Student Performance Prediction using R

INDEX

FIGURE INDEX

TABLE INDEX

1. ABSTRACT: -

➢ EVERYTHING ABOUT OUR PROJECT:

In today's World, living becomes easier with every information beforehand. This is where Data Science plays an important role. We have prepared an algorithm where we can predict Student's performance using various factors such as his previous grades, his current academic performance and then based on these factors we have visualize performance using graphs. We have collected real time data from each student and using these real time data we have applied algorithms on that and then we got our final result.

➢ WHAT WE HAVE DONE:

In this project of ours, we have collected the data from students. We have asked them to provide us with their previous semester pointers, their current state Assignments, Internal assessment marks, practical marks and end semester marks. We have divided the data into two parts which include 70 % of trained data and remaining as test data. We have applied our model on trained data. We have used linear regression in our model.

➢ RESULT VISUALIZATION:

First we will visualize our input data and then we will predict the range within which the grades of students will fit. We have used correlation graph, correlation-gram and histogram in our visualization. And the visualisation will be done with the help of colours for better understanding.

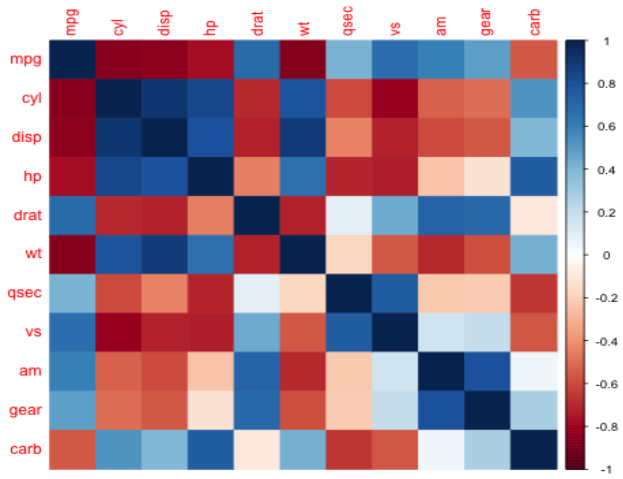

Figure 1 CORRELATION GRAPH

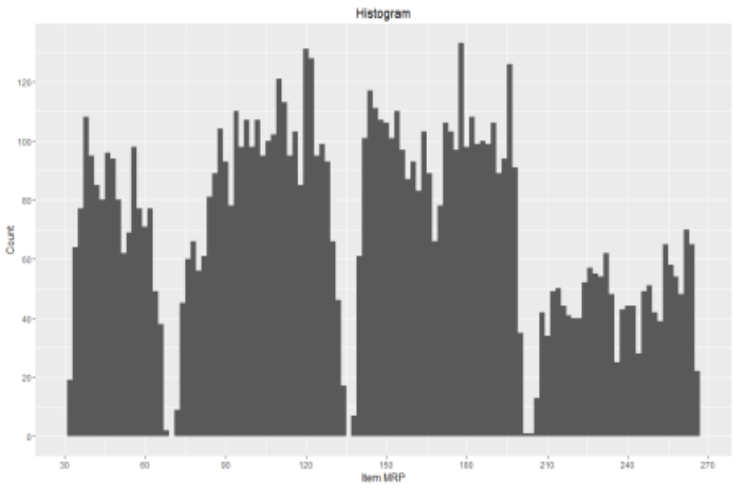

Figure 2 HISTOGRAM GRAPH

2. PROBLEM DEFINITION: -

We have collected data from every student of the class and thus forming the real time dataset. The dataset taken contains the previous semester pointer and current semester pointer. We can segregate the data into trained and test data. The 70% of the data is taken under train data. The linear regression model is applied on this train data. By using the Multivariate Linear Regression model we are able to find out the relation between previous data and current pointer and thus able to find out the equation between input and output.

Before applying the model the data is pre-processed so as to remove non-uniformity. The row having an empty attribute has been removed from the dataset which comes under the data cleaning process. After data-processing, we have used only those attributes which were relevant to the students' performance prediction and columns such as roll no and attendance were removed due to irrelevancy. After this we have calculated the correlation between the attributes and after applying regression model, we found the values of the intercept and the corresponding estimated error, significance code and the R-squared values i.e. accuracy.

Along with this we have visualized the data in the histogram, correlation-gram indicating the number of students in the particular range of pointer. Using the plot graph we have drawn the relation between the different parameter of the statistics representing the accuracy of the regression.

3. LITERATURE SURVEY:-

- **NEED OF THE PROJECT:**

Living in the world which is growing very fast with digital methods and using old methods to do the work is not a good way. Using Modern technology can reduce human efforts and increase efficiency. In this way the smart work process has helped us to implement this "Student Performance Prediction" {for all Students & Teachers}. It is designed for predicting the pointer which can be achieved in coming semester. This application will help to find out the performance of the student and if it is not good then we student have to concentrate in improvement of the result. This application is designed to implement data science and reduce manpower, time and human error in predicting the same using no. of formulas. This system is also helpful for teacher to know about what type of improvement need in student.

Communication gap between student and faculty is been reduced by this system such as students will always been informed about all the notices by manually, apply leave and gate pass. Then for faculty apply leave and outdoor duty applications.

- **EXISTING SYSTEM:**

The system of evaluating student in the college nowadays is done using excel and applying formula. The formula calculation is very much complex and not that much accurate. Also the visualization of that data is not done in proper way. The number of rows and columns in excel sheet is very large and not useful to find out something meaningful from that data.

4. ARCHITECTURE & BLOCK DIAGRAM:-

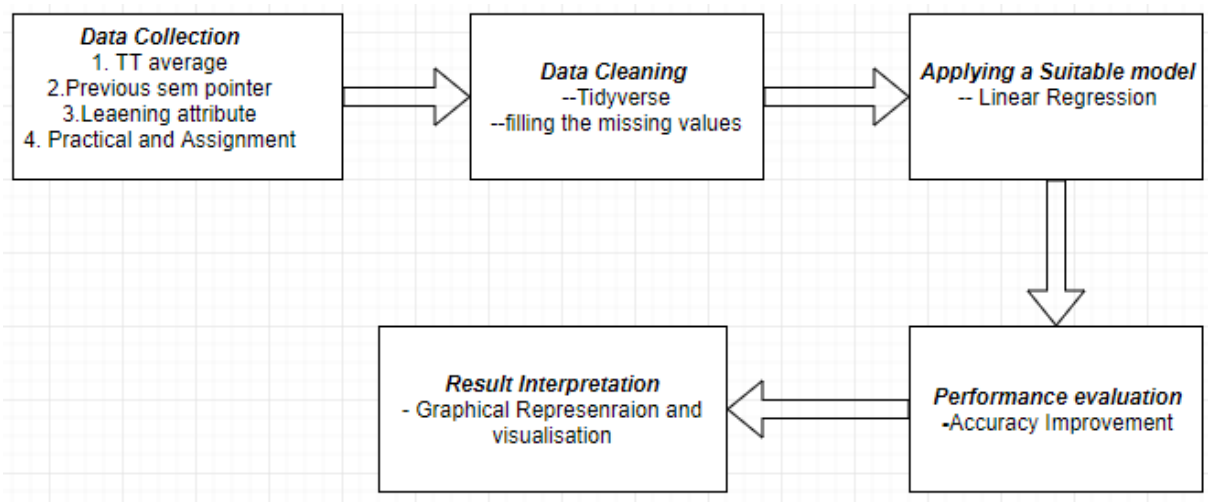

Figure 3 ARCHITECTURE & BLOCK DIAGRAM

- **COLLECTION OF DATA** – Data is collected from the provided excel file. We use the following data:
    - Term Test marks average
    - Previous semester pointer/percentage
    - Learning Attitude
    - Practical Marks
    - Assignment marks
- **CLEANING OF DATA** – We pre-process the data to remove some records with missing values. The following methods are used to clean the dataset:
    - Tidyverse
    - Filling missing value with zero
- **APPLYING A SUITABLE MODEL** – We apply the correct algorithm model on the data. We specifically use algorithms like Multi Variate Linear Regression.

- **PERFORMANCE EVALUATION** – Improving the accuracy of model with the help of true data.
- **RESULT INTERPRETATION** – Use the models developed and the graphs plotted make prediction.

DATA SET:

Table 1 INPUT DATA SET

Roll No	Learning Attitude Avg.	IAT Average	Assignment Avg.	Practical Avg.	Pointer
1	94	11.5	97.33	93	8.72
2	72	0	45	42.5	5.56
3	80.5	11.5	93.33	85	8.76
4	86.5	12.5	85	85	8.4
5	74	11.5	89.33	88.5	8.04
6	86.5	4	89.33	91	6.2

CORRELATIONS:

It's a statistical measure that suggests the level of linear dependence between two variables. If the correlation coefficient is close to 1, it would indicate that the variables are positively linearly related and the scatter plot falls almost along a straight line with positive slope. For -1, it indicates that the variables are negatively linearly related and the scatter plot almost falls along a straight line with negative slope. And for zero, it would indicate a weak linear relationship between the variables.

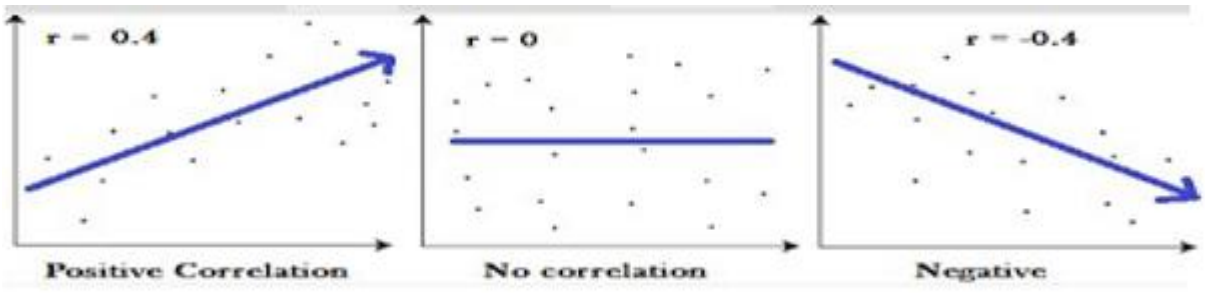

Figure 4 CORRELATION

LINEAR REGRESSION:

Regression methods are used for estimating numeric data. It explores the relationship between a dependent, that is, the target variable and independent variables or the predictor variables. This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables.

Linear regression can be used to find the relationship between one predictor and target variable or between target and more than one predictor variables. Hence, linear regression is divided into 2 types:

- Simple Linear Regression
- Multiple Linear Regression

I. Simple Linear Regression

It establishes the relationship between two variables using a straight line. Hence, only one predictor is used. Simple linear regression draws a line that comes closest to the data by finding the slope and intercept that define the line and minimize regression errors.

It searches for statistical relationship, not deterministic relationship. Relationship between two variables is known to be deterministic if one variable can be accurately expressed by the other. For example, we can easily predict the temperature in Fahrenheit if we already the temperature in Celsius. Statistical relationship is not accurate in determining relationship between two variables. For example, the relationship between height and weight is not the most accurate in all cases.

The fundamental idea is to obtain a line that best fits the data. The best fit line is the one for which total prediction error is as small as possible. Error is the distance between the points to the regression line. Function used for building linear models is **lm ()**.

Example:

```
> linearMod <- lm(dist ~ speed, data=cars) # build linear regression model on full data
> print(linearMod)
```

II. Multiple Linear Regression

Multiple linear regression attempts to model the relationship between two or more explanatory variables and the response variable by fitting a linear equation to observed data. Multiple regression is an extension of simple linear regression. It is used when we want to predict the value of a variable based on the value of two or more other variables. The variable we want to predict is called the dependent variable (or sometimes, the outcome, target or criterion variable). It is also known as Multivariate regression.

If we have 3 dependant variables, (x), the prediction of y is expressed by the following equation:-

$$\mathbf{y = b0 + b1*x1 + b2*x2 + b3*x3}$$

The “b” values are called the regression weights or beta coefficients.

Example:

```
> input <- mtcars[,c("mpg","disp","hp","wt")]
> model <- lm(mpg~disp+hp+wt, data = input)
> summary(model)
```

Here, mpg is the target variable and disp, hp and wt are the predictor variables.

R-SQUARED:

R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination. It is the percentage of the response variable variation that is explained by a linear model.

R-squared = Explained variation / Total variation Where, **R = 0<R<1**

0% indicates that the model explains none of the variability of the response data around its mean. 100% indicates that the model explains all the variability of the response data around its mean. In general, the higher the R-squared, the better the model fits your data.

Formula: $R^2 = 1 – (\sigma^2/Var(Y))$

ADJUSTED R-SQUARED:

Adjusted R-squared adjusts the statistic based on the number of independent variables in the model. Adjusted R2 also indicates how well terms fit a curve or line, but adjusts for the number of terms in a model. If you add more and more useless variables to a model, adjusted r-squared will decrease. If you add more useful variables, adjusted r-squared will increase.

VISUALIZATIONS:

With ever increasing volume of data, it is impossible to tell stories without visualizations. Data visualization is an art of how to turn numbers into useful knowledge.

R Programming lets you learn this art by offering a set of inbuilt functions and libraries to build visualizations and present data. Before the technical implementations of the visualization, let’s see first how to select the right chart type.

Selecting the Right Chart Type

There are four basic presentation types:

- Comparison
- Composition
- Distribution
- Relationship

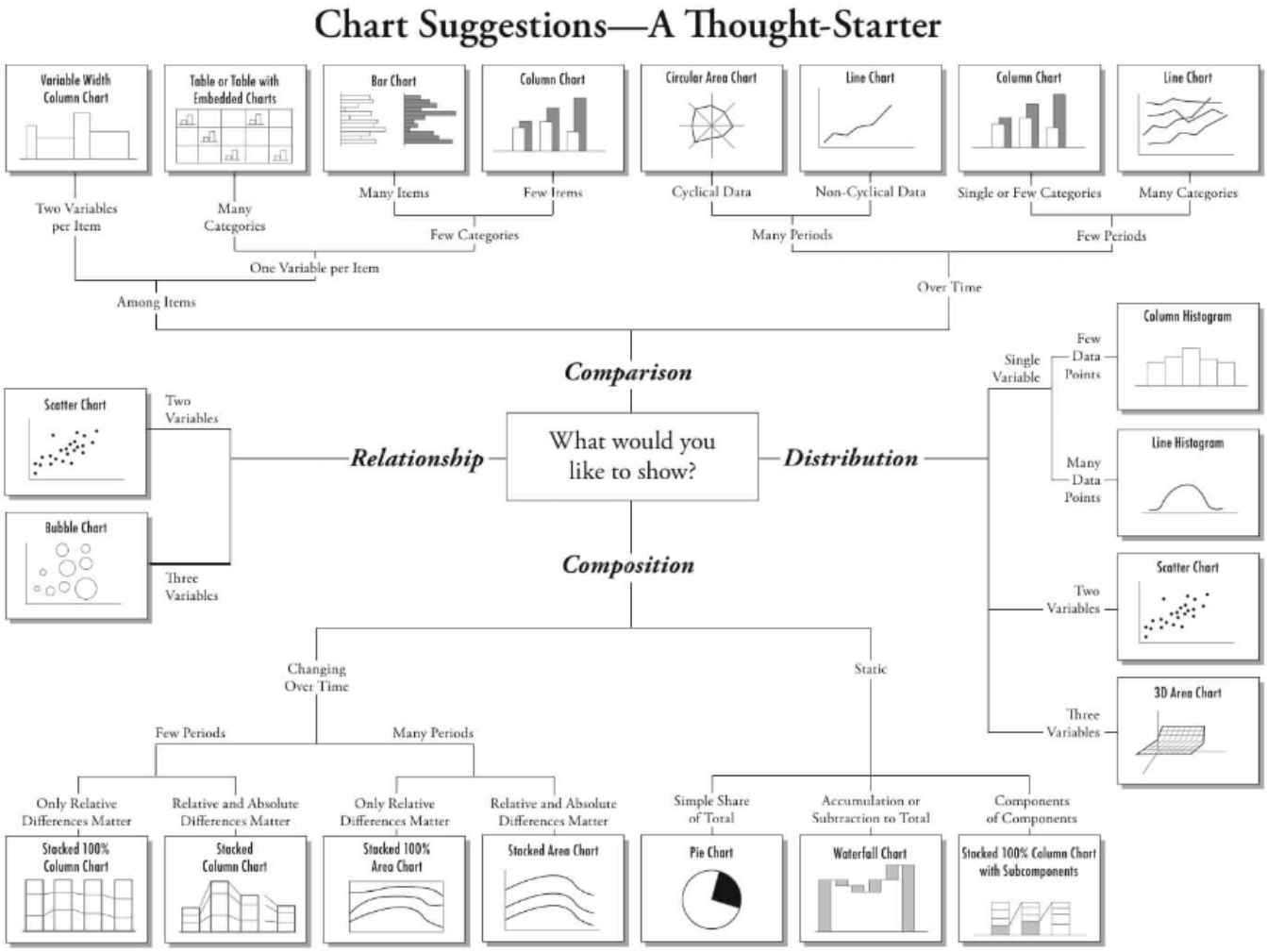

Figure 5 TYPES OF VISUALIZATION

5. FLOWCHART:-

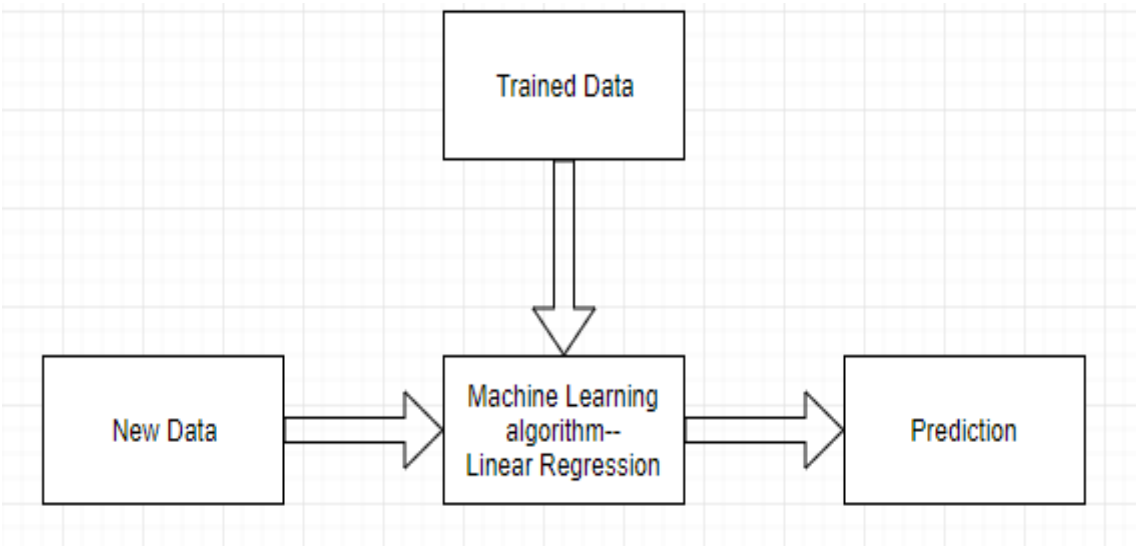

Figure 6 FLOWCHART

In this project, to arrive at a prediction, we use the model of linear regression. In our dataset, we have columns like the term test average, term work, practical and assignments which help us determine a student's pointer. We also have the real time data of the pointers that the students achieved so that we can train the model accordingly and predict the pointers of students in the test dataset.

6. IMPLEMENTATION:-

STEP 1: ALGORITHM

A simple example of regression is predicting Semester 3 pointer of a student when his previous records are known. To do this we need to find the relationship between his different academic year performances.

The steps to create the relationship is:

- Carry out the experiment of gathering a sample of observed values of his previous semester pointers.
- Create a relationship model using the **lm ()** functions in R.
- Find the coefficients from the model created and create the mathematical equation using these
- Get a summary of the relationship model to know the average error in prediction. Also called **residuals**.
- To predict the pointer of current Semester of new persons, use **predict ()** function in R.

STEP 2: SOURCE CODE & OUTPUT

```
# Libraries Requried
library(ggplot2)
library(corrplot)
library(corrgram)
library(dplyr)
library(caTools)
library(Hmisc)
library(tidyr)
library(plotrix)
# Store the csv File data into data as data.frame
data <- read.csv("C:/Users/virag/Desktop/Data Science - R Programming/R Programs/Project.csv")
# Data cleaning and setting Roll_No as character to avoid that column in prediction
data$Roll_No <- as.character(data$Roll_No)
data <- data %>% mutate(LA_1=replace(LA_1,is.na(LA_1),0))
data <- data %>% mutate(IAT_1=replace(IAT_1,is.na(IAT_1),0))
data <- data %>% mutate(Prac_1=replace(Prac_1,is.na(Prac_1),0))
data <- data %>% mutate(Asg_1=replace(Asg_1,is.na(Asg_1),0))
data <- data %>% mutate(Sem_1=replace(Sem_1,is.na(Sem_1),0))
# Printing the summary and type of data
summary(data)
str(data)
# Correlation and Corrplots
# Grab only numeric columns
```

```
num.cols <- sapply(data, is.numeric)
# Filter to numeric columns for correlation
cor.data <- cor(data[,num.cols])
# Correlation Data
corrplot::corrplot(cor.data, method='color')
corrgram(data, order=TRUE, lower.panel=panel.shade, upper.panel=panel.pie,text.panel=panel.txt)
# Histogram
qplot(Sem_1, data=data, geom='histogram', bins=20, fill=..count.., xlab='Sem 1 Pointers', ylab = 'No. of
Students')
# Set the seed
set.seed(101)
# Split the data into test and train sets
sample <- sample.split(data$Sem_1, SplitRatio = 0.83)
# Training data
train <- subset(data[2:6],sample==TRUE)
summary(train)
# Test data
test <- subset(data[2:6],sample==FALSE)
summary(test)
#Apply Multi Variate Regression and create model using train data
model <- lm(Sem_1~., train)
summary(model)
# Grab residuals
res <- residuals(model)
# Convert to data frame for plot
res <- as.data.frame(res)
# Plot residuals
qplot(res, data=res, geom='histogram', bins = 30, xlab = 'Residuals', ylab = 'No. of Students')
par(mfrow=c(2,2))
plot(model)
# Test our model by predicting on our testing set
sem1prec <- predict(model, test)
summary(sem1prec)
# Create a dataset of actual and predicted results to check model performance
results <- cbind(sem1prec, test$Sem_1)
colnames(results) <- c('Predicted', 'Real')
results <- as.data.frame(results)
results
# Plot the predicted pointer
DF <- rbind(data.frame(fill="blue", obs=results$Predicted),
        data.frame(fill="green", obs=results$Real))
ggplot(DF, aes(x=obs, fill=fill)) + ggtitle("Pointers Predicted") +
  xlab("Pointers") + ylab("No. of Students") +
  geom_histogram(binwidth=0.25, colour="black", position="dodge") +
  scale_x_continuous(breaks=seq(0, 10, 0.5)) +
  scale_fill_manual(name="Legend",values=c("red","green"),labels=c("Predicted","Real"))
# Check the performance of our model
sse <- sum((results$Predicted - results$Real)^2)
sst <- sum((mean(data$Sem_1) - results$Real)^2)
R2 <- 1-sse/sst
R2
# Create function to replace negative values with 0
to_zero <- function(x){
  if(x<0){
    return(0)
```

```
  } else {
    return(x)
  }
}
# Apply the function to predicted results
results$Predicted <- sapply(results$Predicted, to_zero)
# Check the range of predicted values
range(results$Predicted)
# Improved Prediction Accuracy
sse <- sum((results$Predicted - results$Real)^2)
sst <- sum((mean(data$Sem_1) - results$Real)^2)
R2 <- 1-sse/sst
RMSE <- sqrt(mean((results$Real - results$Predicted)^2))
R2
RMSE
```

7. RESULT & DISCUSSION:-

After acquiring the real-time data for this project, we have validated the data and then it is pre-processed to remove the dis-continuity in the data. Thus, we are able to derive a proper statistical relation between the various attributes considered in the data like Term-test marks an internal assessment marks, etc. After applying the multivariate linear regression model on the trainee's data set we are able to get the equation between input parameter and the output parameter i.e. the predicted pointer. The trainee's data is obtained by splitting the dataset into train and test data. Over 80% of the data comes under the train data. The applying the machine learning model to trainee's data and putting the output of the model with the test data set containing only the input parameter, we are able to predict the end semester pointer. This pointer is then compared with the real pointer, and the R-squared value for the project i.e. the coefficient of determination, is around 90%, which is pretty good. The R-squared value can further be increased by increasing the number of rows in the dataset so as to get even more accurate value of the intercepts in the relation. The predicted value is comparable to predicted values. We have able to able to visualize the output even with the help of histogram and correlation-gram. The histogram shows the number of the students in the particular pointer range and the correlation-gram is able to show the accuracy. In this way the project has full-filled its aim.

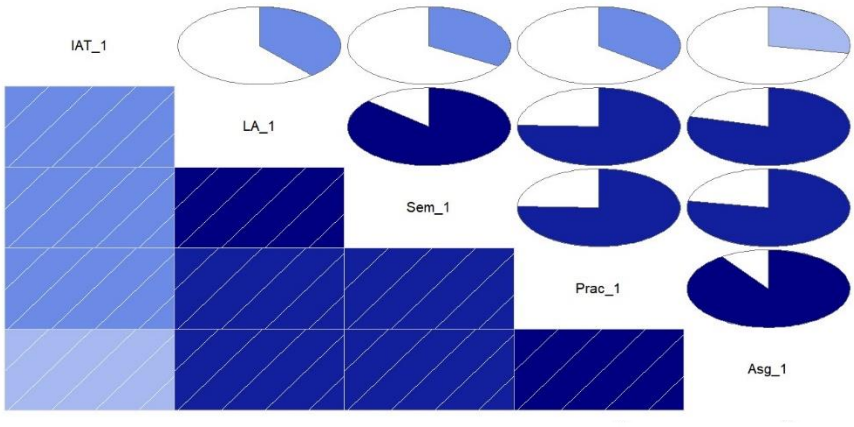

Figure 7 DATA.FRAME IN CORRGRAM

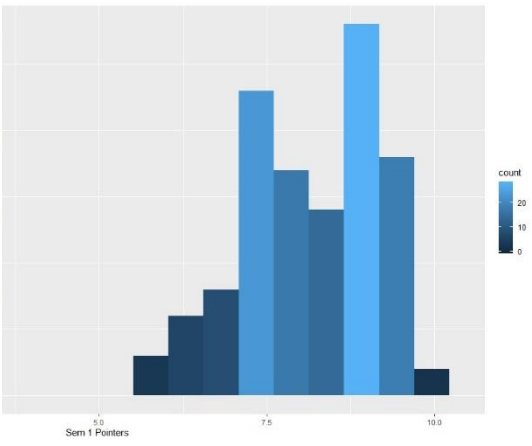

Figure 8 HISTOGRAM

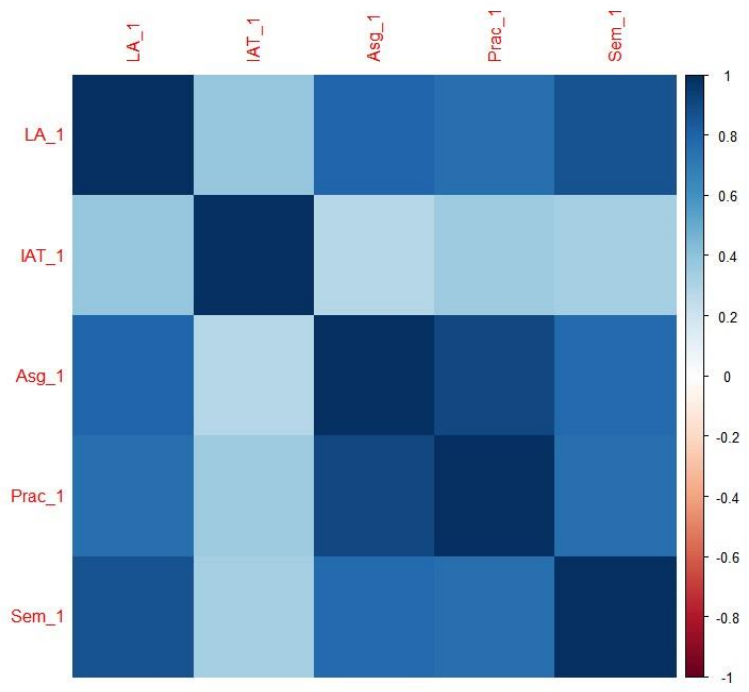

Figure 9 DATA.FRAME IN CORRELATION

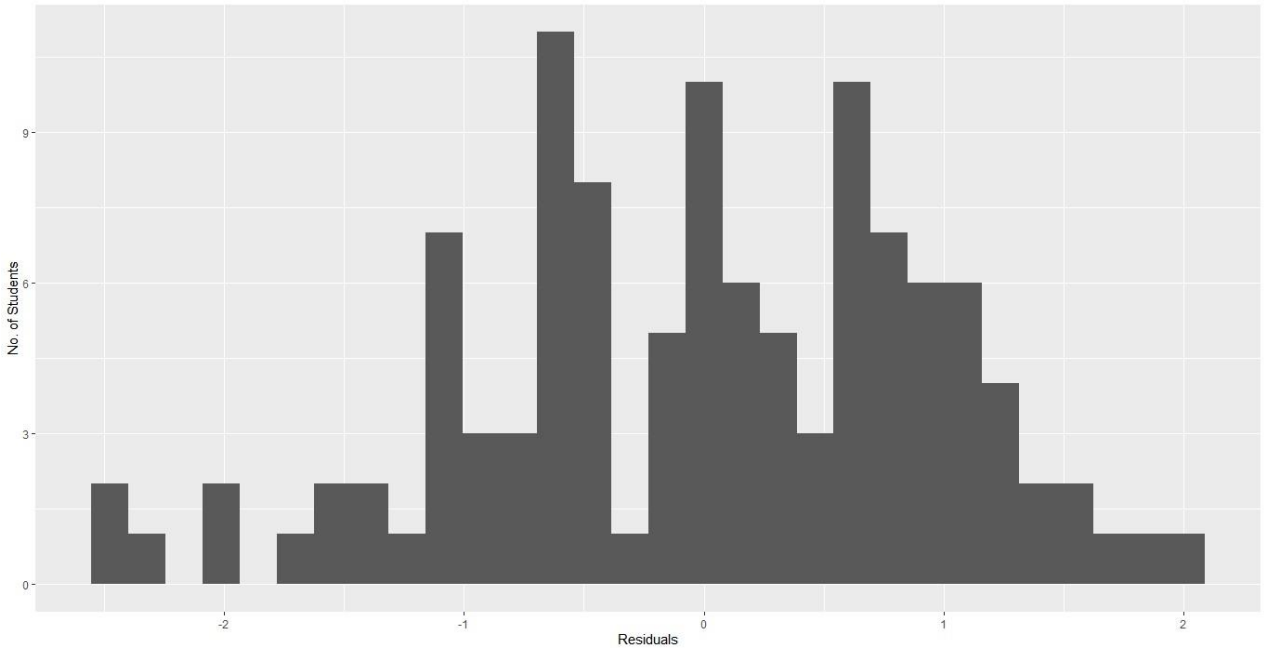

Figure 10 RESIDUAL OUTPUT

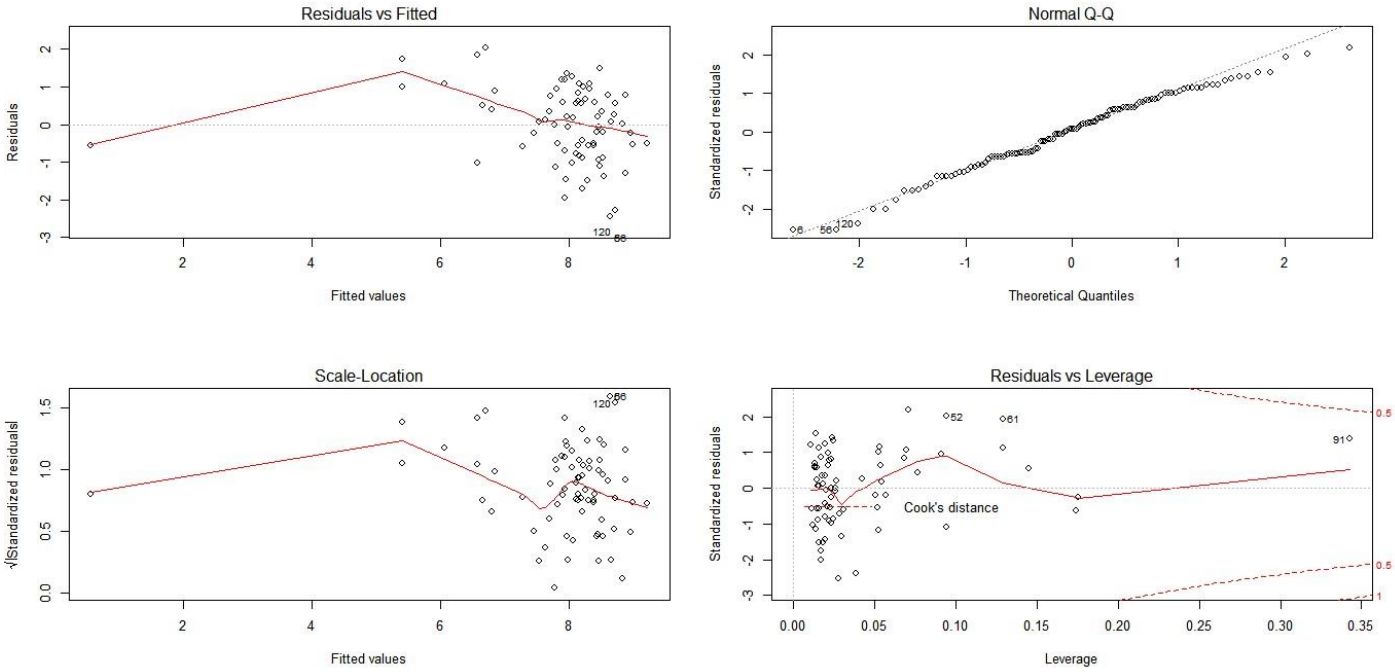

Figure 11 MODEL OUTPUT

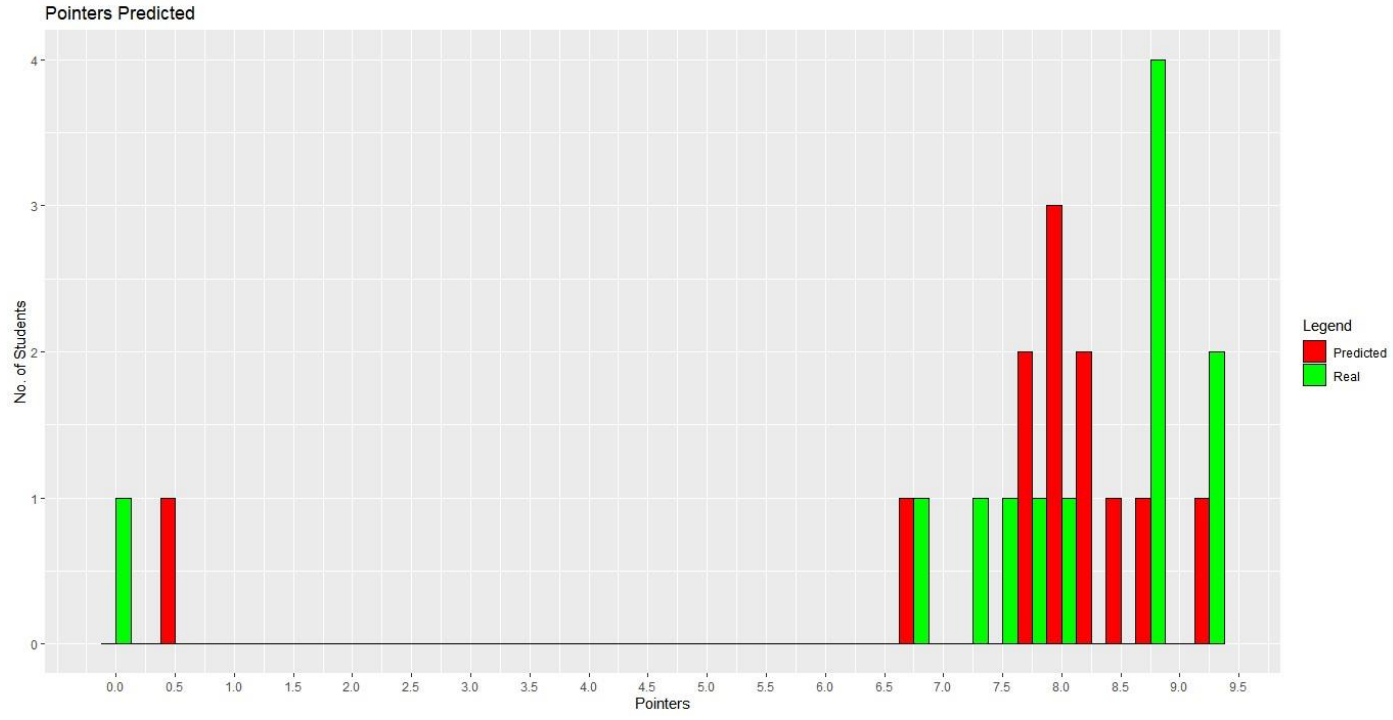

Figure 12 PREDICTED POINTERS Vs REAL POINTERS

Table 2 PREDICTED POINTERS Vs REAL POINTERS

Roll No	Predicted Pointers	Real Pointers
14	8.5064501	8.84
38	7.8242863	7.33
50	8.1139504	8.67
51	9.2212334	8.72
62	8.1893634	8.76
66	8.1422941	7.60
82	8.0497049	9.33
90	6.8556565	7.75
93	8.8362857	7.88
94	8.1058232	9.24
95	0.5693761	0.00
96	7.7706271	6.80

Prediction Accuracy: 0.8933519 i.e. **89%.**

8. CONCLUSION:-

The project is pretty much successful. We are able to understand the basics of the data science, get a quick insight of this field and explore the different aspects of the same. Over this period of time we have understood a few concepts and have obtained good knowledge of how the project should be done i.e. the critical steps that should be considered while working on the project. The data cleaning using suitable algorithm and different ways to represent the data graphically so as to get an overview of the data distribution and applying suitable machine learning model to get desired result are some of the areas which we have learned during the internship.

9. FUTURE SCOPE:-

Off course, there is so much to learn in this vast field. Also the scope of improvisation is there in our project since we have not be able to put it into a proper product. We will try to make it more user-friendly by applying the user-interface and increase the accuracy. We should be able to add even more input attributes. In future, we will surely be able to accomplish it by learning and clearing the concepts more and off course with the help of our mentors of the internship. Also with this project we can predict how the student will perform in coming semester by using his previous semesters' pointers. In this way we can know which students requires attention and could also prove helpful for the college in the prediction of the passing percentage of the students of the college and will be working on the same in future.

10. REFERENCES:-

- We have referred "www.kaggle.com" for the data set for practicing algorithms.
- "https://www.tutorialspoint.com/r/r_multiple_regression.htm" for learning our algorithm.
- https://www.hackerearth.com/practice/machine-learning/linear-regression/multivariate-linear-regression-1/tutorial/
- https://www.r-bloggers.com/r-tutorial-series-multiple-linear-regression/
- https://www.guru99.com/r-simple-multiple-linear-regression.html
- https://towardsdatascience.com/linear-regression-detailed-view-ea73175f6e86
- https://rahilthakur.in/

GROUP MEMBERS:

1. **Virag Dosani (SE CMPN A 77)** _____
2. **Anuja Somthankar (FE CMPN B 35)** _____
3. **Mukul Gharpure (SE CMPN A 32)** _____
4. **Swati Dubey (SE CMPN A 27)** _____

SUPERVISOR:

1. **Dr. Anand Khandare** _____