**“Student Performance Predictor”**

**ARTIFICIAL INTELLIGENCE PROJECT REPORT**

**GROUP MEMBERS:**

Syed Muhammad Asad CS-026

Hafiz M. Saad Waqar CS-031

Bilal Shamim CS-307

**COURSE:**

Artificial Intelligence

**SUBMITTED TO:**

Dr. Saad Qasim Khan  
Assistant Professor (Ph.D Scholar)

**Computer & Information Systems Engineering Department**

**NED University of Engineering & Technology**

**Table of Contents**

1. [INTRODUCTION-----------------------------------------------------------------------------------2](#_Toc535350832)
2. [DATASET OVERVIEW---------------------------------------------------------------------------2](#_Toc535350832)
3. Linear Regression -----------------------------------------------------------------------------------2

* Overview
* Cost Function--------------------------------------------------------------------------------2
* Gradient Decent-----------------------------------------------------------------------------3

1. Graphs ------------------------------------------------------------------------------------4

* GPA vs Course Plot------------------------------------------------------------------------4
* Linear Model--------------------------------------------------------------------------------4
* Residual vs fitted Value Plot---------------------------------------------------------------4
* Normal QQ Plot-----------------------------------------------------------------------------4
* Scale Location Plot-------------------------------------------------------------------------4
* Residual vs Leverage Plot------------------------------------------------------------------4

1. Implementation Of Algorithm -------------------------------------------------------------------10
2. CODE------------------------------------------------------------------------------------------------20
3. OUTPUT --------------------------------------------------------------------------------------------29
4. CONCLUSION-------------------------------------------------------------------------------------29

**INTRODUCTION:**

LINEAR REGRESSION:

In [statistics](https://en.wikipedia.org/wiki/Statistics), linear regression is a [linear](https://en.wikipedia.org/wiki/Linearity) approach to modelling the relationship between a scalar response (or [dependent variable](https://en.wikipedia.org/wiki/Dependent_variable)) and one or more [explanatory variables](https://en.wikipedia.org/wiki/Explanatory_variable) (or [independent variables](https://en.wikipedia.org/wiki/Independent_variable)). The case of one explanatory variable is called [simple linear regression](https://en.wikipedia.org/wiki/Simple_linear_regression). For more than one explanatory variable, the process is called multiple linear regression.[[1]](https://en.wikipedia.org/wiki/Linear_regression#cite_note-Freedman09-1) This term is distinct from [multivariate linear regression](https://en.wikipedia.org/wiki/Multivariate_linear_regression), where multiple correlated dependent variables are predicted, rather than a single scalar variable.[[2]](https://en.wikipedia.org/wiki/Linear_regression#cite_note-2)

In linear regression, the relationships are modeled using [linear predictor functions](https://en.wikipedia.org/wiki/Linear_predictor_function) whose unknown model [parameters](https://en.wikipedia.org/wiki/Parameters) are [estimated](https://en.wikipedia.org/wiki/Estimation_theory)from the [data](https://en.wikipedia.org/wiki/Data). Such models are called [linear models](https://en.wikipedia.org/wiki/Linear_model).[[3]](https://en.wikipedia.org/wiki/Linear_regression#cite_note-3) Most commonly, the [conditional mean](https://en.wikipedia.org/wiki/Conditional_expectation) of the response given the values of the explanatory variables (or predictors) is assumed to be an [affine function](https://en.wikipedia.org/wiki/Affine_transformation) of those values; less commonly, the conditional [median](https://en.wikipedia.org/wiki/Median) or some other [quantile](https://en.wikipedia.org/wiki/Quantile) is used. Like all forms of [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis), linear regression focuses on the [conditional probability distribution](https://en.wikipedia.org/wiki/Conditional_probability_distribution) of the response given the values of the predictors, rather than on the [joint probability distribution](https://en.wikipedia.org/wiki/Joint_probability_distribution) of all of these variables, which is the domain of [multivariate analysis](https://en.wikipedia.org/wiki/Multivariate_analysis).

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications.[[4]](https://en.wikipedia.org/wiki/Linear_regression#cite_note-4) 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.

**WHAT IS LINEAR REGRESSION??**

Linear regression is a basic and commonly used type of predictive analysis.  The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable?  (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?  These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables.  The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b\*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

APPLICATION OF LINEAR REGRESSION:

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 error reduction,[[clarification needed](https://en.wikipedia.org/wiki/Wikipedia:Please_clarify)] linear regression can be used to fit a predictive model to an observed [data set](https://en.wikipedia.org/wiki/Data_set) of values of the response and explanatory variables. After developing such a model, if additional values of the explanatory variables are collected without an accompanying response value, the fitted model can be used to make a prediction of the response.
* If the goal is to explain variation in the response variable that can be attributed to variation in the explanatory variables, linear regression analysis can be applied to quantify the strength of the relationship between the response and the explanatory variables, and in particular to determine whether some explanatory variables may have no linear relationship with the response at all, or to identify which subsets of explanatory variables may contain redundant information about the response

**HISTORY**

The earliest form of regression was the method of least squares, which was published by Legendre in 1805,[6] and by Gauss in 1809.[7] Legendre and Gauss both applied the method to the problem of determining, from astronomical observations, the orbits of bodies about the Sun (mostly comets, but also later the then newly discovered minor planets). Gauss published a further development of the theory of least squares in 1821,[8] including a version of the Gauss–Markov theorem.

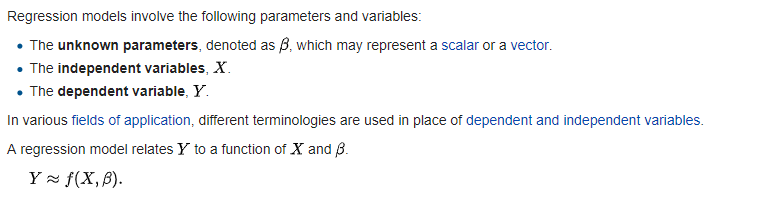
The term "regression" was coined by Francis Galton in the nineteenth century to describe a biological phenomenon. The phenomenon was that the heights of descendants of tall ancestors tend to regress down towards a normal average (a phenomenon also known as regression toward the mean).[9][10] For Galton, regression had only this biological meaning,[11][12] but his work was later extended by Udny Yule and Karl Pearson to a more general statistical context.[13][14] In the work of Yule and Pearson, the joint distribution of the response and explanatory variables is assumed to be Gaussian. This assumption was weakened by R.A. Fisher in his works of 1922 and 1925.[15][16][17] Fisher assumed that the conditional distribution of the response variable is Gaussian, but the joint distribution need not be. In this respect, Fisher's assumption is closer to Gauss's formulation of 1821.

In the 1950s and 1960s, economists used electromechanical desk "calculators" to calculate regressions. Before 1970, it sometimes took up to 24 hours to receive the result from one regression.[18]

Regression methods continue to be an area of active research. In recent decades, new methods have been developed for robust regression, regression involving correlated responses such as time series and growth curves, regression in which the predictor (independent variable) or response variables are curves, images, graphs, or other complex data objects, regression methods accommodating various types of missing data, nonparametric regression, Bayesian methods for regression, regression in which the predictor variables are measured with error, regression with more predictor variables than observations, and causal inference with regression.

**MATHEMATICALLY,**

Regression models involve the following parameters and variables:



**TYPES OF LINEAR REGRESSION:**

There are several types of linear regression analyses available to researchers.

* Simple linear regression  
  1 dependent variable (interval or ratio), 1 independent variable (interval or ratio or dichotomous)

* [Multiple linear regression](http://www.statisticssolutions.com/data-analysis-plan-multiple-linear-regression/)  
  1 dependent variable (interval or ratio) , 2+ independent variables (interval or ratio or dichotomous)

* [Logistic regression](http://www.statisticssolutions.com/data-analysis-plan-logistic-regression/)  
  1 dependent variable (dichotomous), 2+ independent variable(s) (interval or ratio or dichotomous)

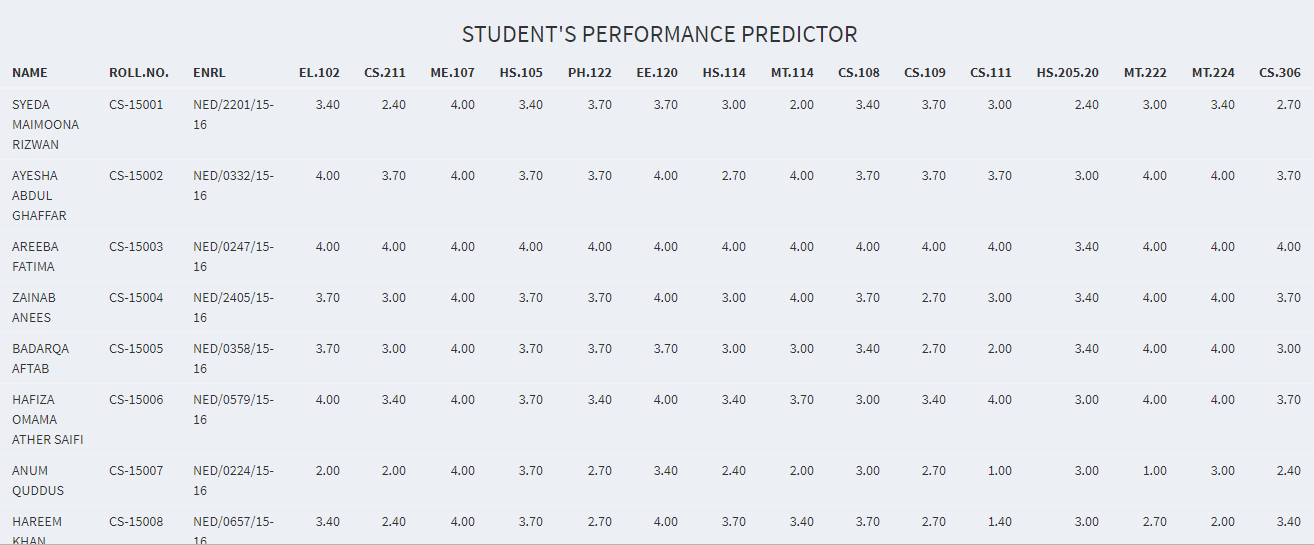
* [Ordinal regression](http://www.statisticssolutions.com/data-analysis-plan-ordinal-regression/)  
  1 dependent variable (ordinal), 1+ independent variable(s) (nominal or dichotomous)

* [Multinominal regression](http://www.statisticssolutions.com/data-analysis-plan-multinominal-logistic-regression/)  
  1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio or dichotomous)

* [Discriminant analysis](http://www.statisticssolutions.com/discriminant-analysis-independent-variables/)  
  1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio)

When selecting the model for the analysis, an important consideration is model fitting.  Adding independent variables to a linear regression model will always increase the explained variance of the model (typically expressed as R²).  However, overfitting can occur by adding too many variables to the model, which reduces model generalizability.  Occam’s razor describes the problem extremely well – a simple model is usually preferable to a more complex model.  Statistically, if a model includes a large number of variables, some of the variables will be statistically significant due to chance alone.

**DATASET OVERVIEW:**

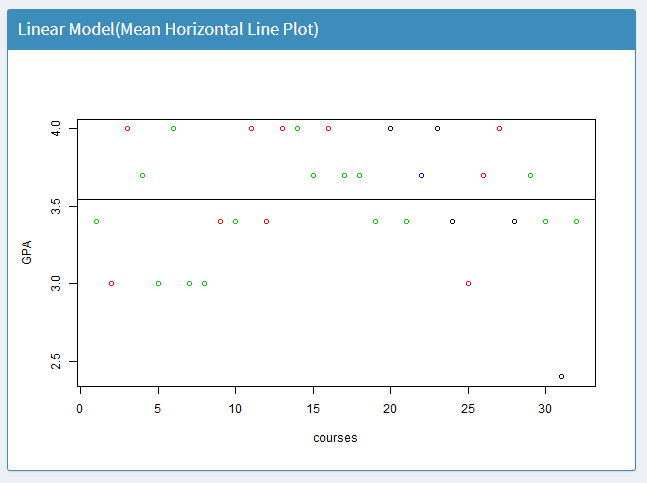
* The dataset that we have used in this project is **CIS Department’s Final Year students of NED University of Engineering and Technology.**
* This dataset contains course wise GPA of every student with their CGPAs per semester.

OVERVIEW:

linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent(y) variable. The line in the graph is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best. The line can be modelled based on the linear equation shown below.

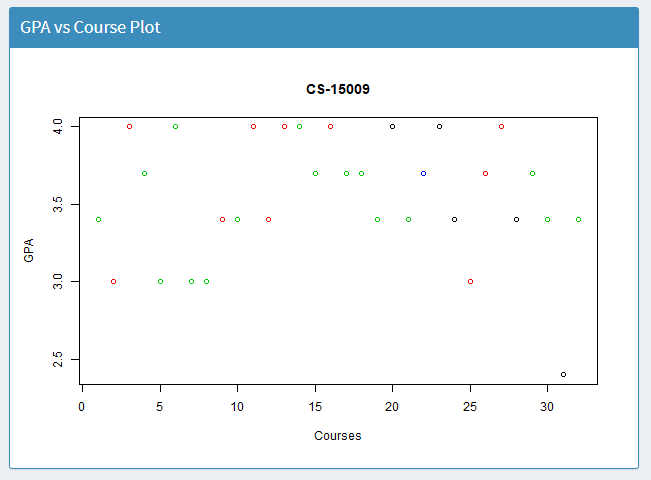
**Y = MX + C**

The motive of the linear regression algorithm is to find the best values for C(intercept) and M(slope). Before moving on to the algorithm, let’s have a look at two important concepts you must know to better understand linear regression

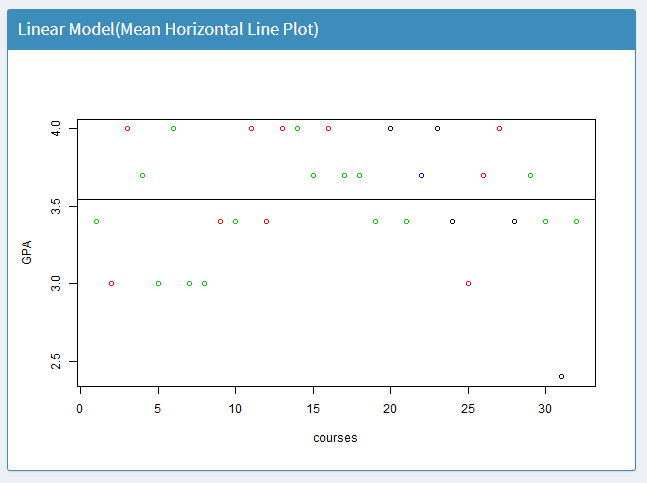


**GPA vs course PLOT**

Example of a student, Roll number CS-009



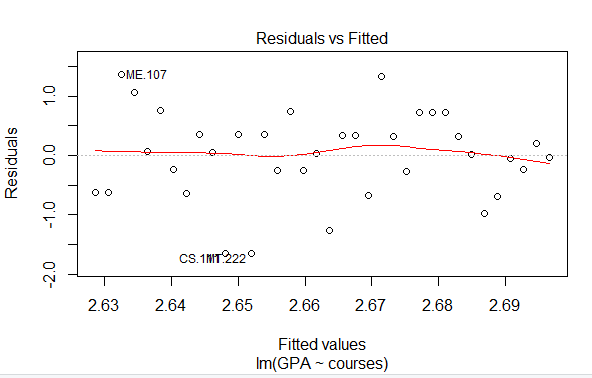
**LINEAR MODEL  
 (MEAN HORIZONTAL LINE PLOT)**

Example of a student, Roll number CS-009

Residual VS Fitted values plot

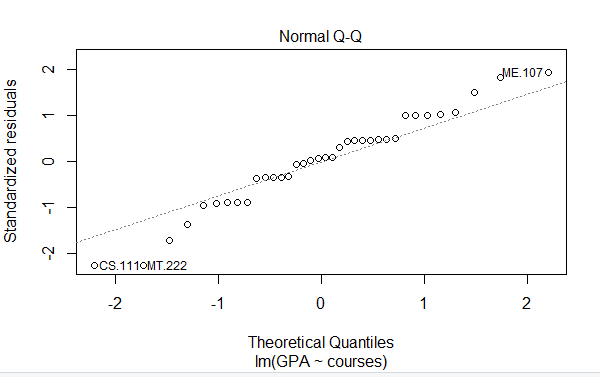
A residual is the difference between the observed value of the dependent variable (y) and the predicted value (ŷ). The first one is the residuals vs. fitted plot. This plot tests the assumptions of whether the relationship between your variables is linear (i.e. linearity) and the whether there is equal variance along the regression line (i.e. homoscedasticity).

A “good” residuals vs. fitted plot should be relatively shapeless without clear patterns in the data, no obvious outliers, and be generally symmetrically distributed around the 0 line without particularly large residuals.



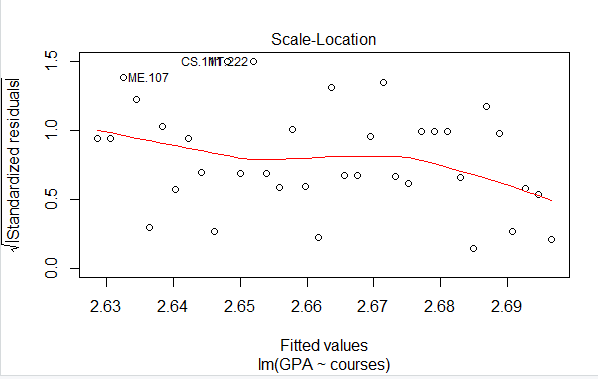
Normal QQ Plot:

The normal qq plot helps us determine if our dependent variable is normally distributed by plotting quantiles (i.e. percentiles) from our distribution against a theoretical distribution. If our data is normally distributed, it will be plotted in a generally straight line on the qq plot.



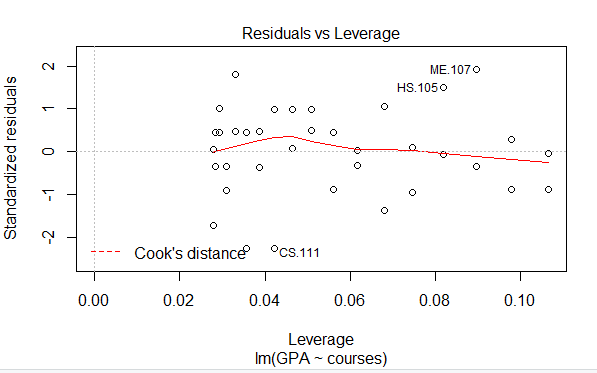
Scale Location Plot:

The Scale-Location plot shows whether our residuals are spread equally along the predictor range, i.e. homoscedastic. We want the line on this plot to be horizontal with randomly spread points on the plot.



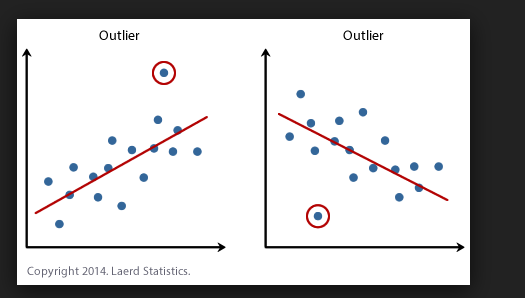
Residuals VS Leverage Plot:

The Residuals vs. Leverage plots helps you identify influential data points on your model. Outliers can be influential, though they don’t necessarily have to it and some points within a normal range in your model could be very influential. The points we’re looking for(or not looking for) are values in the upper right or lower right corners, which are outside the red dashed Cook’s distance line. These are points that would be influential in the model and removing them would likely noticeably alter the regression results.



OUTLIER:

An **outlier** is a single data point that goes far outside the average value of a group of statistics. **Outliers** may be exceptions that stand outside individual samples of populations as well.



COOK’S DISTANCE:

**Cook's distance** measures the effect of deleting a given observation. Points with a large **Cook's distance** are considered to merit closer examination in the analysis.

CODE

UI.r

library(shiny)

library(shinydashboard)

ui <- dashboardPage(

dashboardHeader(title = "ABS"),

dashboardSidebar(

sidebarMenu(

img(src="p.png",width="95%"),

menuItem("Data Set", tabName = "dataset",icon = icon("dashboard")),

menuItem("Select Student", tabName = "selectstd",icon = icon("dashboard")),

menuItem("Students Coursewise GPA", tabName = "stdGPA",icon = icon("dashboard")),

menuItem("Predicted BE GPA", tabName = "BEprediction",icon = icon("dashboard")),

menuItem("Histogram", tabName = "histogram",icon = icon("dashboard")),

h5(textOutput("name")),

h5(textOutput("rollNo"))),

imageOutput("myImage")

),

dashboardBody(

titlePanel(title= h3("STUDENT'S PERFORMANCE PREDICTOR", align="center")),

tabItems(

tabItem(tabName = "dataset",tableOutput("mydataframe")),

tabItem(tabName = "BEprediction",

box(width=12,status = "primary",solidHeader = TRUE,collapsible = FALSE,title = " Final Year Predicted GPA ",verbatimTextOutput("be"))),

tabItem(tabName = "histogram",plotOutput("myhist")),

tabItem(tabName = "stdGPA",

tableOutput("selstd" ),

tableOutput("selstda" ),

tableOutput("selstdb" )

),

tabItem(tabName = "selectstd",

box(width=4,status = "primary",solidHeader = TRUE,collapsible = FALSE,title = "Selection ",selectInput("std", " Select the Student", choices = c("CS-15001", "CS-15002", "CS-15003" ,"CS-15004" ,"CS-15005" ,"CS-15006", "CS-15007", "CS-15008" ,"CS-15009", "CS-15010", "CS-15011", "CS-15012", "CS-15013", "CS-15014", "CS-15015", "CS-15016", "CS-15017", "CS-15018", "CS-15019", "CS-15020", "CS-15021","CS-15022","CS-15022","CS-15023","CS-15024","CS-15025","CS-15026","CS-15027","CS-15028","CS-15029","CS-15030","CS-15031","CS-15032","CS-15033","CS-15034","CS-15035","CS-15306","CS-15307"))),

box(width = 5,status = "primary",solidHeader = TRUE,collapsible = FALSE,title = "Details",tableOutput("selstddetail" )),

box(width = 3,status = "primary",solidHeader = TRUE,collapsible = FALSE,title = "GPA ",h6(textOutput("fe")) ,h6(textOutput("se")) ,h6(textOutput("te")),h6(textOutput("meanCGPA"))),

br(),

br(),br(),

#tableOutput("selstda" ),

#tableOutput("selstdb" ),

box(width = 6,status = "primary",solidHeader = TRUE,collapsible = FALSE,title = "GPA vs Course Plot",plotOutput("myplot")),

box(width = 6,status = "primary",solidHeader = TRUE,collapsible = FALSE,title = " Linear Model(Mean Horizontal Line Plot) ",plotOutput("prediction")),

box(width = 6,status = "primary",solidHeader = TRUE,collapsible = FALSE,title = " Linear Model(Residuals VS Leverage) ",plotOutput("model")),

box(width = 6,status = "primary",solidHeader = TRUE,collapsible = FALSE,title = " Best Fitted Line For Linear Model ",plotOutput("bestline"))

)

)

)

)

**SERVER.r**

library(shiny)

library(shinydashboard)

library(shinythemes)

shinyServer(

function(input, output) ({

output$tabset1Selected <- renderText({

input$tabset1})

output$myImage <- renderImage({

if(aa() == 307){ #condition for roll number 306 and 307

list(src = "www/m.jpg",

contentType = 'image/png',

width = 225,

height = 225,

alt = "")

}else if(aa()== 1 || aa()== 2 || aa()== 3 || aa()== 4 || aa()== 5 || aa()== 6 || aa()== 7 || aa()== 8 || aa()== 9 || aa()== 10

|| aa()== 11 || aa()== 12 || aa()== 13 || aa()== 14 || aa()== 15 || aa()== 16 || aa()== 17 || aa()== 18 || aa()== 19 || aa()== 20

|| aa()== 21 || aa()== 25

)

{

list(src = "www/f.jpg",

contentType = 'image/png',

width = 225,

height = 225,

alt = "")

}else{

list(src = "www/m.jpg",

contentType = 'image/png',

width = 225,

height = 225,

alt = "")

}

},deleteFile = FALSE)

output$mystd <- renderText(

paste ("You selected quantitive variable: ", input$std))

mydataframe <- data.frame(data) #making the dataframe

aa <- reactive ({ as.numeric(substr(input$std,6,9))}) #slicing the string e.g from CS-15004 to 004

seca <- reactive({ if(aa() == 307){ #condition for roll number 306 and 307

aa <- 37;

}else if(aa()== 306)

{

aa<- 36;

}else{

aa <- aa();

}

} )

output$selstd <- renderTable({

azb <- mydataframe[seca(),]

azb[,1:15]

})

output$mydataframe <- renderTable({

mydataframe})

output$selstddetail <- renderTable({

azb <- mydataframe[seca(),]

azb[,1:3]

})

output$selstda <- renderTable({

azb <- mydataframe[seca(),]

azb[,16:32]

})

output$selstdb <- renderTable({

azb <- mydataframe[seca(),]

azb[,33:39]

})

output$selstd1 <- renderTable({

t(mydataframe[seca(),] )

})

output$mydataframe <- renderTable({ #printing all the student's information

mydataframe})

toplot1 <- reactive ({toplot()[4:35,1]})

output$myhist <- renderPlot({

#hist(mydataframe$CGPA,xlab = "GPA",ylab = "frequency",col='blue',main= "Histogram")

hist(as.numeric(toplot1()),breaks = seq(0, max(8,l=6)), col='blue', main="Histogram of CGPA",

xlab="CGPA")

})

datadatatext.features <- mydataframe

datadatatext.features$NAME <- NULL

datadatatext.features$ROLL.NO. <- NULL

datadatatext.features$ENRL <- NULL

results <- kmeans(datadatatext.features,4)

d<-reactive({

d<-t(mydataframe[seca(),])

d<-data.frame(d)

names(d)<-c("GPA")

d$GPA<-as.numeric(as.character(d$GPA))

library(tibble)

add\_column(d,courses=as.numeric(c(NA,NA,NA,1:36)),.after = 1)

#d$GPA<-as.numeric(as.character(d$GPA))

#d$courses<-as.numeric(as.character(d$courses))

})

output$dd <- renderTable({

d() })

output$prediction <- renderPlot({

plot(GPA~courses,data=d()[4:35,],col=results$cluster)

mean.GPA=mean(d()$GPA,na.rm = T)

abline(h=mean.GPA)

model1=lm(GPA~courses,data=d())

summary(model1)

})

output$bestline <- renderPlot({

plot(GPA~courses,data=d()[4:35,],col=results$cluster)

mean.GPA=mean(d()$GPA,na.rm = T)

model1=lm(GPA~courses,data=d())

abline(model1)

summary(model1)

})

toplot <- reactive ({ t(mydataframe[seca(),])})

name <- reactive({ toplot()[1,1]})

output$name <- renderText(

paste ("Name : ", name()))

rollNo <- reactive({ toplot()[2,1]})

output$rollNo <- renderText(

paste ("Roll No. : ", rollNo()))

fe <- reactive ({toplot()[36,1]})

output$fe <- renderText(

paste ("FE GPA : ", fe()))

se <- reactive ({toplot()[37,1]})

output$se <- renderText(

paste ("SE GPA : ", se()))

te <- reactive ({toplot()[38,1]})

output$te <- renderText(

paste ("TE GPA : ", te()))

gpa\_mean <- reactive({

mean(d()$GPA,na.rm = T)

})

output$meanCGPA <- renderText(

paste ("Mean CGPA : ", gpa\_mean()))

cgpa <- reactive ({toplot()[39,1]})

output$be <- renderPrint ({

model2=lm(GPA~courses,data=d())

summary(model2)

})

output$model <- renderPlot({

model1=lm(GPA~courses,data=d())

plot(model1)

})

output$myplot <- renderPlot({

plot(toplot1(), col=results$cluster ,main = input$std ,xlab = "Courses",ylab = "GPA")

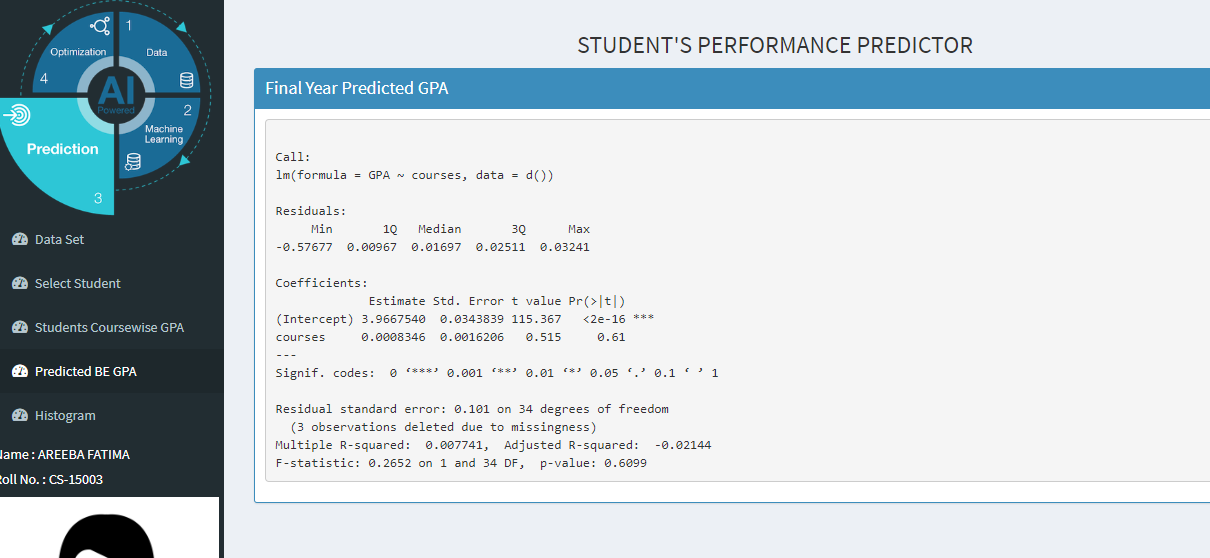
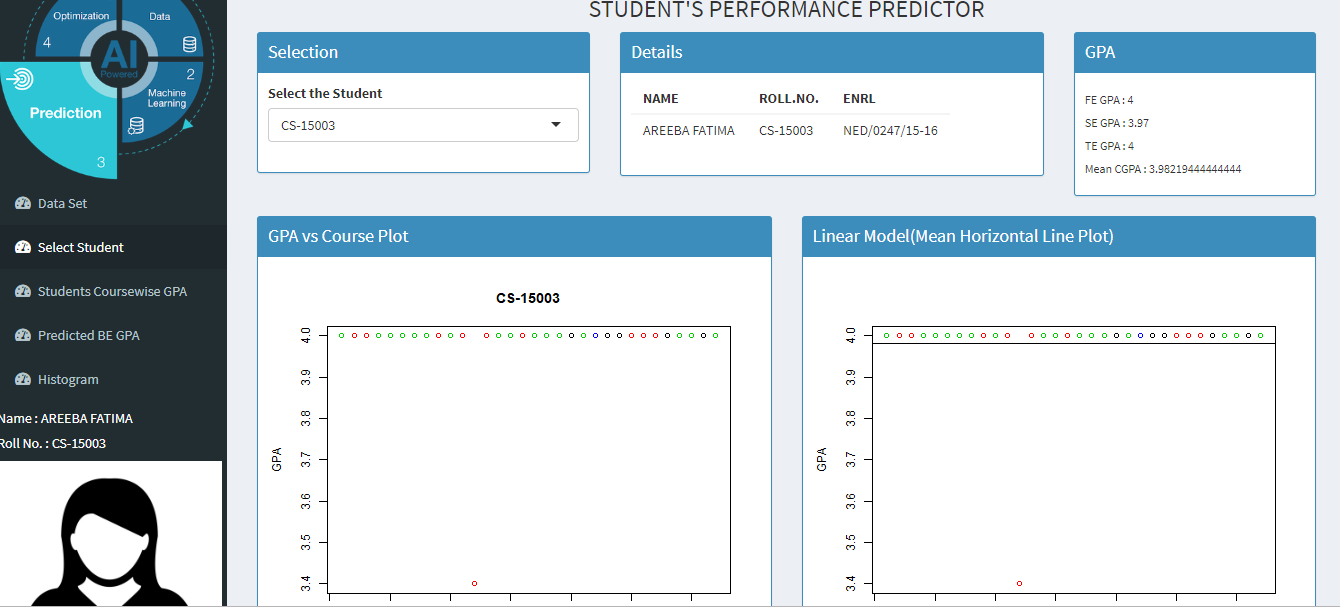
})

}

))

OUTPUT

Example of a student, Roll number : CS-003



Example of a student, Roll number : CS-003

