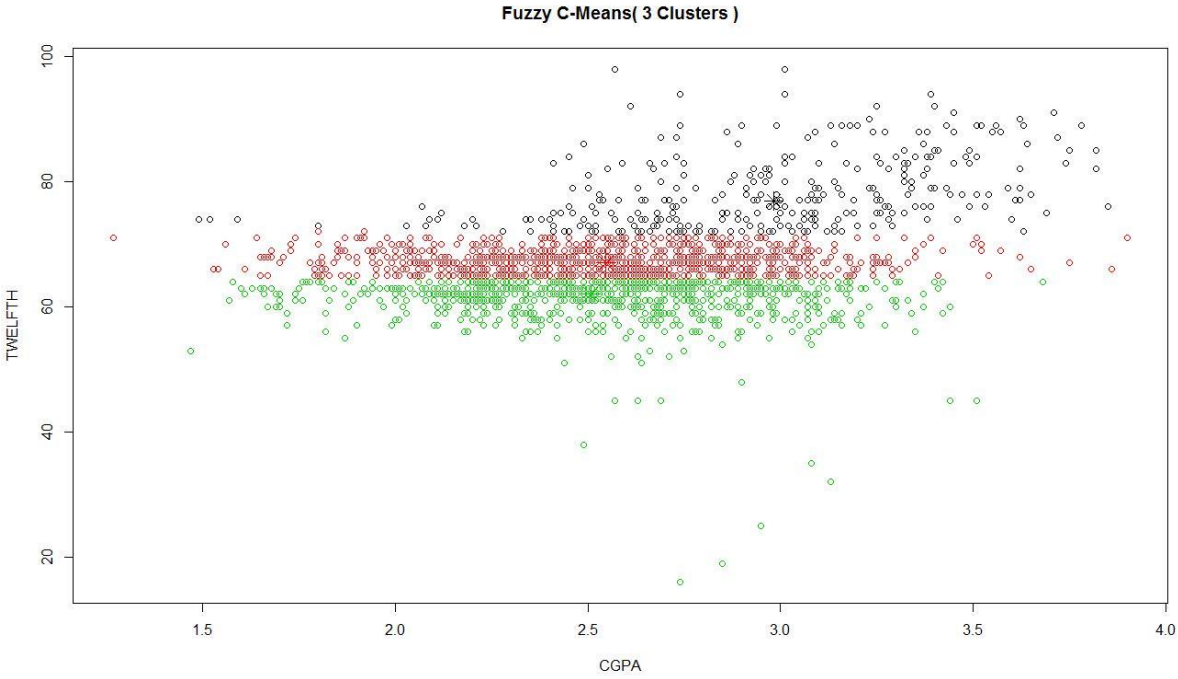


Figure6.11 Fuzzy C-means analysis of CGPA and 12th

The result obtained is almost like k means, the centroids are assigned according to the probability which is determined by fuzzy logic. The numbers of clusters formed are 3 which are depicted by different colours.

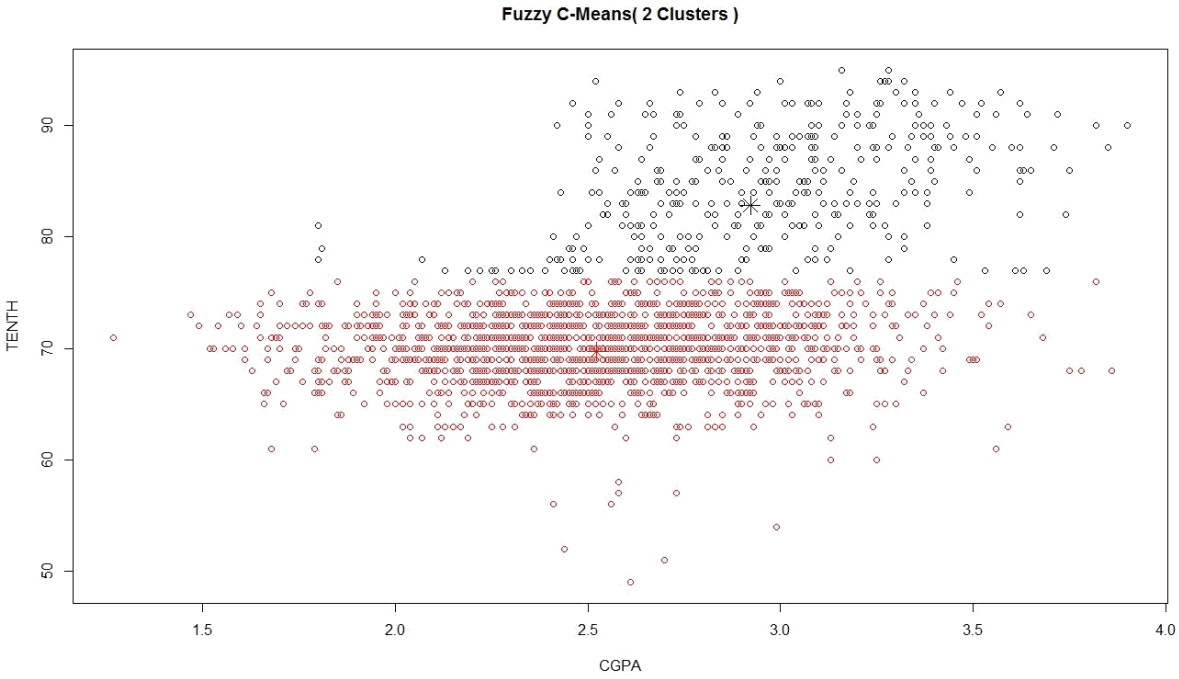


Figure6.12 Fuzzy C-means analysis of CGPA and 10th

The result obtained is almost like k means, the centroids are assigned according to the probability which is determined by fuzzy logic. The numbers of clusters formed are 3 which are depicted by different colours.

6.5 Mixed Gaussian Method

Intuitively it is a tightly packed ball-shape like thing. Given a set of data points, and assume that we know there are k clusters in the data, we need to:

- Assign the data points to the k clusters (soft assignment)
- Learn the Gaussian distribution parameters for each cluster: μ and 5.

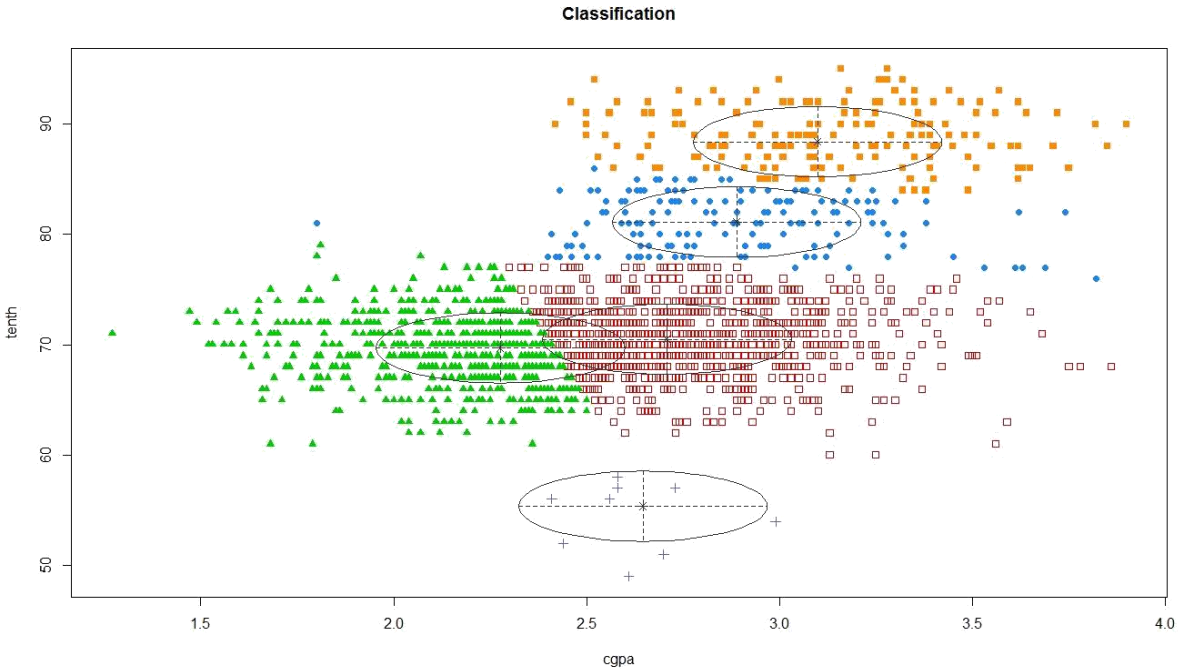


Figure6.13 Gaussian Clustering Method for CGPA and 10th

As we can see above this is the best method and gives the best analysis of our data. It uses ellipses to find out the clusters. Hence, it gives better result as compared previous algorithms.

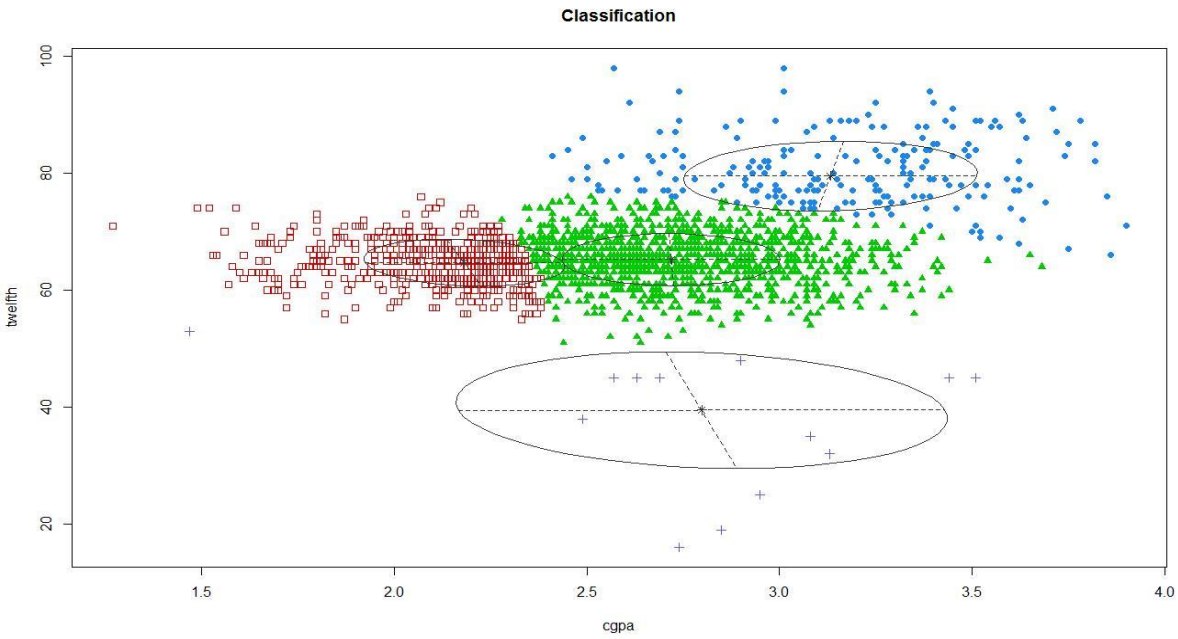


Figure6.14 Gaussian Clustering Method for CGPA and 12th

As we can see above this is the best method and gives the best analysis of our data. It uses ellipses to find out the clusters. Hence, it gives better result as compared previous algorithms.

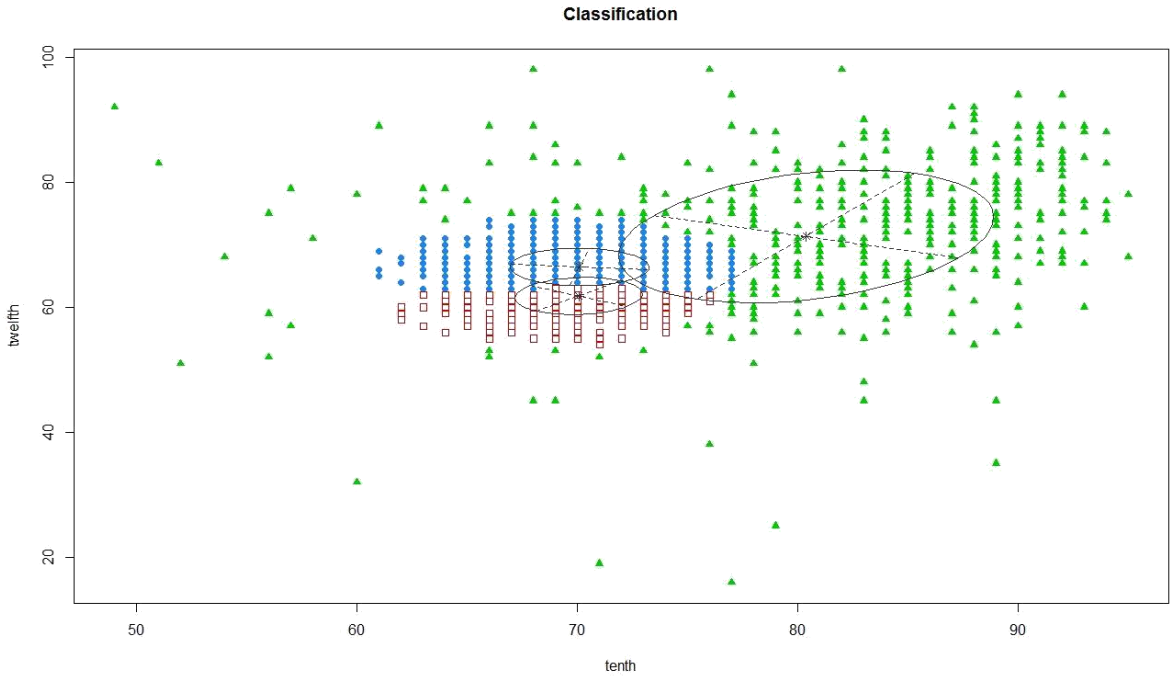


Figure6.15 Gaussian Clustering Method for 10^{th} and 12^{th}

As we can see above this is the best method and gives the best analysis of our data. It uses ellipses to find out the clusters. Hence, it gives better result as compared previous algorithms.

CHAPTER 7

RESULTS AND INFERENCES

After our base for the analysis was developed we performed a comparative study of all the methods used based on different comments and categories.

Table7.1. Algorithms Comparison

	K-MEANS	**FUZZY CMEANS**	**DENDROGRAM (HIERARCHICAL)**	**DENSITY BASED**	**MULTI GUASSIAN**
Description	Partition based clustering algorithm.	Partition based on probability of belonging to each cluster.	Build a tree-based hierarchical taxonomy (dendrogram) from a set of unlabelled examples.	Clustering based on density (local cluster criterion), such as density-connected points.	Clusering with elliptical shape and different orientation.
Technique Used	Each cluster is represented by the center of the cluster using centorid method.	Weighted centroid based is on those probabilities using fuzzy algorithm.	Recursive application of a standard clustering algorithm can produce a hierarchical clustering.	Each cluster has a considerable higher density of points than outside of the cluster	The order of complexity is similar to K-Means with a larger constant. It also requires K to be specified.
Results based on initial choice	Result can vary significantly depending on initial choice of seeds	Different centroids are formed when using different degree of fuzziness.	Decompose data objects into a several levels of nested partitioning (tree of clusters), and varies based on input data.	In density based cluster, a cluster is extend along the density distribution. We can select the amount of density initially according to our needs.	Multi-Gaussian can discover cluster with elliptical shape with different orientation and hence it is more general than K-Means.

Other Facts	Can get trapped in local minimum. To increase chance of finding global optimum: restart with different random seeds	Processes of initialization, iteration, and termination are the same as the ones used in k-means. The resulting clusters are best analyzed as probabilistic distributions rather than a hard assignment of labels.	A clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster.	Unlike other cluster, density based cluster can have some outliers (data points that doesn't belong to any clusters). On the other hand, it can detect cluster of arbitrary shapes (doesn't have to be circular at all)	Unlike K-Means whose cluster is always in circular shape this has elliptical shape and gives us the best results.
Extra Comments				In density based cluster, a cluster is extend along the density distribution. Two parameters is important: "eps" defines the radius of neighborhood of each point, and "minpts" is the number of neighbors within my "eps" radius.	

Table7.2. Speed Analysis (All values are in seconds)

Data Size	KMEANS	HIERARCHICAL	DENSITY	FUZZY CMEANS
1550	3.9	5.0	3.5	2.5
3550	10.6	12.3	10.9	9.1
7500	21.6	24.1	39.5	17.8
10250	29.2	32.4	78.4	25.1

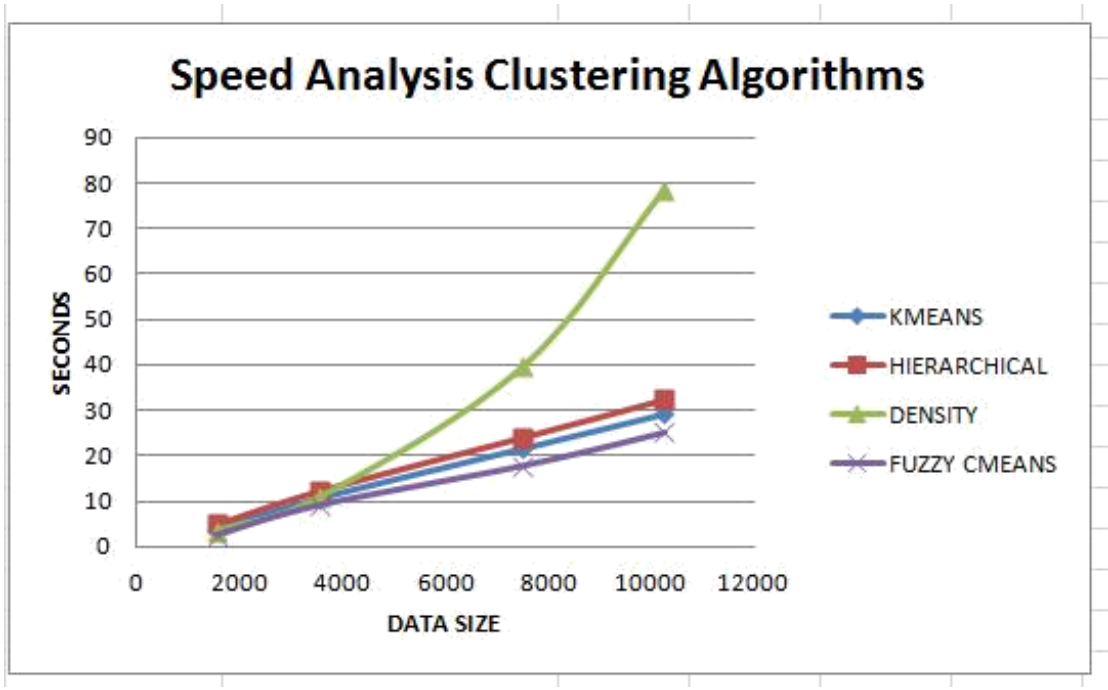


Figure7.1 Speed Analysis Graph

A speed comparison is shown which describes that density increases exponentially as the data size increases and the others show a linear increase.

Table7.3. Inferences from the Algorithms

General Comments	KMEANS	FUZZY CMEANS	DENDROGRAM (HIERARCHICAL)	DENSITY BASED	MULTIGUASSIAN
Favourability	Produces clusters based on distance hence results not favourable in our case.	Similar results to K-means only the centroids seemed to be a little misplaced compared to K-means and hence not too favourable for us.	This method is helpful to pinpoint and extract outliers for further processing.	We mainly implemented this technique to show outlier analysis and this is a very efficient technique which shows the outliers clearly.	This technique is a relatively new technique and works extremely well when used with our data.
Outlier Analysis	Outliers are not identified at all in this analysis and they are forcefully grouped into various clusters	Outliers are not identified at all in this analysis and they are forcefully added into clusters.	Outlier analysis can be very clearly seen in this analysis by looking at the dendrogram formed.	As we can see the outliers are shown in different colour or as black dots.	This algorithm still does not show outlier analysis clearly but outlier groups are formed as we can see at the bottom.

Efficiency of Algorithm	The analysis seems like its segregating the data based on tenth standard marks and hardly taking into account the CGPA.	It cannot form clusters of a variety of shapes and hence patterns of certain shape may not be analysed.	We can choose the level which looks optimal hence using this algorithm we can see at which level suitable, appropriate clusters are being formed.	The density of nodes inside the cluster will be more than the density outside. This can be used for dense areas of a variety of shapes like donut shape etc.	However this algorithm is unstable and can give varied results based on the dataset and sometimes ellipses can overlap to produce awkward results.
Further Analysis	This type of technique won't be helpful to form favourable clusters in our data as we are looking for students who may need further attention or who stand as a benchmark for not admitting further students like them.	This algorithm produces different results depending on the fuzziness and comes close to K-means when probability is zero and produces varied results as probability decreases.	The nodes that join the tree at the higher levels indicate the outliers. This method is helpful to pinpoint and extract outliers for further processing.	This technique is fast when the dataset is small but as the dataset increases the time taken for this algorithm to run increases tremendously.	Instead of forming rigid linear boundaries in uses ellipses to divide the data. The groups formed divide the data based on both CGPA and tenth unlike fuzzy and K-means. We can see a well formed groups and we can definitely use this algorithm for our purpose.

CHAPTER 8

PREDICTIVE ANALYSIS

We have experimented on predictive analysis of data and tried out various algorithms on actual data of our college students. We used correlation analysis to see the relationship between variables like CGPA, tenth and twelfth. We also tried to use regression to predict the potential GPA of students and the results are shown in the graphs and charts below.

Table 4. Correlation and Regression summary report

	10TH	*12TH*	*CGPA*
10TH	1		
12TH	0.446498	1	
CGPA	0.426577	0.5407	1

Correlation analysis between tenth, twelfth and CGPA.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.609473
R Square	0.371458
Adjusted R	0.360234
Standard E	0.245849
Observatic	115

Regression Statistics of our student data analysis.

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regressior	2	4.00063	2.000315	33.09505	5.09E-12
Residual	112	6.769451	0.060442		
Total	114	10.77008			

Anova Statistics of data analysis of CGPA, Tenth and Twelfth.

	Coefficients	*tandard Erı*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*
Intercept	0.104926	0.054402	1.928715	0.056297	-0.00286	0.212717	-0.00286	0.212717
10th_NOR	0.300655	0.091056	3.301873	0.001289	0.120239	0.48107	0.120239	0.48107
12th_NOR	0.449552	0.084863	5.297382	5.94E-07	0.281406	0.617697	0.281406	0.617697

Table 5. Tenth, Twelfth and CGPA Percentile by normalized distribution

10th_NOR	12th_NOR	CGPA_NOR
87%	71%	44%
88%	84%	71%
53%	12%	19%
42%	58%	49%
91%	98%	96%
81%	80%	84%
31%	68%	90%
69%	94%	78%
76%	95%	71%
60%	90%	99%
71%	39%	21%
23%	66%	31%
68%	95%	97%
91%	56%	55%
82%	68%	9%
88%	74%	88%
55%	16%	75%
82%	92%	87%
64%	73%	74%
58%	86%	70%
46%	97%	45%
91%	84%	93%
78%	81%	67%
70%	91%	88%
83%	85%	82%
84%	59%	94%
48%	43%	29%
81%	92%	81%
53%	37%	6%
85%	50%	29%

It represents a Normalized distribution of percentile of our computer batch to get a better idea of the relative standing of individuals in order to implement predictive analysis.

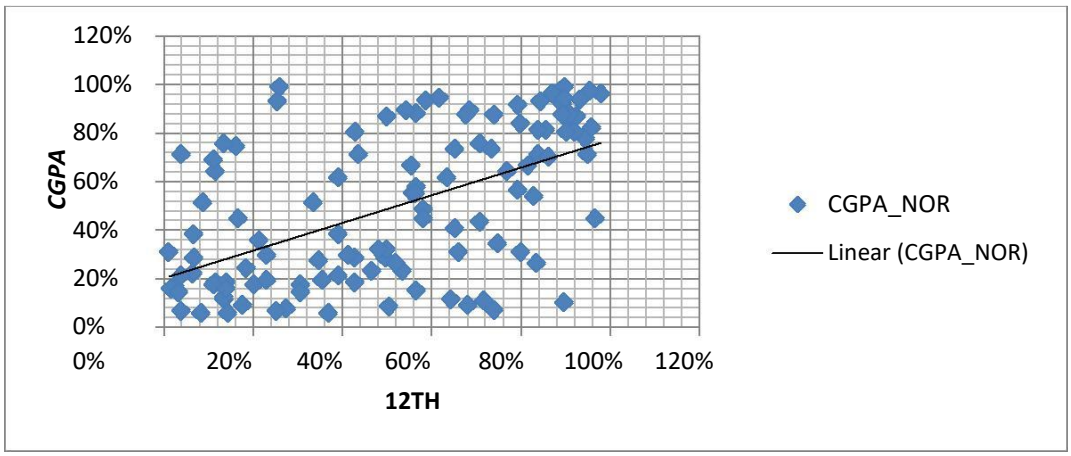


Figure8.1. Twelfth vs CGPA: basis for predictive analysis

This figure shows the trend based on correlation and aids us in the prediction of data. This forms the bases of predicting the data for CGPA using 12^{th} standard marks.

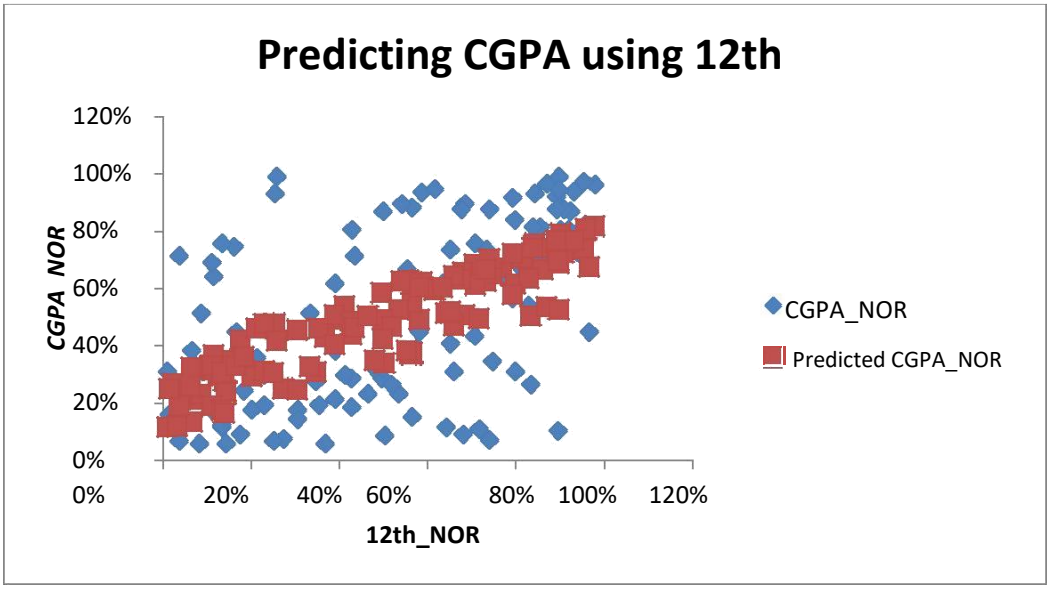


Figure8.2. Predicting CGPA using 12th

This graph shows the prediction of CGPA based on tenth. This prediction is more accurate than tenth as the correlation between CGPA and tenth was higher.

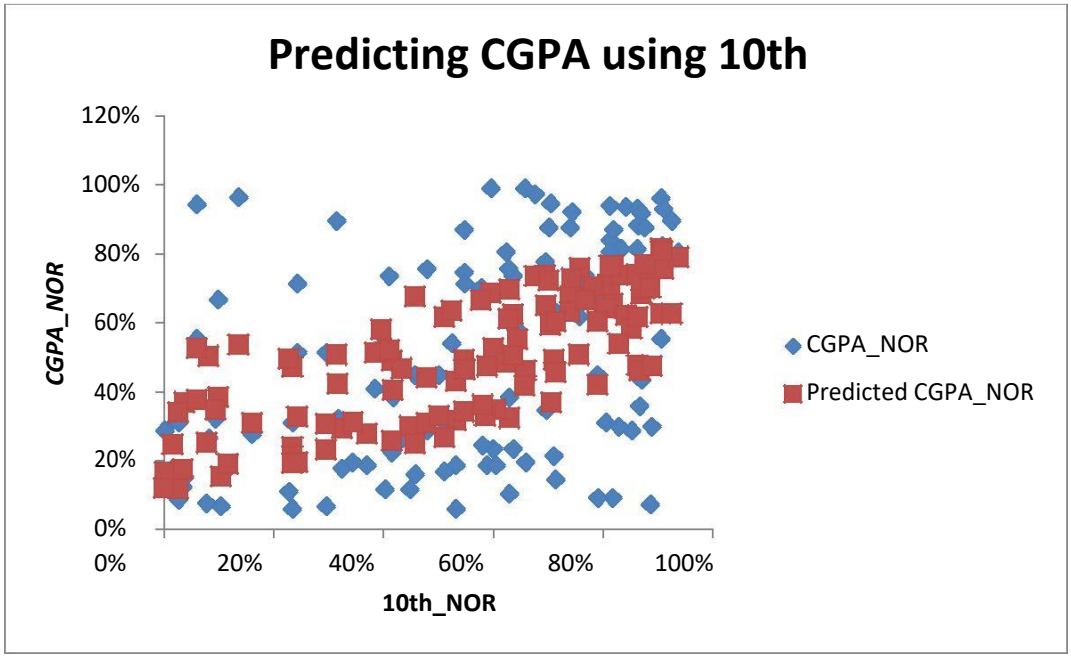


Figure8.3. Predicting CGPA using 10th

The prediction of CGPA via 12th standard marks is more accurate compared to 10th standard marks as the correlation is higher between CGPA and 12th than CGPA and 10th.

The second phase of our project includes the prediction of the increment in college enrolment over the years depending on multiple independent variables. Beyond this we have performed an extensive predictive analysis to try and find out variables that are highly correlated to get better, more accurate predictions. Using this we have discovered an extensive co-relation relation between the student's CGPA over the three years with his/her aptitude mock test marks and the final placement package received at the end of four years.

The main aim in doing is to help the institute make better decisions regarding student acceptance and help students personally analyse and improve performance to prepare for better companies.

The techniques used are:

- Linear model functions (LM)
- Data frames
- Shiny ui.r (user interface)
- Shiny Server.r (for server files)

CHAPTER 9

REGRESSION ANALYSIS

Regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for **modelling and analysing** several variables, when the focus is on the relationship between a dependent variable and one or more independent variables

Linear Model Function:

- **The lm() function**
- In R, the lm(), or "linear model," function can be used to create a multiple regression model. The lm() function accepts a number of arguments ("Fitting Linear Models(called objects)“, Data). The following list explains the two most commonly used parameters.
    - **formula: describes the model**
        - Note that the formula argument follows a specific format. For multiple linear regression, this is "YVAR ~ XVAR1 + XVAR2 + … + XVARi" where YVAR is the dependent, or predicted, variable and XVAR1, XVAR2, etc. are the independent, or predictor, variables.
    - **data: the variable that contains the dataset**

9. 1 Linear Regression Analysis

A simple linear regression model that describes the relationship between two variables x and y can be expressed by the following equation. The numbers α and β are called parameters, and ϵ is the error term.

Estimated Simple Regression Equation

$$y = \alpha + \beta x + \epsilon$$

In linear regression, data are modelled using linear predictor functions, and unknown model parameters are estimated from the data. Such models are called linear models. Most commonly, linear regression refers to a model in which the conditional mean of y given the value of X is an affine function of X. Less commonly, linear regression could refer to a model in which the median, or some other quantile of the conditional distribution of y given X is expressed as a linear function of X. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of y given X, rather than on the joint probability distribution of y and X, which is the domain of multivariate analysis.

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications. This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

Linear regression has many practical uses. Most applications fall into one of the following two broad categories:

- If the goal is prediction, or forecasting, or reduction, linear regression can be used to fit a predictive model to an observed data set of y and X values. After developing such a model, if an additional value of X is then given without its accompanying value of y, the fitted model can be used to make a prediction of the value of y.
- Given a variable y and a number of variables X1, ..., Xp that may be related to y, linear regression analysis can be applied to quantify the strength of the relationship between y and the Xj, to assess which Xj may have no relationship with y at all, and to identify which subsets of the Xj contain redundant information about y.

For example, in the data set faithful, it contains sample data of two random variables named waiting and eruptions. The waiting variable denotes the waiting time until the next eruptions, and eruptions denote the duration. Its linear regression model can be expressed as:

$$\text{Eruptions} = \alpha + \beta * \text{Waiting} + \epsilon$$

	A	B	C	D	E
1	NAME	CGPA	avgmarks	SALARY	percen
2	Desirae	3.08	33.6	409009	71
3	Jaquelyn	2.57	27.2	300108	75
4	Isaac	3.34	28.4	419778	69
5	Lynn	2.44	35.2	122166	70
6	Knox	3.37	28.8	509774	74
7	Carolyn	3.73	23.6	663607	86
8	Kai	3.42	33.4	485055	95
9	Ivy	2.66	30.2	300626	67
10	Patrick	3.6	24.2	717715	73
11	Martena	2.96	27.6	326461	68
12	Kirby	3.23	30.2	450332	81
13	Claire	2.53	22.8	310653	66
14	Jessica	2.68	26.8	281118	73
15	Kelly	3.47	33.2	467931	80
16	Pandora	3.75	28.8	713250	78
17	Caesar	2.42	24	252877	75
18	Sacha	3.63	30	777578	68
19	Denise	2.81	35.4	394281	90
20	Carolyn	4	26	838469	69
21	Felix	3.76	26.8	800971	75
22	Trevor	3.54	31.2	610440	72
23	Herman	3.27	39.4	543655	80

Fig 9.1.1 The Database

The Database containing one dependent variable and many independent variables

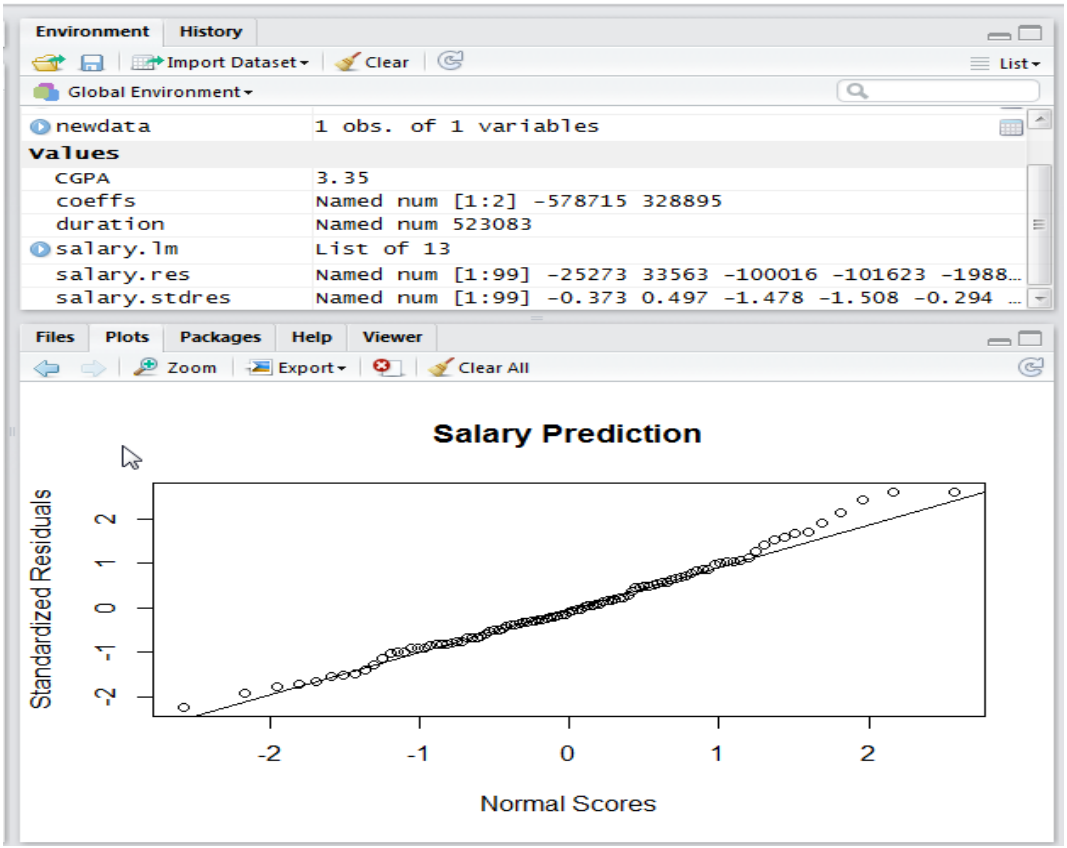


Fig 9.1.2 A plot

A graphical plot of the prediction (the snake shows the trend)

predictions		
NAME	CGPA	Predicted salary
Siddhant	2.5	243522.4
Devansh	2.98	401392
sushant	3.2	473748.9
Anuj	3.5	572417.4
John	3.9	703975.4
viraj	2.2	144853.9

Fig 9.1.3 (a) Predicted Values

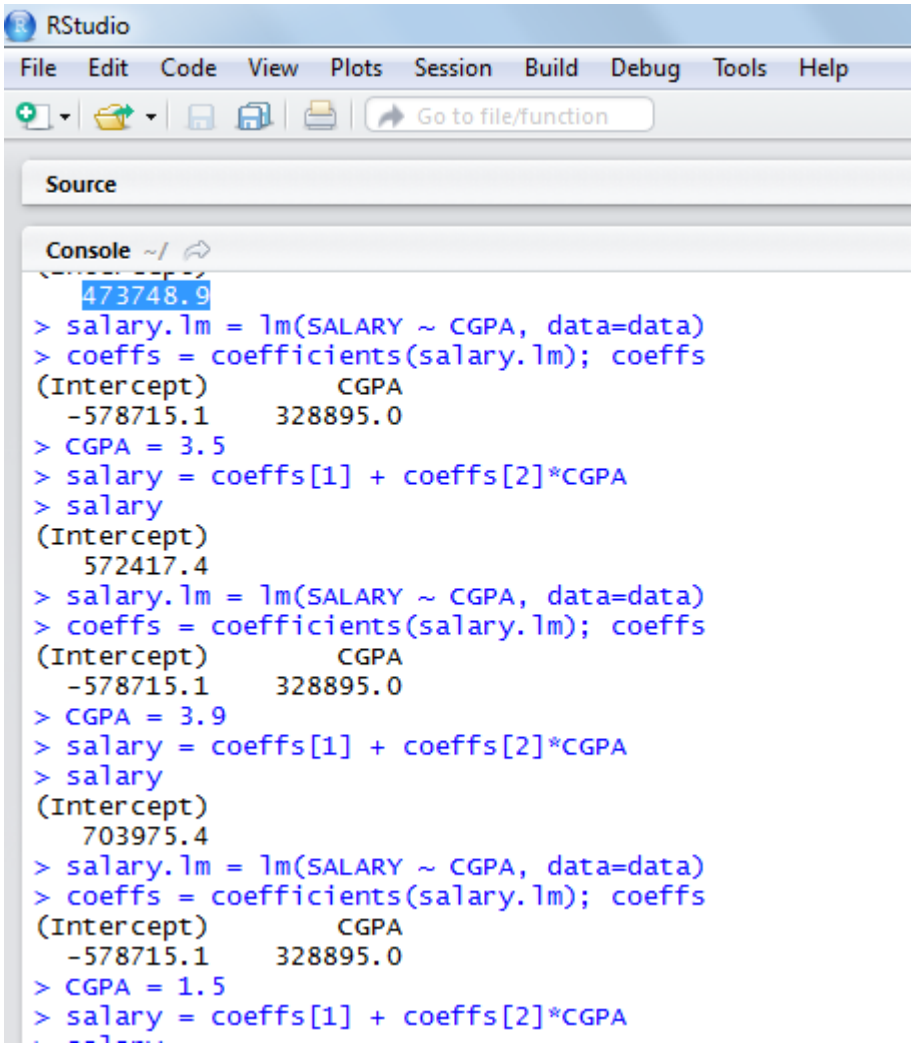


Fig 9.1.3(b) Predicted values

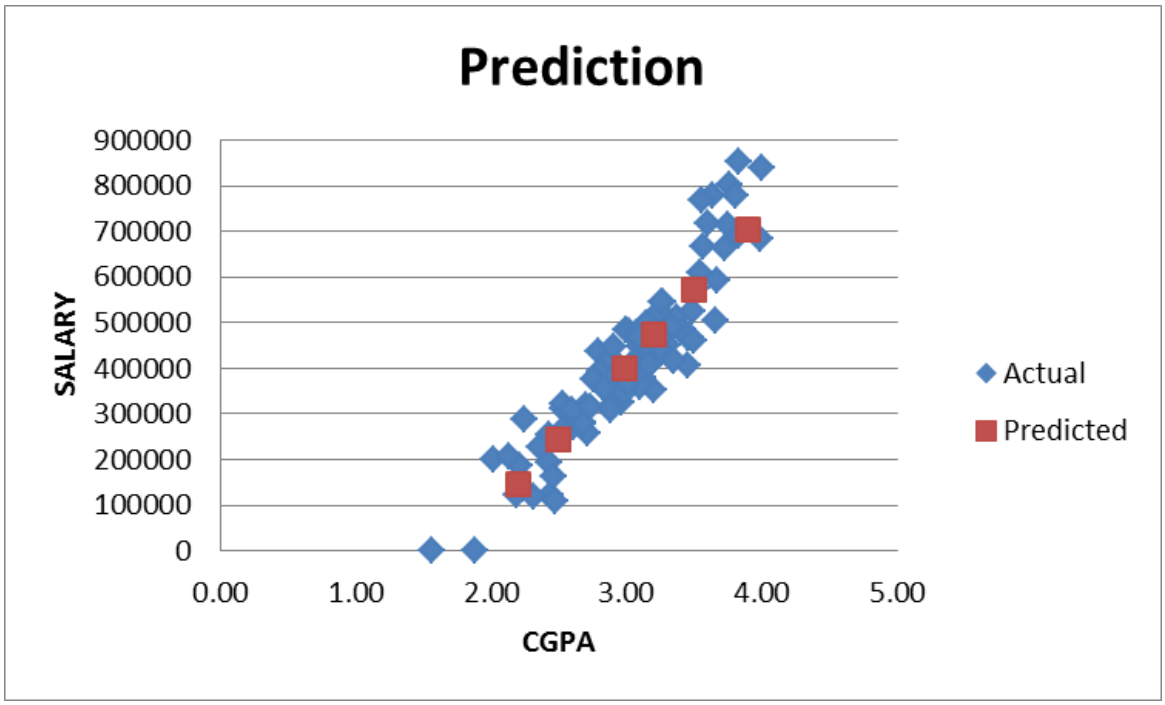


Fig 9.1.4 The graph

A graphical Plot showing the predicted values across the CGPA and Salary axis

9. 2. Multiple Regression Analysis

Multiple regression analysis is a powerful technique used for predicting the unknown value of a variable from the known value of two or more variables- also called the predictors.

More precisely, multiple regression analysis helps us to predict the value of Y for given values of X1, X2, …, Xk.

By multiple regression, we mean models with just one dependent and two or more independent (exploratory) variables. The variable whose value is to be predicted is known as the dependent variable and the ones whose known values are used for prediction are known independent (exploratory) variables.
Multiple regression technique does not test whether data are linear. On the contrary, it proceeds by assuming that the relationship between the Y and each of Xi's is linear. Hence as a rule, it is prudent to always look at the scatter plots of (Y, Xi), i= 1, 2,…,k.

The general purpose of multiple regression is to learn more about the relationship between several independent or predictor variables and a dependent or criterion variable. For example, a real estate agent might record for each listing the size of the house (in square feet), the number of bedrooms, the average income in the respective neighbourhood according to census data, and a subjective rating of appeal of the house. Once this information has been compiled for various houses it would be interesting to see whether and how these measures relate to the price for which a house is sold. For example, you might learn that the number of bedrooms is a better predictor of the price for which a house sells in a particular neighbourhood than how "pretty" the house is (subjective rating). You may also detect "outliers," that is, houses that should really sell for more, given their location and characteristics.

THE REGRESSION EQUATION
Represented as $y = a + \sum bx$

A line in a two dimensional or two-variable space is defined by the equation Y=a+b*X; in full text: the Y variable can be expressed in terms of a constant (a) and a slope (b) times the X variable. The constant is also referred to as the intercept, and the slope as the regression coefficient or B coefficient. For example, GPA may best be predicted as 1+.02*IQ. Thus, knowing that a student has an IQ of 130 would lead us to predict that her GPA would be 3.6

The Process:

- create a linear model using lm(FORMULA, DATAVAR)
- predict the fall enrolment (ROLL) using the unemployment rate (UNEM) and number of spring high school graduates (HGRAD)

- twoPredictorModel <- lm(ROLL ~ UNEM + HGRAD, datavar)
- display model: twoPredictorModel
- **For enrolment case 2 variables:**
- > #what is the expected fall enrollment (ROLL) given this year's unemployment rate (UNEM) of 9% and spring high school graduating class (HGRAD) of 100,000
- > -8255.8 + 698.2 * 9 + 0.9 * 100000
- [1] 88028
- > #the predicted fall enrollment, given a 9% unemployment rate and 100,000 student spring high school graduating class, is 88,028 students

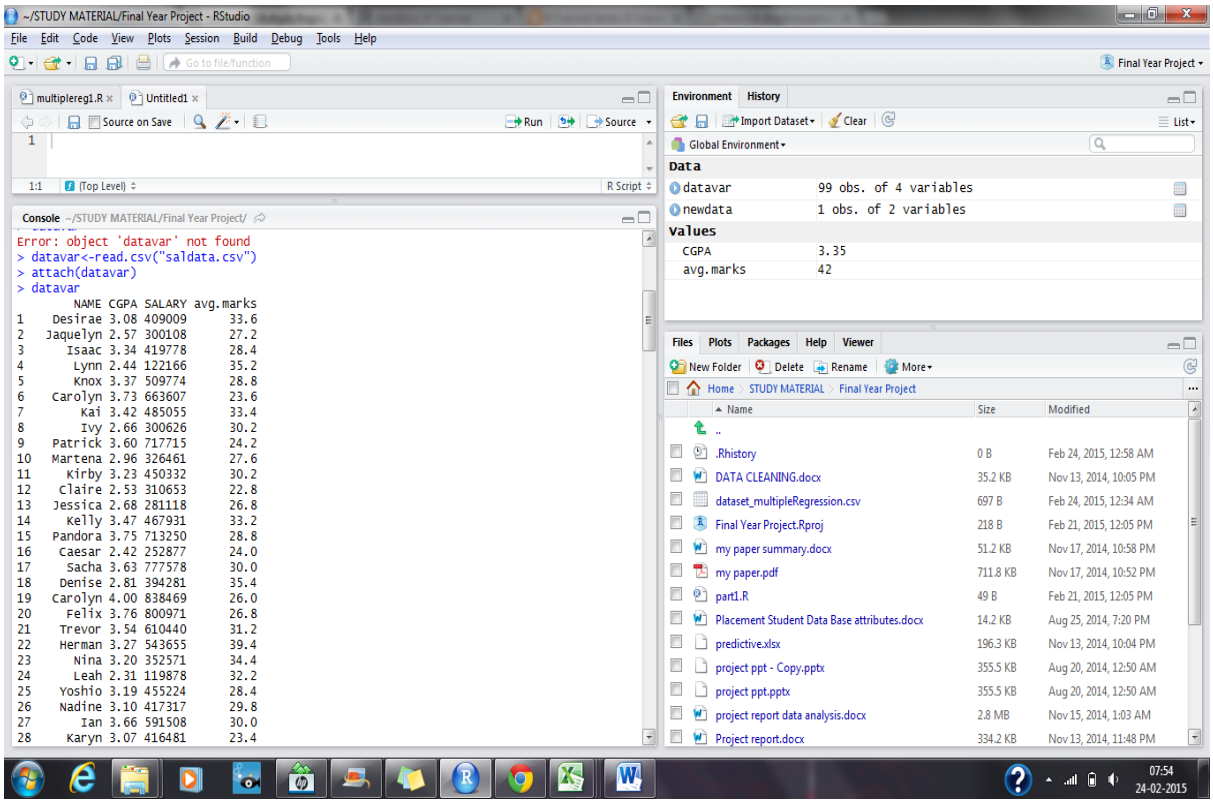
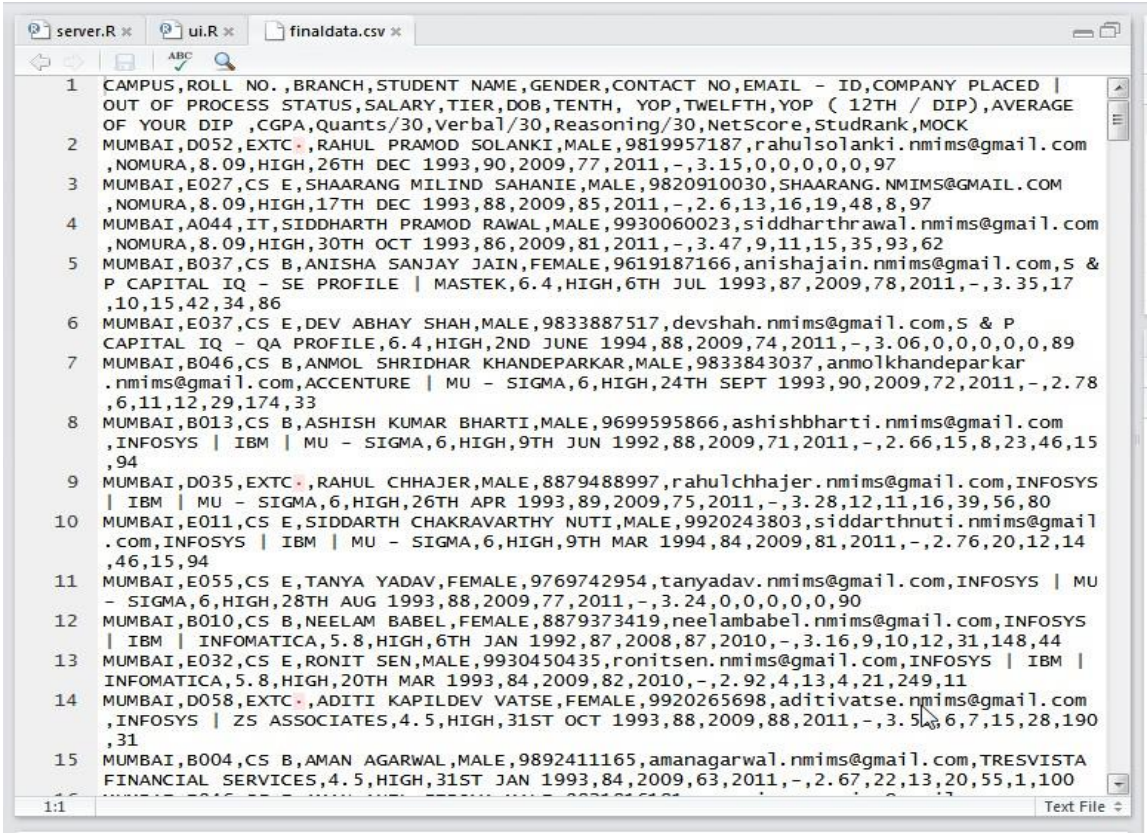


Fig 9.2.1 The Database

The call of the database to the R studio for analyzing



9.2.2 Database read in a .csv format

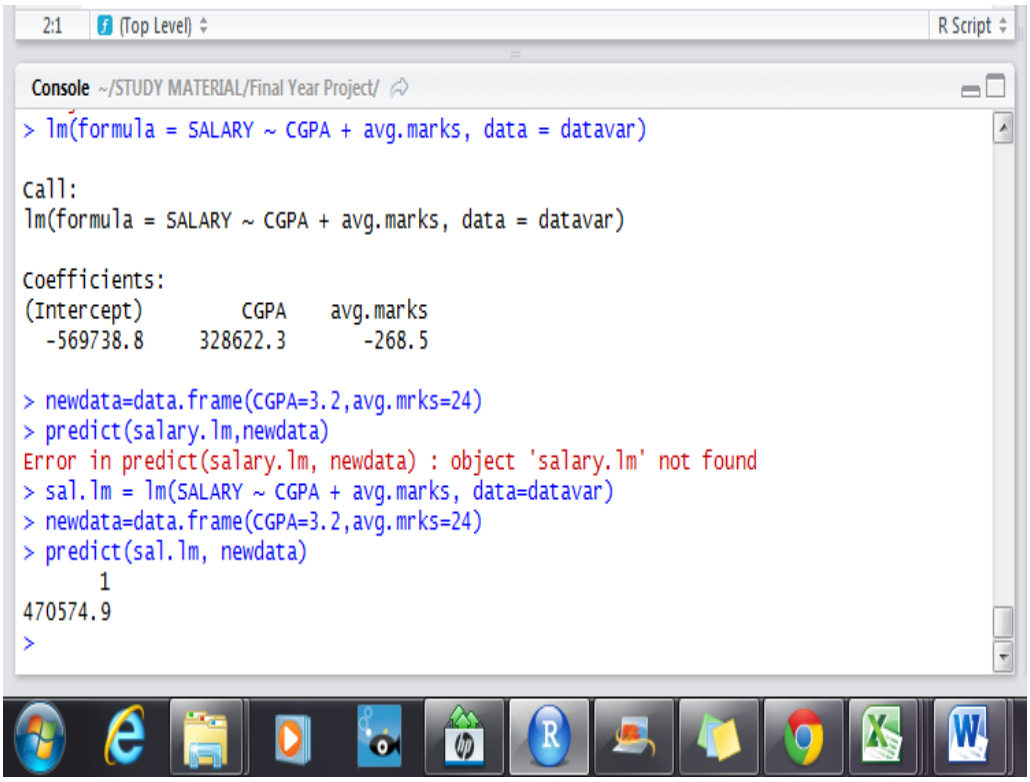


Fig 9.2.3 The Intercepts and co-efficient

The Intercept and co-efficient calculations for a two variable dependence

```
Console ~/STUDY MATERIAL/Final Year Project/
       1
470574.9
> twoPredictorModel <- lm(SALARY ~ CGPA + avg.marks, datavar)
> twoPredictorModel

Call:
lm(formula = SALARY ~ CGPA + avg.marks, data = datavar)

Coefficients:
(Intercept)         CGPA    avg.marks
  -569738.8     328622.3       -268.5

> summary(twoPredictorModel)

Call:
lm(formula = SALARY ~ CGPA + avg.marks, data = datavar)

Residuals:
    Min      1Q  Median      3Q     Max
-154410  -47836   -7311   41119  175890

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -569738.8    61117.0  -9.322 4.26e-15 ***
CGPA         328622.3    14202.1  23.139  < 2e-16 ***
avg.marks      -268.5     1297.1  -0.207    0.836
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 68530 on 96 degrees of freedom
Multiple R-squared:  0.8493, Adjusted R-squared:  0.8462
F-statistic: 270.5 on 2 and 96 DF,  p-value: < 2.2e-16
```

Fig 9.2.4 The efficiency test

A Summary using two variables shows positive significance coeff indicating that there is a relation possible. The Pr(>|t|) values can be extracted to test if the data under investigation are auto correlated

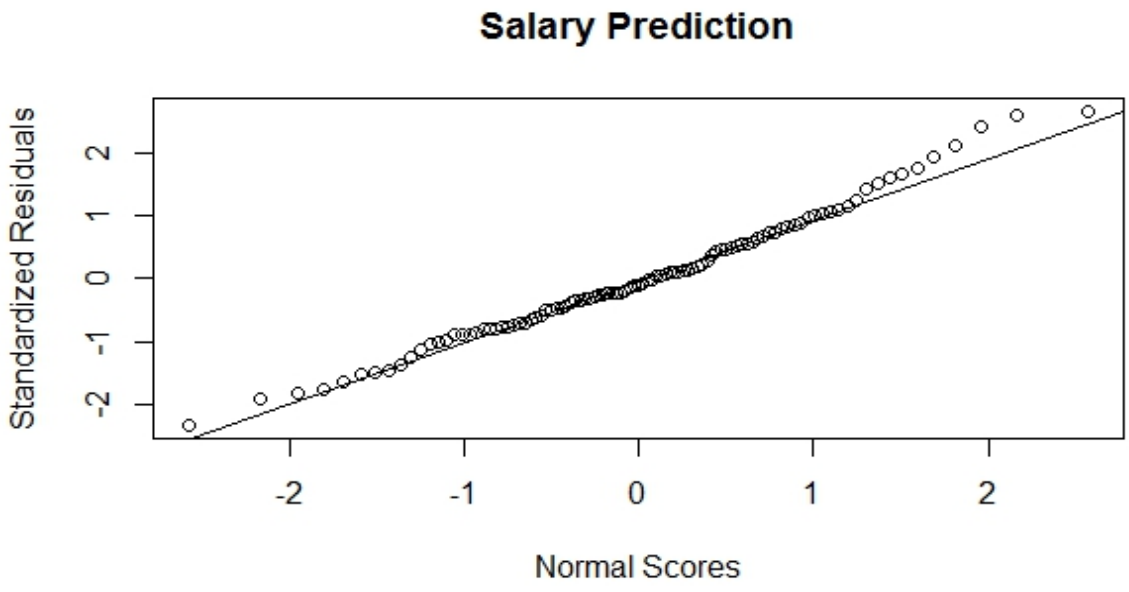


Fig 9.2.5 Comparison

Slightly better plots than the linear dependence

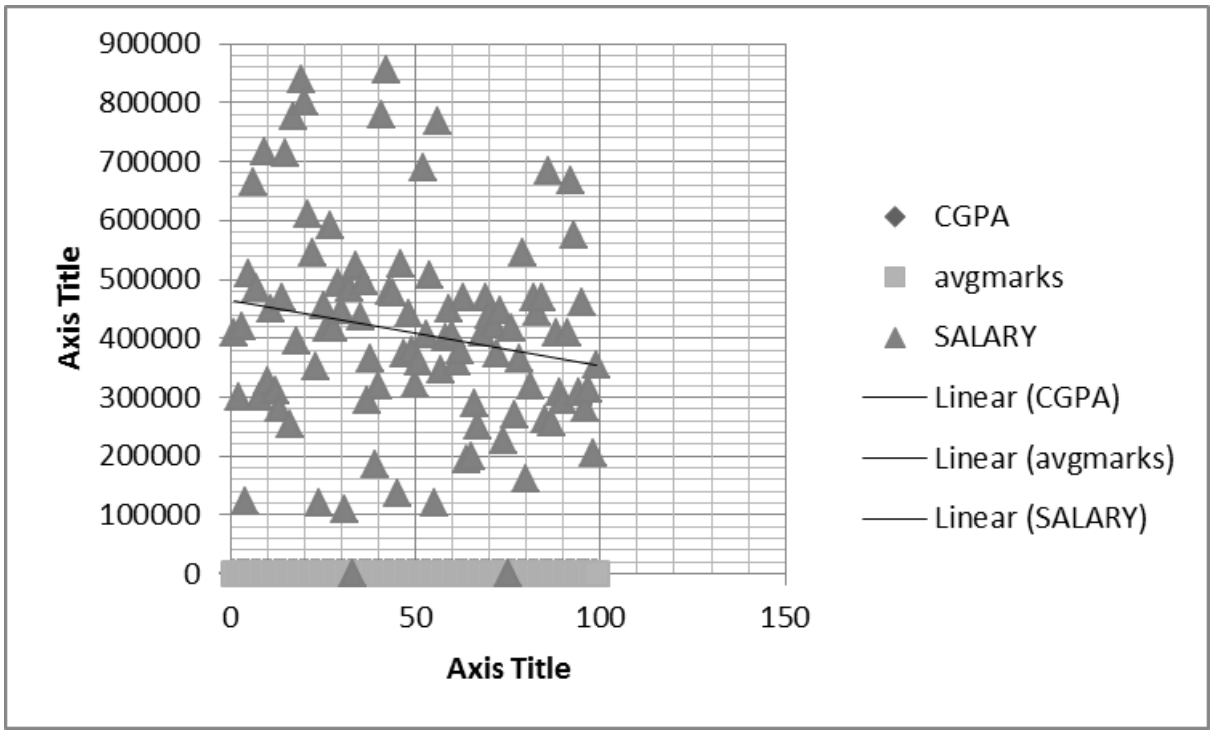


Fig 9.2.6 A linear plot in the dense areas

	Predicted Data			
Name	CGPA	avgmarks	percen	Salary
Anuj Chop	2.9	32	79	374846.9
Devansh	3.2	29	91	460499.5
Sid	3.4	41	75.3	540893.4
Sushant	3.14	39	87	441284.7

Fig 9.2.7 (a) Predicted Values

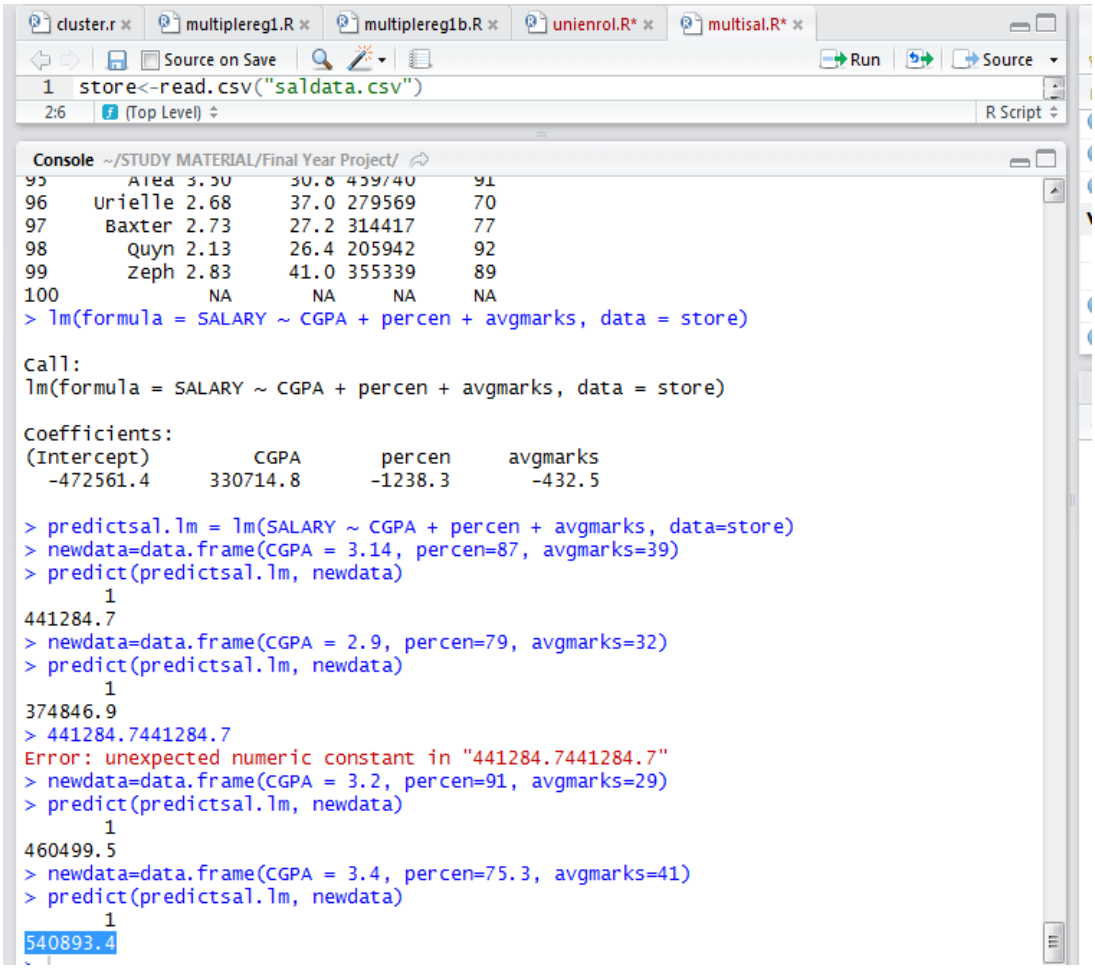


Fig 9.2.7 (b) Example of Predicted values for any entry possible

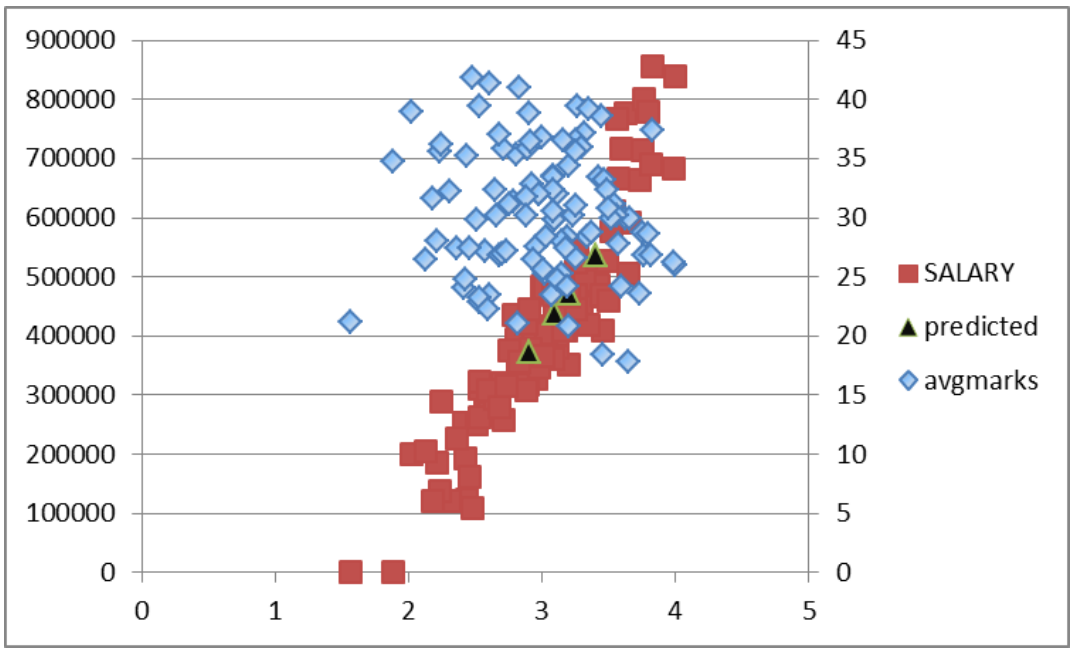


Fig 9.2.8 Graph of the dependent values with axis representing the independent variables respectively

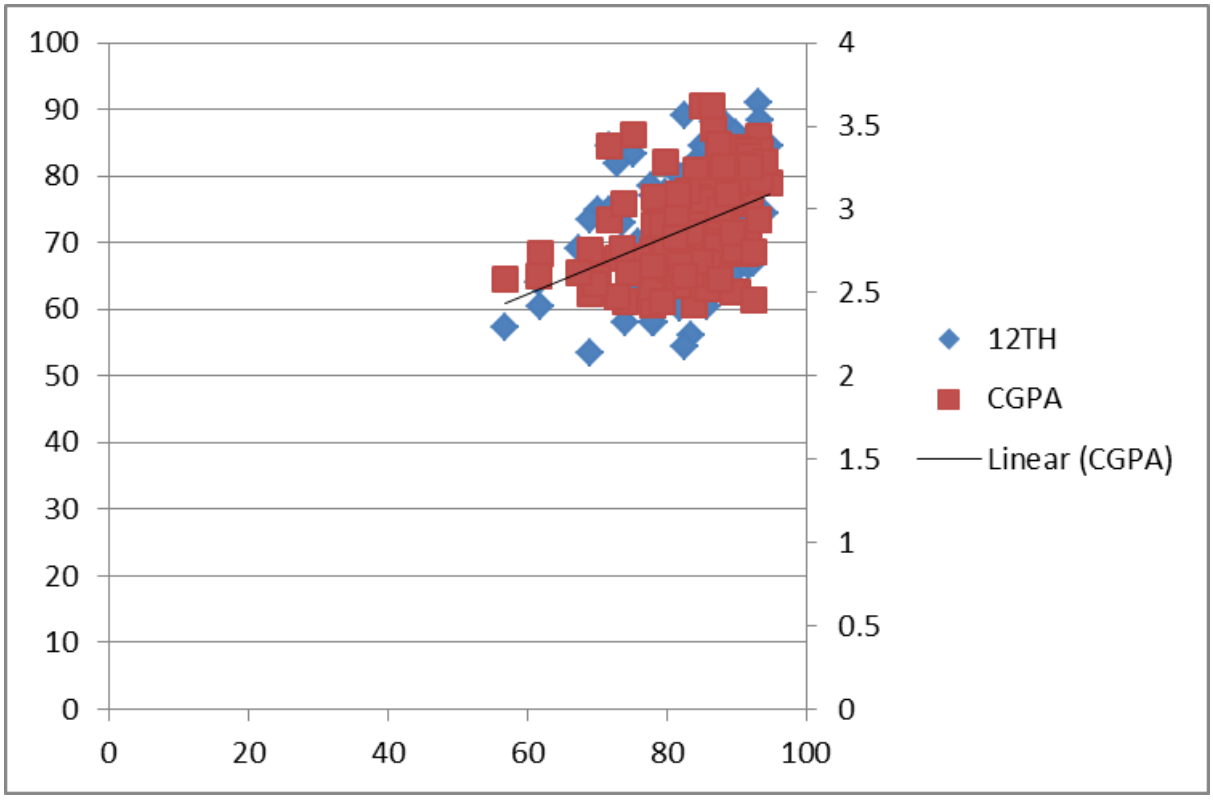


Fig 9.2.9 CGPA vs 12th for multiple regression

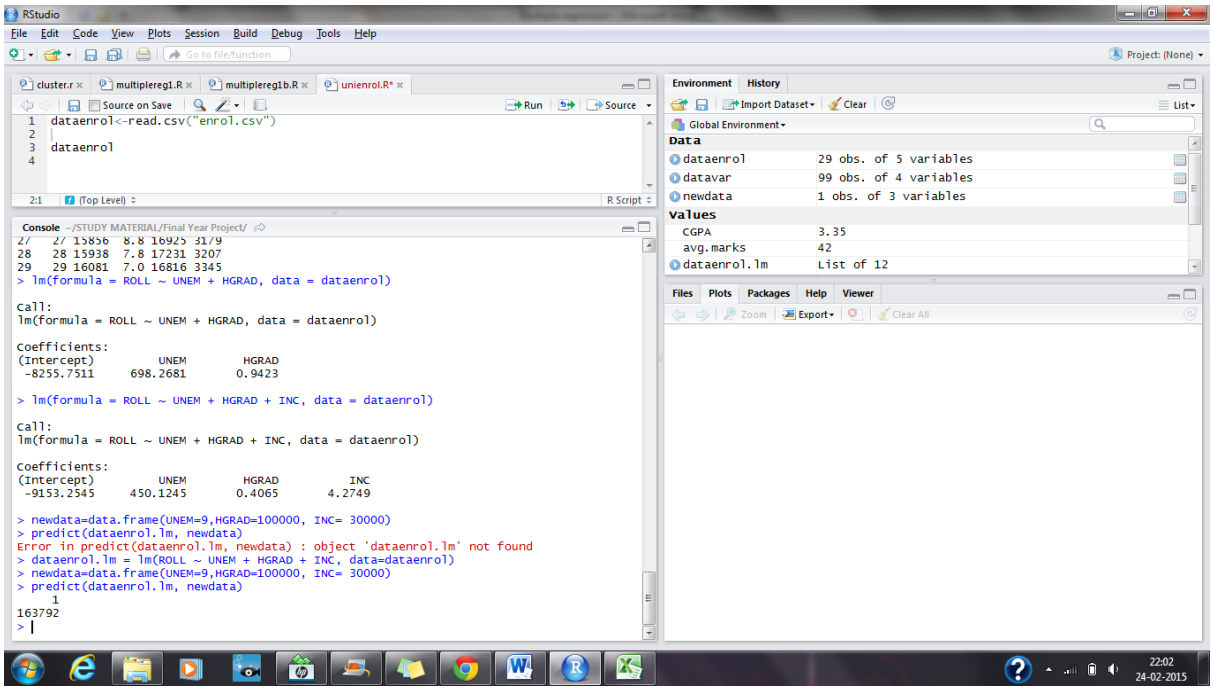


Fig 9.2.10 Another implementation for university student enrolment

9. 3. Logistic Regression Analysis

Logistic regression is a direct probability model that measures the relationship between the categorical dependent variable and one or more independent variables, which are usually (but not necessarily) continuous, by estimating probabilities. Thus, it treats the same set of problems as does prohibit regression using similar techniques; the first assumes a logistic function and the second a standard normal distribution function.

Logistic regression can be seen as a special case of generalized linear model and thus analogous to linear regression. The model of logistic regression, however, is based on quite different assumptions (about the relationship between dependent and independent variables) from those of linear regression. In particular the key differences of these two models can be seen in the following two features of logistic regression. First, the conditional distribution p(y | x) is a Bernoulli distribution rather than a Gaussian distribution, because the dependent variable is binary. Second, the estimated probabilities are restricted to [0, 1] through the logistic distribution function because logistic regression predicts the probability of the instance being positive.

Logistic regression is used widely in many fields, including the medical and social sciences. Logistic regression may be used to predict whether a patient has a given disease (e.g. diabetes; coronary heart disease), based on observed characteristics of the patient (age, sex, body mass index, results of various blood tests, etc.; age, blood cholesterol level, systolic blood pressure, relative weight, blood haemoglobin level, smoking (at 3 levels), and abnormal electrocardiogram.)

Logistic regression can be binomial or multinomial. Binomial or binary logistic regression deals with situations in which the observed outcome for a dependent variable can have only two possible types (for example, "dead" vs. "alive").

In binary logistic regression, the outcome is usually coded as "0" or "1", as this leads to the most straightforward interpretation. If a particular observed outcome for the dependent variable is the noteworthy possible outcome (referred to as a "success" or a "case") it is usually coded as "1" and the contrary outcome (referred to as a "failure" or a "non-case") as "0". Logistic regression is used to predict the odds of being a case based on the values of the

independent variables (predictors). The odds are defined as the probability that a particular outcome is a case divided by the probability that it is a non-case.

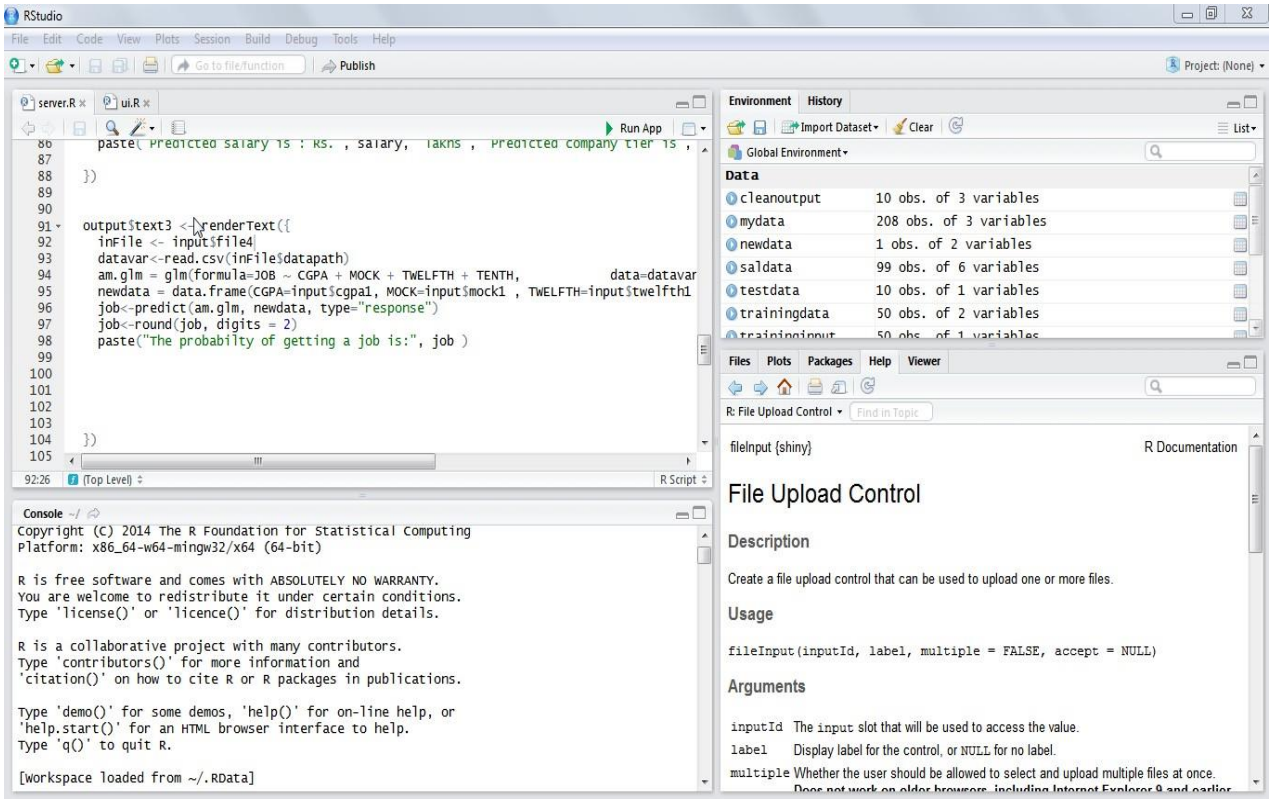


Fig 9.3.1 Probability function calculation code

CHAPTER 10

THE GRAPHICAL USER INTERFACE

Steps to interact with the application:

1. User opens the application by using the specified address or in R Studio.

2. User will be greeted with a welcome tab followed by a number of other tabs out of which he will select the appropriate tab to initiate the task to be completed

3. Before beginning the evaluation process the user will be prompted to upload his data that he wants to use for prediction analysis.

4. Clustering

- User selects the appropriate options from the drop down menus for the X and Y axis labels.
- User can also select the number of cluster he would like to view in the scatter plot
- The scatter plot will be displayed on the right hand corner of the applications screen

5. Linear regression

- The user selects the appropriate independent variable from the list of options made available.
- User can slide the slider input and select precisely the value he wants to be considered.
- The output will be displayed in the right hand side of the screen.
- The output will consist of the predicted salary and the tier of company into which a student might fall.
- The user will also be able to see the derived conclusions below the slider input.

6. Multiple Regression

- The user selects the appropriate independent variable from the list of options made available.
- In this case the user can select multiple variables to get more accurate prediction.
- User can slide the slider input and select precisely the value he wants to be considered.
- The output will be displayed in the right hand side of the screen.
- The output will consist of the predicted salary and the tier of company into which a student might fall.
- The user will also be able to see the derived conclusions below the slider input.

7. Logistic regression

- The user selects the appropriate independent variable from the list of options made available.

- In this case as well the user can select multiple variables to get more accurate prediction.
- User can slide the slider input and select precisely the value he wants to be considered.
- The output will be displayed in the right hand side of the screen.
- The output will consist of the probability of a student getting a job.
- The user will also be able to see the derived conclusions below the slider input.

Fig 10.1 R runtime environment facilitating the running of shiny application

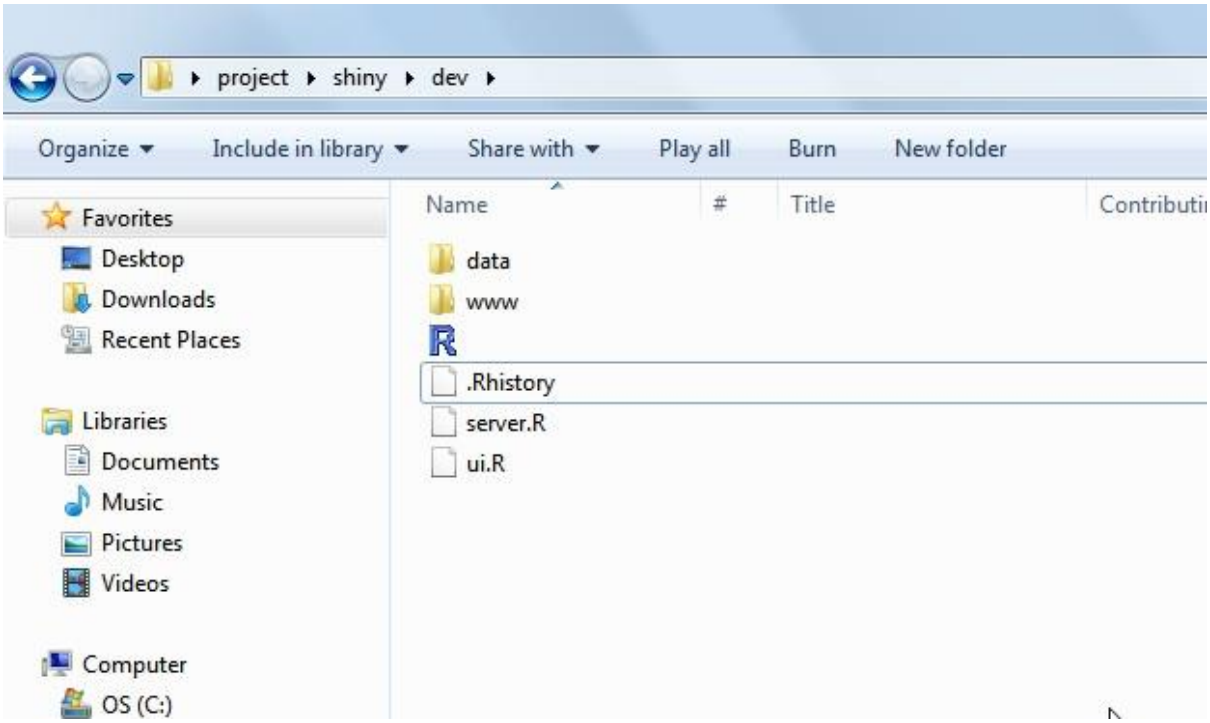


Fig 10.2 Server UI data www folders and files

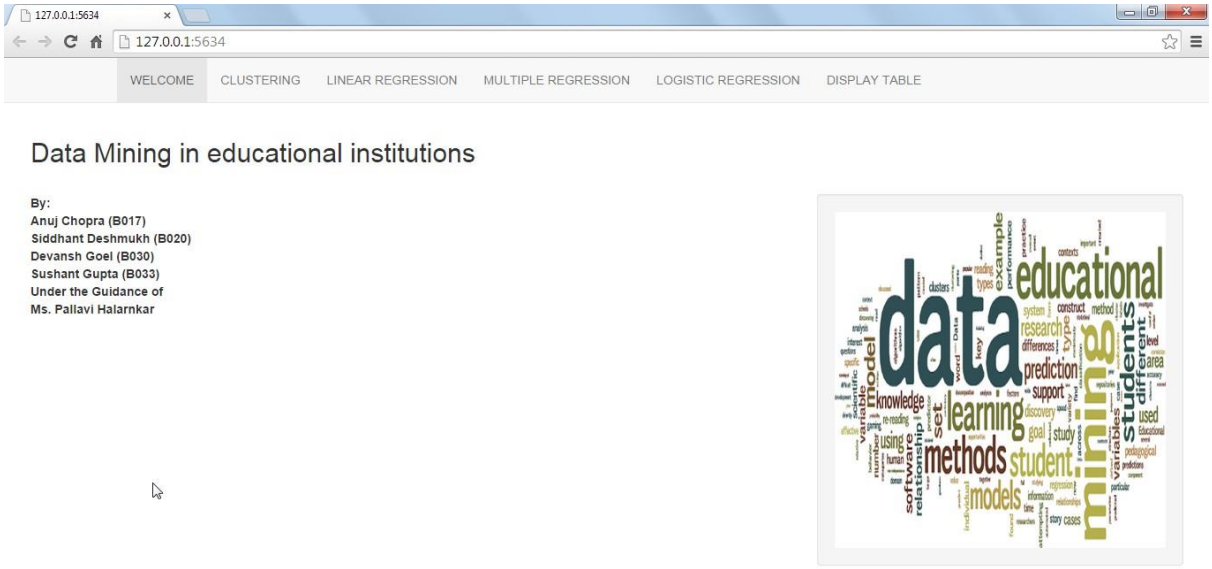


Fig 10.3 Home Screen showing multiple tabs

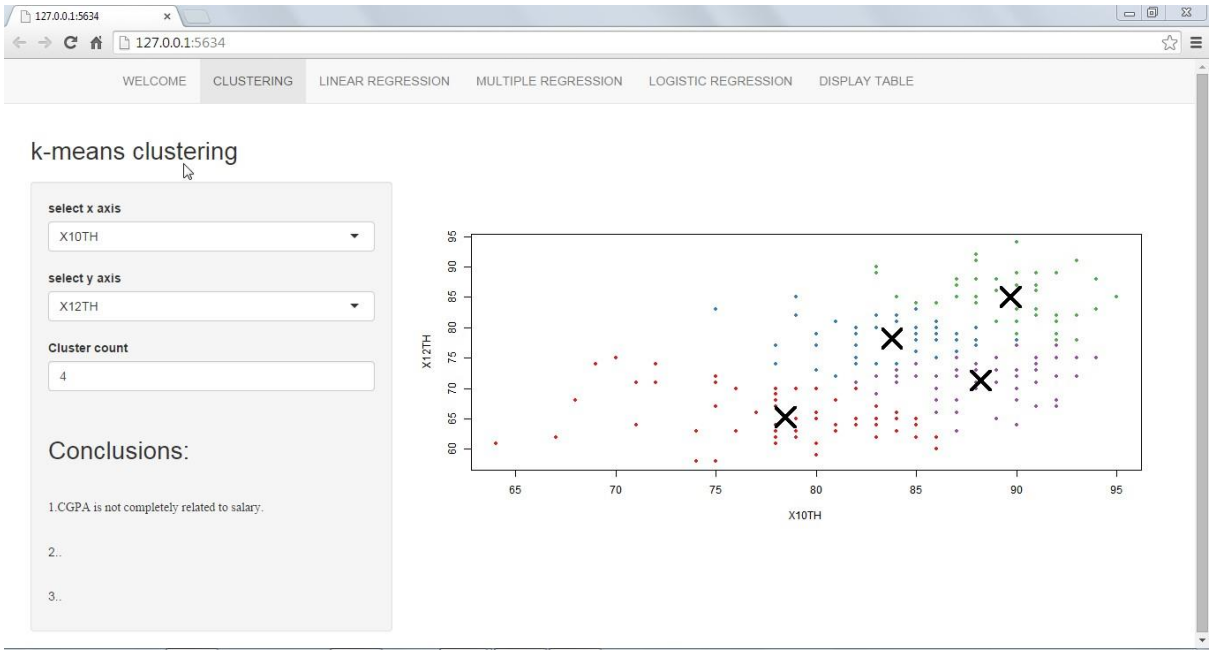


Fig 10.4 Clustering Tab

This tab option enables the user to choose the varied attributes to plot as per requirements for the analysis in the k-means clustering.

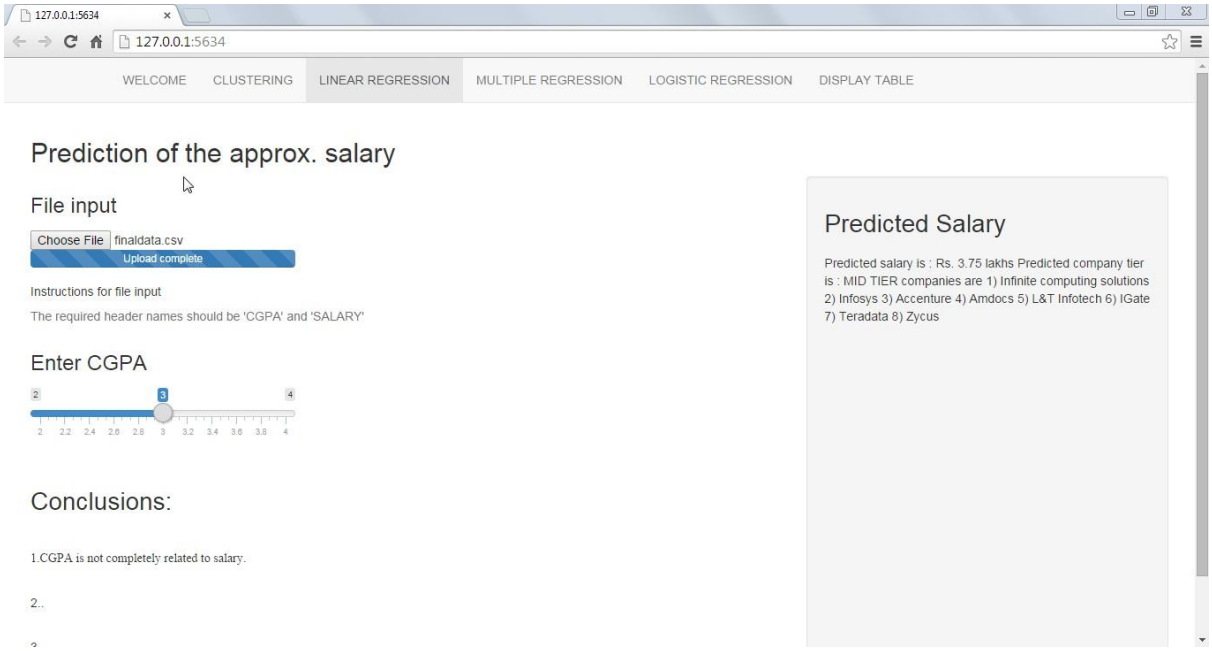


Fig 10.5 Linear Regression

The results are dynamic and get updated every time the slider for the CGPA is moved

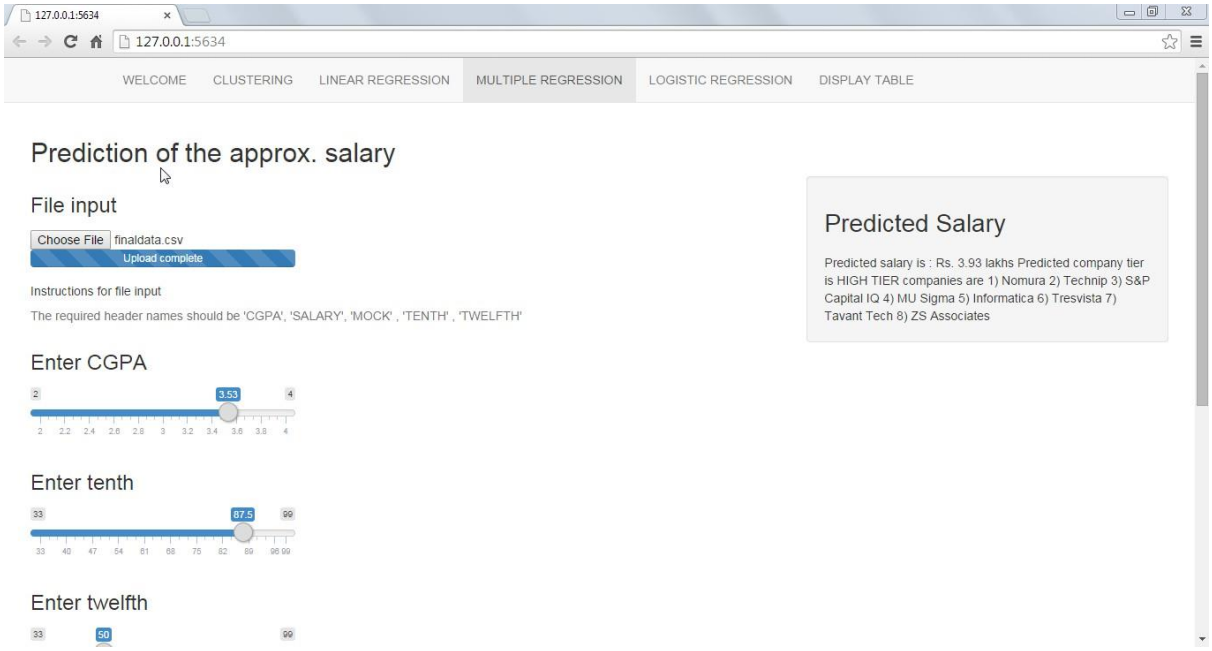


Fig 10.6 Multiple Regression Tab

The results are dynamic and get updated every time the slider for either of the CGPA, Tenth, twelfth, or mock test marks is moved.

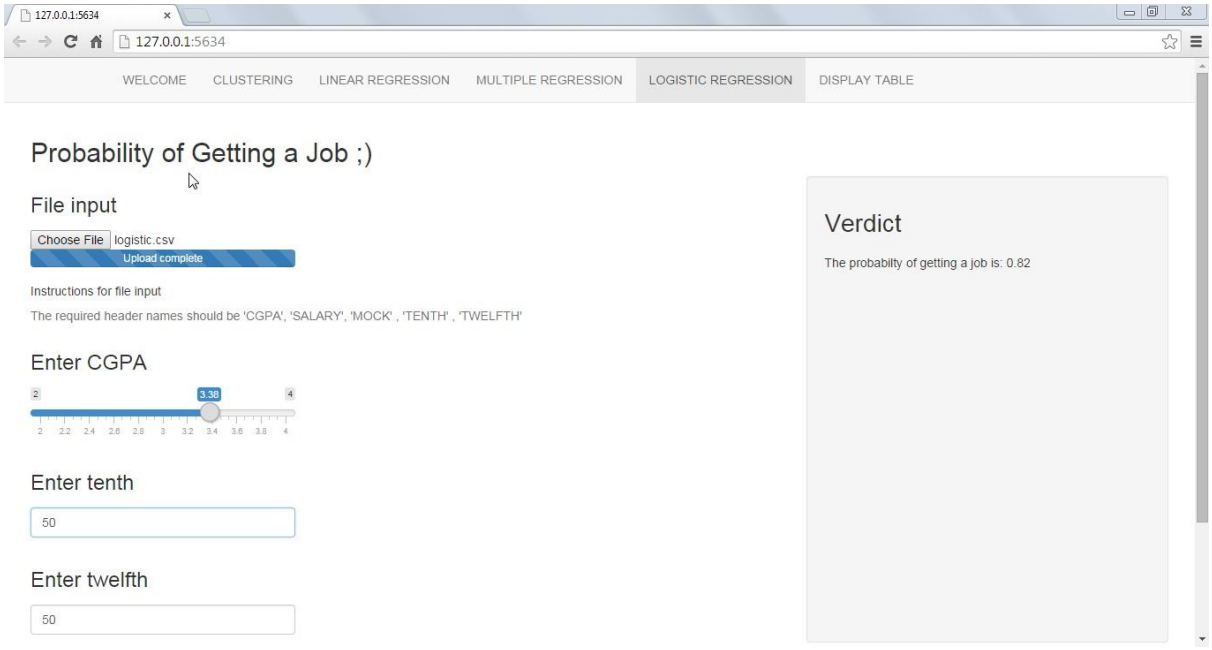
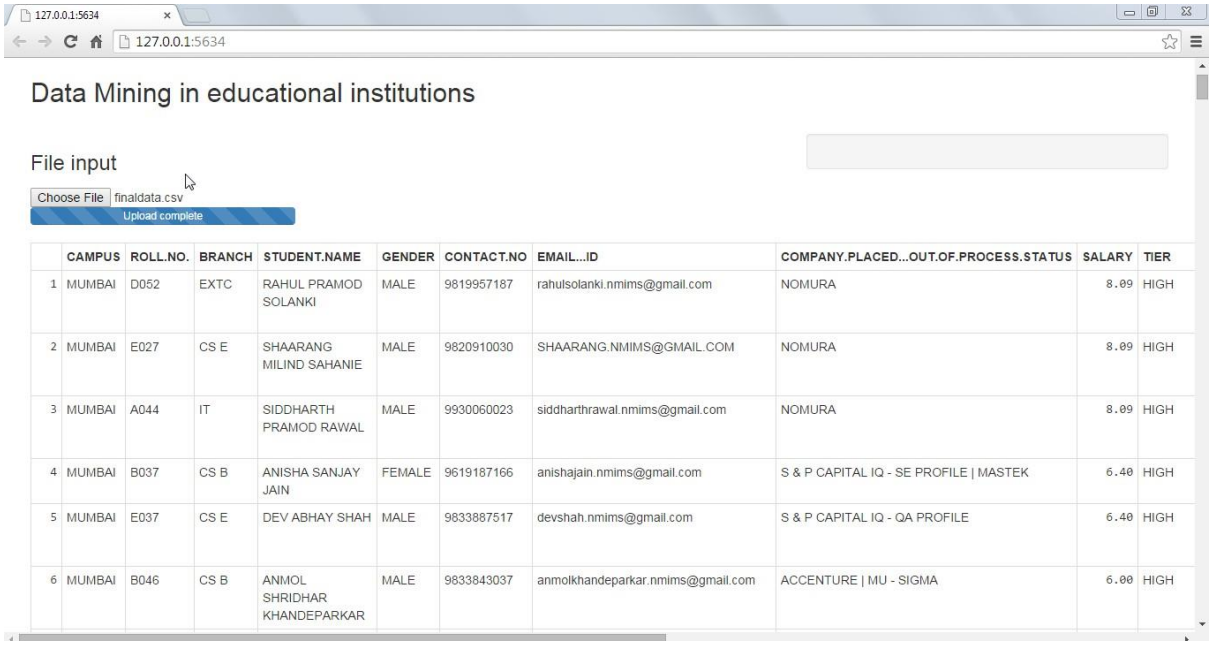


Fig 10.7 Logistic Regression Tab

The probability for attaining a job is giving in this tab. It logically answers a 'yes' or a 'no' and computes a probability using the nominal graph

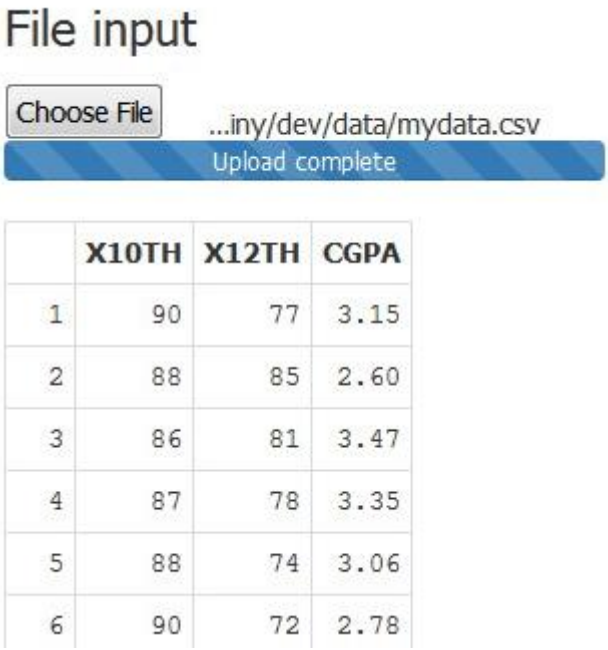


	CAMPUS	ROLL.NO.	BRANCH	STUDENT.NAME	GENDER	CONTACT.NO	EMAIL...ID	COMPANY.PLACED...OUT.OF.PROCESS.STATUS	SALARY	TIER
1	MUMBAI	D052	EXTC	RAHUL PRAMOD SOLANKI	MALE	9819957187	rahulsolanki.nmims@gmail.com	NOMURA	8.09	HIGH
2	MUMBAI	E027	CS E	SHAARANG MILIND SAHANIE	MALE	9820910030	SHAARANG.NMIMS@GMAIL.COM	NOMURA	8.09	HIGH
3	MUMBAI	A044	IT	SIDDHARTH PRAMOD RAWAL	MALE	9930060023	siddharthrawal.nmims@gmail.com	NOMURA	8.09	HIGH
4	MUMBAI	B037	CS B	ANISHA SANJAY JAIN	FEMALE	9619187166	anishajain.nmims@gmail.com	S & P CAPITAL IQ - SE PROFILE | MASTEK	6.40	HIGH
5	MUMBAI	E037	CS E	DEV ABHAY SHAH	MALE	9833887517	devshah.nmims@gmail.com	S & P CAPITAL IQ - QA PROFILE	6.40	HIGH
6	MUMBAI	B046	CS B	ANMOL SHRIDHAR KHANDEPARKAR	MALE	9833843037	anmolkhandeparkar.nmims@gmail.com	ACCENTURE | MU - SIGMA	6.00	HIGH

Fig 10.8 The Display Tab

This tab enables to user to view a tabular form of his .csv format database uploaded for analysis

Fig 10.9 Welcome Page



	X10TH	X12TH	CGPA
1	90	77	3.15
2	88	85	2.60
3	86	81	3.47
4	87	78	3.35
5	88	74	3.06
6	90	72	2.78

Fig 10.10 Any data can be generated dynamically

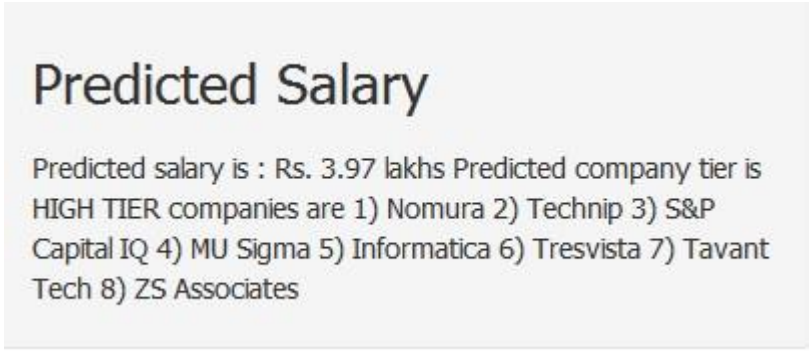


Fig 10.11 High Tier Companies

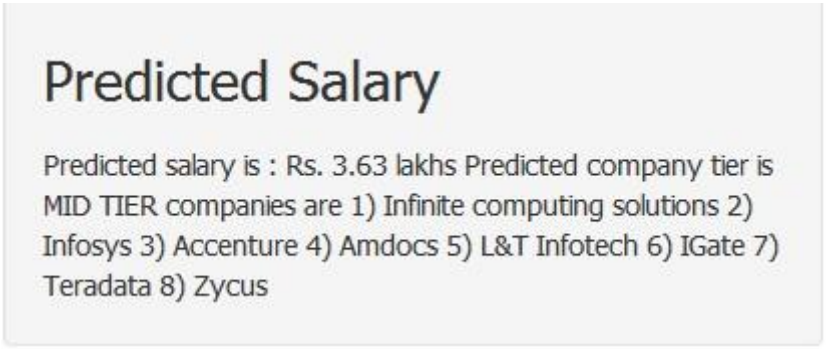


Fig 10.12 Mid-Tier companies

Fig 10.13 Low Tier Companies

Fig 10.14 Company tiers and list of predicted companies change dynamically

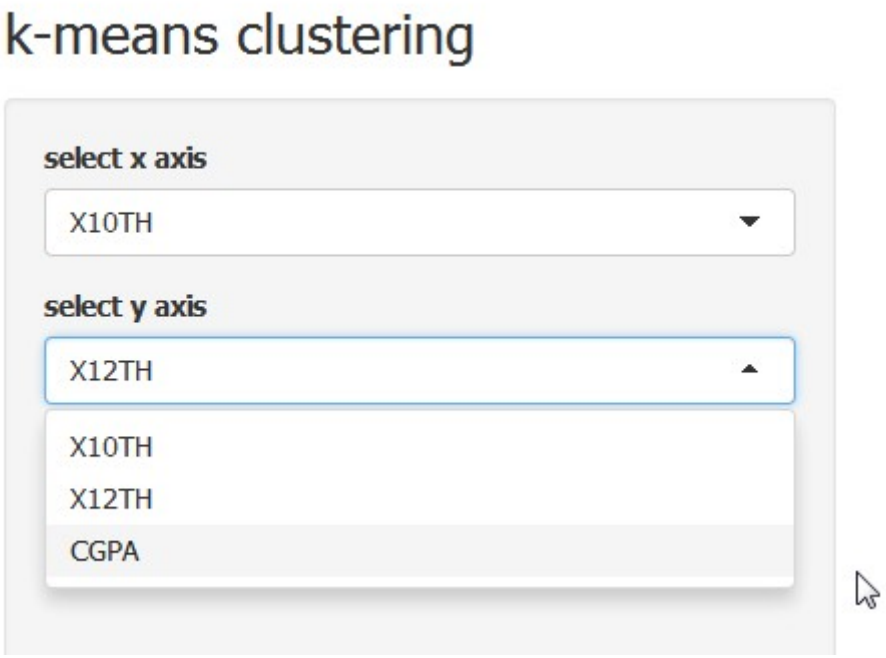


Fig 10.15 Drop down menu gives an option to change clusters dynamically

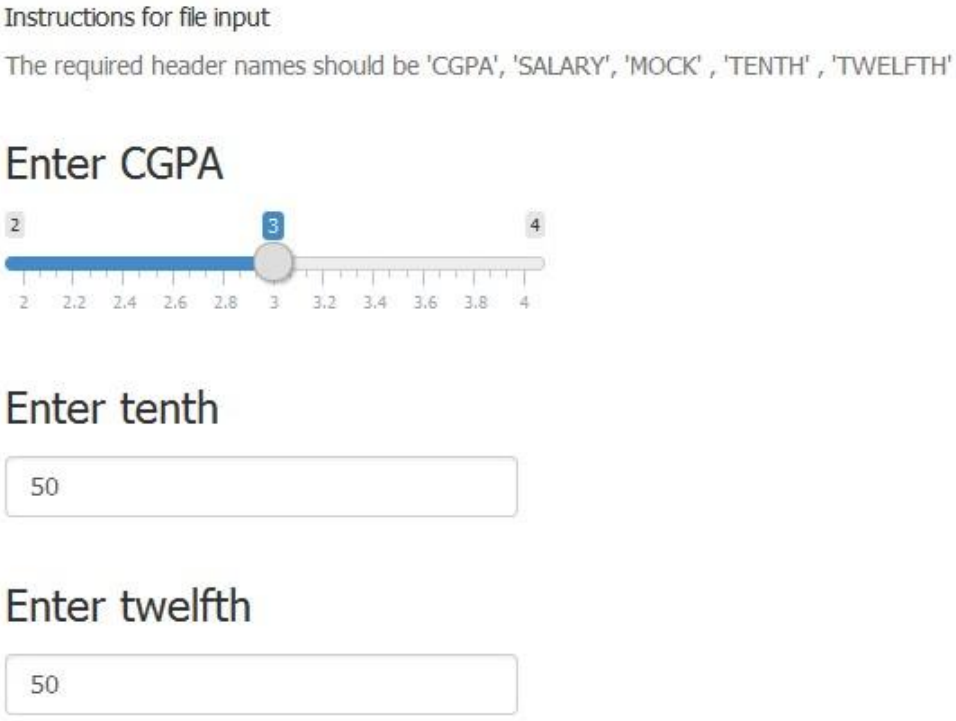


Fig 10.16 'Help' object and different input methods used

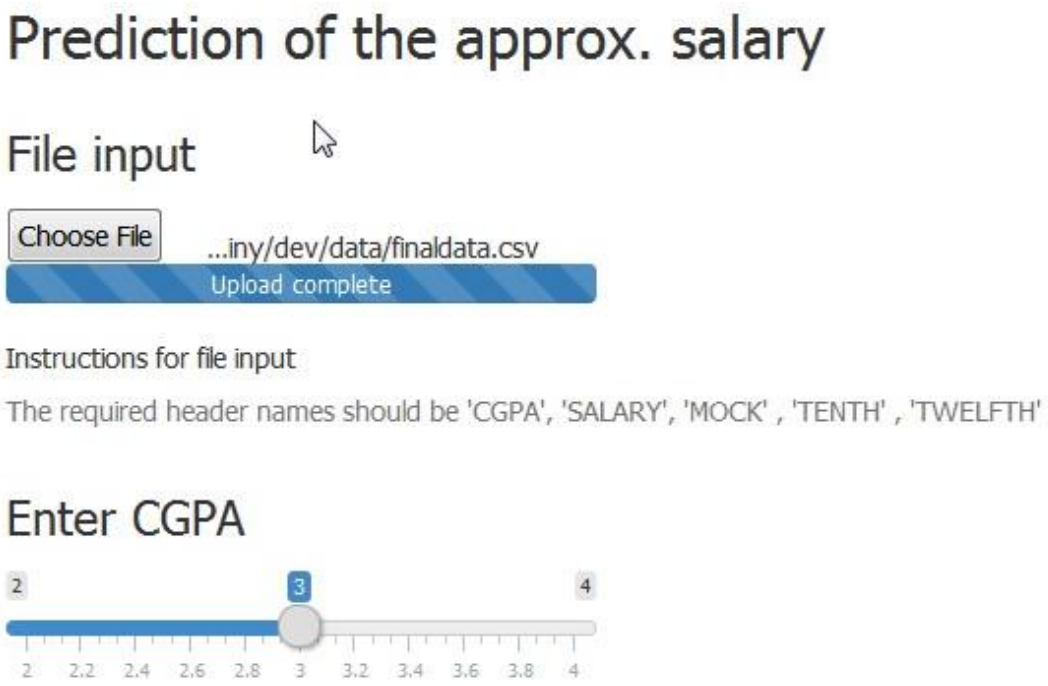


Fig 10.17 Dynamic Data uploading facility

CONCLUSION

Finally, we have implemented a data warehousing and mining project for educational systems in which we acquired a relatively large amount of data, incorporated a data warehouse, cleaning and transformation algorithms for data and extensive comparison and analysis of 5 different algorithms as a primary analysis on our data. We did various comparisons and made various inferences based on the results seen by us. We used the algorithms like K-means, Dbscan, Fuzzy C means, multi Gaussian and hierarchical which gave us clear graphical views describing each cluster, thus enabling us to analyse and make our inferences about them. When we made the data warehouse, from the graphs and pie charts we could easily conclude that there are specific percentages which separate some students from the other based on tenth, twelfth and CGPA. Apart from this we also found a statistical fact giving us a distinction of the number of students separated across campus, branch and gender.

The current stage

- Enables us to clearly judge a student's future performance during a placement procedure taking into account his past results
- Obtain a graphical view for analysing 'n' clusters to set boundary values for relevant data

Predictive analysis and data is the future of technology and we have tried to help incorporate this system in educational institutes.

Using data from our colleges placement cell we have predicted a number of facts; salary, companies to be potentially placed in based on various independent variables like 10^{th}, 12^{th}, mock test and CGPA marks.

Implementing algorithms and getting the desired results was just half of the story. We also implemented a user interface that is easy to use and intuitive. This user interface allows the user to enter custom values for various inputs and also upload their own database.

On a concluding note, **you can have data without information, but you cannot have information without data.**

REFERENCES

[1]. K.Shanmuga Priya and A.V.Senthil Kumar "Improving the Student's Performance Using Educational Data Mining" Int. J. Advanced Networking and Applications Volume: 04 Issue: 04 Pages:1680-1685 (2013) ISSN : 0975-0290

[2]. Kalyani M Raval "Data Mining Techniques" Volume 2, Issue 10, October 2012 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering

[3]. S. Anupama Kumar and M. N. Vijayalakshmi" Relevance of Data Mining Techniques in Edification Sector" International Journal of Machine Learning and Computing, Vol. 3, No. 1, February 2013

[4]. Mohammed M. Abu Tair and Alaa M. El-Halees "Mining Educational Data to Improve Students' Performance: A Case Study" Volume 2 No. 2, February 2012 ISSN 2223-4985 International Journal of Information and Communication Technology Research ©2012 ICT Journal. All rights reserved

[5]. Bhise R.B.,Thorat S.S. and Supekar A.K. "Importance of Data Mining in Higher Education System" IOSR Journal Of Humanities And Social Science (IOSR-JHSS) ISSN: 2279-0837, ISBN: 2279-0845. Volume 6, Issue 6 (Jan. - Feb. 2013), PP 18-21

[6]. Manpreet Singh Bhullar and Amritpal Kaur "Use of Data Mining in Education Sector" Proceedings of the World Congress on Engineering and Computer Science 2012 Vol I WCECS 2012, October 24-26, 2012, San Francisco, USA