

**Department of Computer Science**

FAST – National University of Computer & Emerging Sciences

Chiniot-Faisalabad Campus.

**MT200-Probability and Statistics**

Course Instructor

**Dr. Haris Khurram**

****

**Group Members:**

|  |  |
| --- | --- |
| **Name** | **Roll Number** |
| M. Zain ul Waqas | 21F-9274 |
| Rai Muhammad Saqlain | 21F-9255 |
| Rimsha Sarfraz | 21F-9244 |
| Maheen Naeem | 22F-3145 |
| Sadaf Tahir | 22F-3195 |

**Section: 5D**

**Desktop App**

**For Statistical Methods that Analyze the Data**

**Diabetes Data Analysis**Date: **Friday, December 8, 2023**

**1. Problem Statement**

Diabetes, a prevalent and potentially life-threatening metabolic disorder, impacts millions of individuals globally. Unraveling the factors linked to favorable or adverse outcomes in diabetes patients is imperative for crafting efficient prevention, early diagnosis, and treatment strategies. Despite advancements in diabetes research and healthcare, several challenges persist, prompting the need for innovative solutions to enhance the overall management of this significant health issue.

**2. Objective**

This study aims to analyze a dataset of diabetes patients to identify factors associated with survival outcomes, disease recurrence, and other clinical features of the disease.

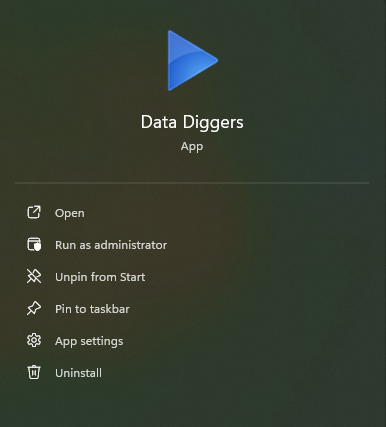
**3. Data Description**

The dataset used in this analysis is taken from Kaggle (<https://www.kaggle.com/datasets/aravindpcoder/diabetes-dataset>).

The data were collected from the Iraqi society, as they data were acquired from the laboratory of Medical City Hospital and (the Specializes Center for Endocrinology and Diabetes-Al-Kindy Teaching Hospital). Patients' files were taken and data extracted from them and entered in to the database to construct the diabetes dataset. The data consist of medical information, laboratory analysis. The data that have been entered initially into the system are: No. of Patient, Sugar Level Blood, Age, Gender, Creatinine ratio(Cr), Body Mass Index (BMI), Urea, Cholesterol (Chol), Fasting lipid profile, including total, LDL, VLDL, Triglycerides(TG) and HDL Cholesterol , HBA1C, Class (the patient's diabetes disease class may be Diabetic, Non-Diabetic, or Predict-Diabetic.

**4. Results**

**Display on Search bar:**

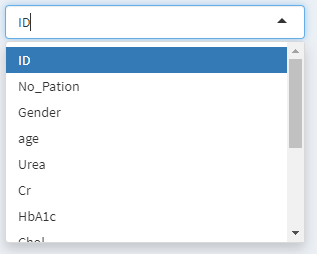
****

**Display on Desktop:**

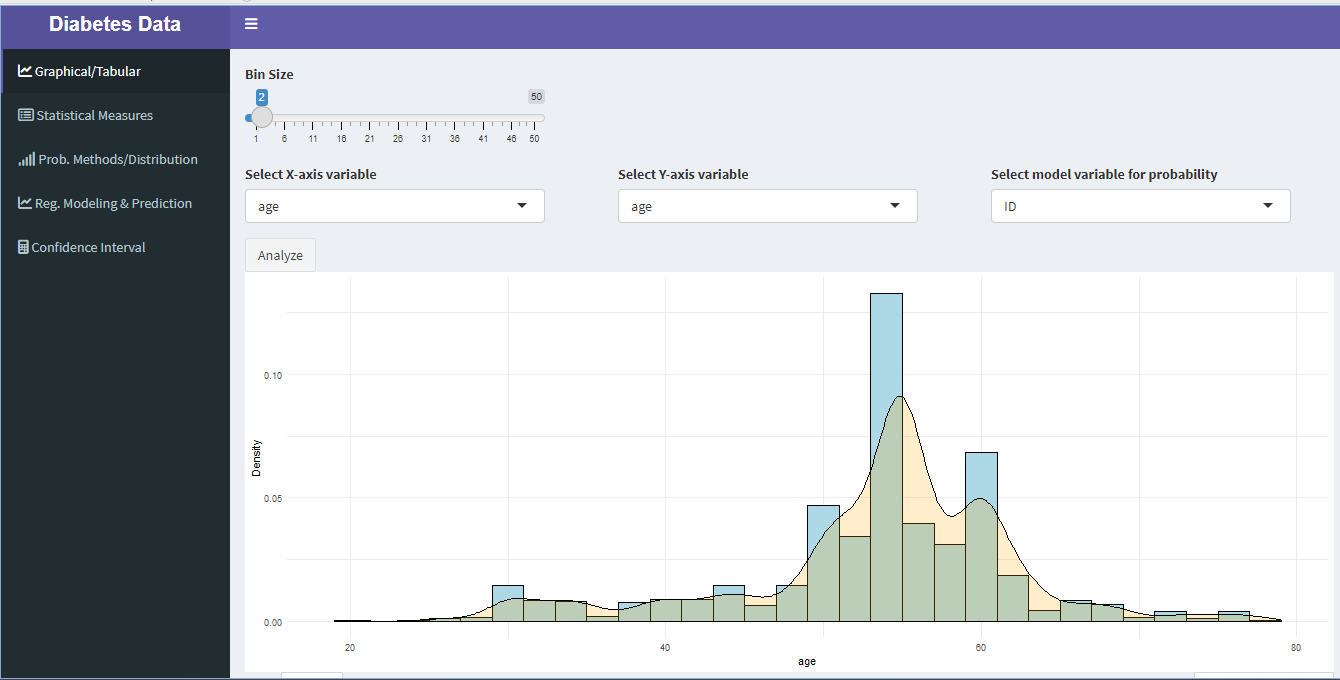
**A screen shot of a computer

Description automatically generated**

**Display for selection of column to draw Graphs:**



**Graphical Representation:**



**Description:**

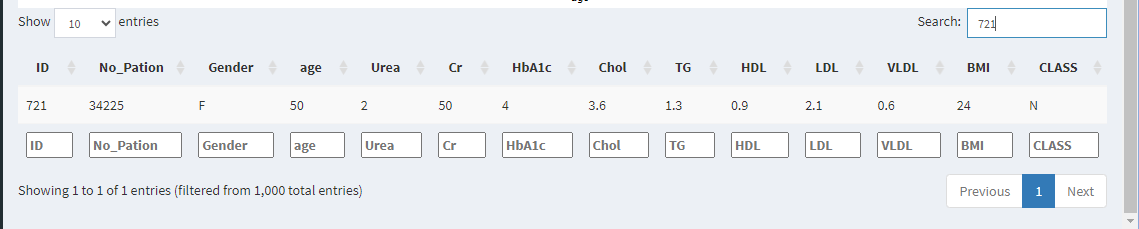
Histograms were created to depict the distributions of age, Urea, and HbA1c variables in a diabetes dataset. Summary statistics, including mean and standard deviation, were computed for these variables. Furthermore, a stacked bar chart illustrates the association between gender and diabetes classification, providing a concise overview of the relationships within the dataset.

**Table that shows Data:**

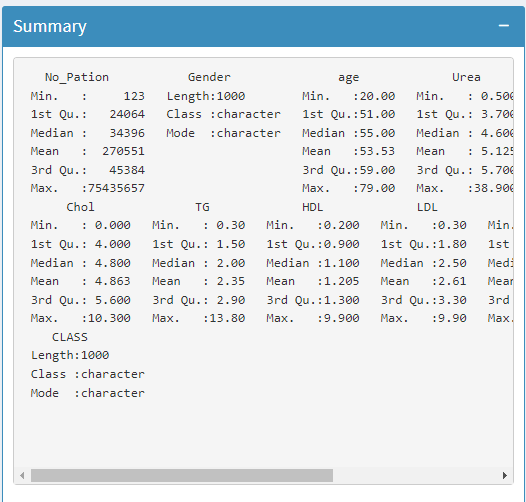
A screenshot of a computer

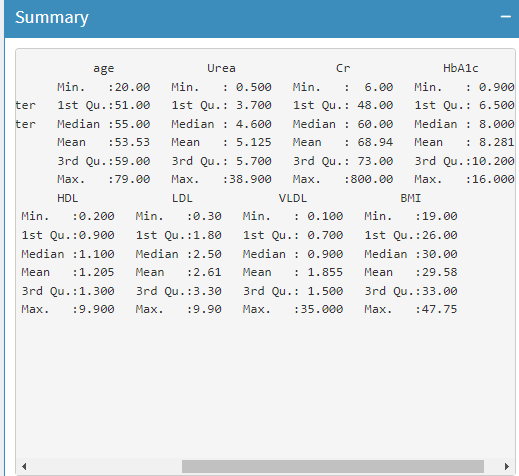
Description automatically generated

**Search the patient for details:**



**Summary:**





**Probability Methods/Distribution:**

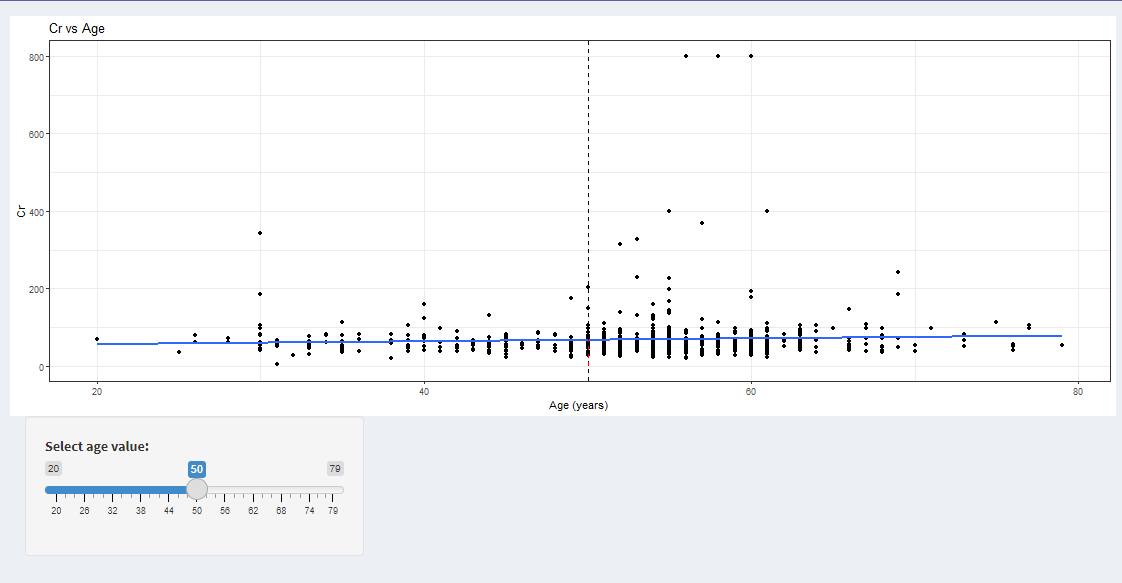
A graph on a screen

Description automatically generated

**Description:**

Graphs were utilized to illustrate the survival probabilities of diabetes patients based on Urea levels, lymph node involvement, and BMI. Cox proportional hazards models were applied to pinpoint factors associated with favorable or adverse survival outcomes.

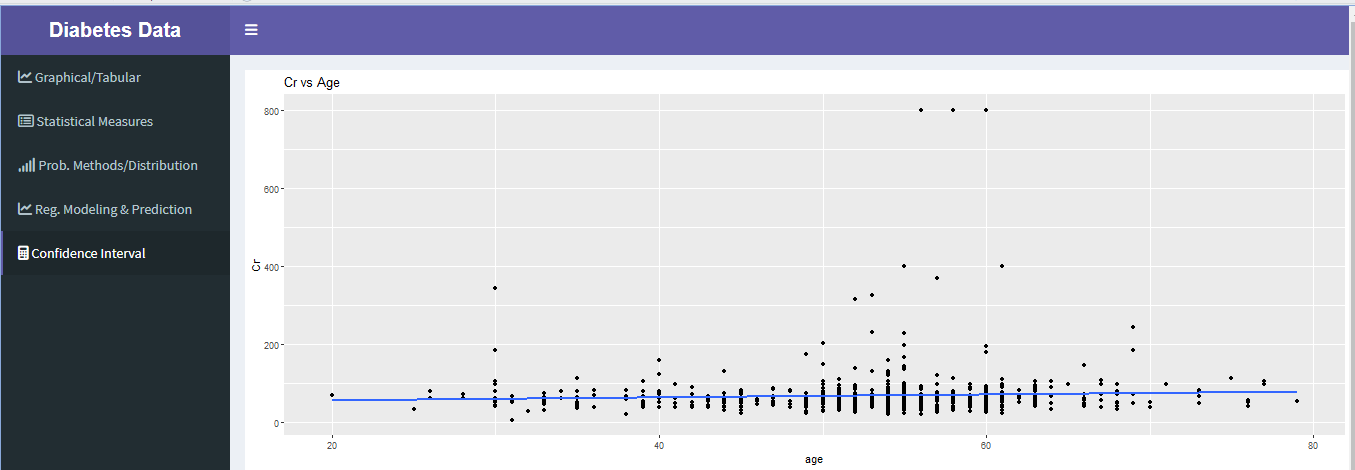
**Regression Modeling and Predictions:**



**Description:**

A logistic regression model was developed to predict the likelihood of adverse outcomes in diabetes patients based on their clinical and demographic characteristics. With a moderate accuracy level, the model highlighted age, Urea levels, and BMI as the most crucial predictors for recurrence or mortality in the dataset. This concise analysis provides key insights into the influential factors impacting adverse outcomes in diabetes

**Confidence Interval of Descriptive measures and Regression Estimates:**



A screenshot of a computer

Description automatically generated

**5. Codes**

library(shiny)

library(readxl)

library(ggplot2)

library(dplyr)

library(tidyr)

library(MASS)

library(lmtest)

library(shinydashboard)

Diabetes <- read\_excel("data.xlsx")

data\_path <- "data.xlsx"

data <- read\_excel(data\_path)

datanew <- read\_excel(data\_path)

lm\_model <- lm(Cr ~ age, data = data)

age\_CI <- t.test(data$age)$conf.int

ui <- dashboardPage(

skin = "purple",

dashboardHeader(

title = tags$div(

style = "white; color: #ffffff; font-weight: bold;",

"Diabetes Data "

)

),

dashboardSidebar(

sidebarMenu(

menuItem(" Graphical/Tabular", tabName = "graphical", icon = icon("chart-line")),

menuItem(" Statistical Measures", tabName = "descriptive", icon = icon("list-alt")),

menuItem(" Prob. Methods/Distribution", tabName = "probability", icon = icon("signal")),

menuItem(" Reg. Modeling & Prediction", tabName = "regression", icon = icon("line-chart")),

menuItem(" Confidence Interval", tabName = "confidence", icon = icon("calculator"))

)

),

dashboardBody(

tabItems(

tabItem(

tabName = "graphical",

fluidRow(

column(width = 4, sliderInput("bin\_size", "Bin Size", min = 1, max = 50, value = 10)),

),

fluidRow(

column(width = 4, selectInput("x\_axis", "Select X-axis variable", choices = names(Diabetes))),

column(width = 4, selectInput("y\_axis", "Select Y-axis variable", choices = names(Diabetes))),

column(width = 4, selectInput("model\_variable", "Select model variable for probability", choices = names(Diabetes)))

),

fluidRow(

column(width = 12, actionButton("analyze", "Analyze"))

),

fluidRow(

column(width = 12, plotOutput("plot"))

),

fluidRow(

column(width = 12, dataTableOutput("table"))

)

),

tabItem(

tabName = "descriptive",

fluidRow(

column(

width = 4,

valueBoxOutput("n\_box")

),

column(

width = 4,

valueBoxOutput("mean\_box")

),

column(

width = 4,

valueBoxOutput("sd\_box")

)

),

fluidRow(

column(

width = 4,

valueBoxOutput("min\_box")

),

column(

width = 4,

valueBoxOutput("median\_box")

),

column(

width = 4,

valueBoxOutput("max\_box")

)

),

fluidRow(

column(

width = 12,

box(

title = "Summary",

status = "primary",

solidHeader = TRUE,

collapsible = TRUE,

collapsed = TRUE,

verbatimTextOutput("summary")

)

)

)

),

tabItem(

tabName = "probability",

plotOutput("density\_plot"),

verbatimTextOutput("distribution\_summary")

),

tabItem(

tabName = "regression",

plotOutput("scatterplot"),

sidebarPanel(

sliderInput(

"age",

"Select age value:",

min = min(datanew$age),

max = max(datanew$age),

value = 50

)

)

),

tabItem(

tabName = "confidence",

plotOutput("scatterplot2"),

fluidRow(

column(

width = 12,

box(

title = "Confidence Interval for Age Mean",

status = "primary",

solidHeader= TRUE,

verbatimTextOutput("confidence\_interval")

),

box(

title = "Confidence Interval for Age Mean",

status = "primary",

solidHeader= TRUE,

verbatimTextOutput("lm\_CI"),

),

box(

title = "Confidence Interval for Age Mean",

status = "primary",

solidHeader= TRUE,

verbatimTextOutput("age\_CI"),

)

)

)

)

)

)

)

server <- function(input, output) {

data <- eventReactive(input$analyze, {

Diabetes

})

output$plot <- renderPlot({

ggplot(data(), aes(x = !!sym(input$x\_axis))) +

geom\_histogram(binwidth = input$bin\_size, aes(y = ..density..), fill = "lightblue", color = "black") +

geom\_density(alpha = .2, fill = "orange") +

labs(x = input$x\_axis, y = "Density") +

theme\_minimal()

})

output$table <- renderDataTable({

data()

})

output$summary <- renderPrint({

summary(data()[, -1])

})

output$density\_plot <- renderPlot({

ggplot(data(), aes(x = !!sym(input$model\_variable))) + geom\_density()

})

output$distribution\_summary <- renderPrint({

summary(data()[, input$model\_variable])

})

output$regression\_summary <- renderPrint({

model <- lm(!!sym(input$model\_variable) ~ ., data = data())

summary(model)

})

# Scatter plot

output$scatterplot2 <- renderPlot({

ggplot(Diabetes, aes(x = age, y = Cr)) +

geom\_point() +

geom\_smooth(method = "lm", se = FALSE) +

ggtitle("Cr vs Age")

})

# Confidence intervals for age mean

output$age\_CI <- renderPrint({

paste("Confidence Intervals for Age Mean:",

paste(round(age\_CI, 2), collapse = " - "))

})

# Confidence intervals for linear regression coefficients

lm\_CI <- confint(lm\_model)

output$lm\_CI <- renderPrint({

paste("Confidence Intervals for Linear Regression Coefficients:",

paste(round(lm\_CI, 2), collapse = " - "))

})

output$confidence\_interval <- renderPrint({

summary(lm\_model)

})

# Define the data subset based on the selected age value

data\_subset <- reactive({

datanew %>% filter(age == input$age)

})

# Define the linear regression model based on the data subset

my\_lm\_model <- reactive({

lm(Cr ~ age, data = data\_subset())

})

# Define the predicted values based on the selected age value

predictions <- reactive({

predict(my\_lm\_model(), newdata = data.frame(age = input$age))

})

# Define the scatterplot

output$scatterplot <- renderPlot({

ggplot(datanew, aes(x = age, y = Cr)) +

geom\_point() +

geom\_smooth(method = "lm", se = FALSE) +

geom\_vline(xintercept = input$age, linetype = "dashed") +

geom\_segment(aes(x = input$age, xend = input$age, y = 0, yend = predictions()), linetype = "dashed", color = "red") +

labs(x = "Age (years)", y = "Cr", title = "Cr vs Age") +

theme\_bw()

})

}

shinyApp(ui, server)

**6. Conclusion**

In conclusion, data analysis plays a pivotal role in our understanding of diabetes, enabling researchers, healthcare professionals, and policymakers to derive valuable insights for improved prevention, diagnosis, and treatment strategies. The comprehensive examination of diabetes-related data provides a foundation for identifying risk factors, tracking trends, and tailoring interventions to individual needs. Leveraging advanced analytics, machine learning, and data-driven approaches offers the potential to uncover hidden patterns and enhance predictive capabilities, fostering a more personalized and effective approach to diabetes management. As we continue to harness the power of data, collaboration between the scientific community, healthcare practitioners, and technology innovators will be essential in advancing our collective efforts to combat the global diabetes epidemic and enhance the quality of life for those affected by this condition.