VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

Project Submission for
Probability and Statistics Lab [BMAT202P]
Titanic Survival Probability Calculator
Lab Slot: L9 + L10

SUBMITTED BY:

1.**Shreya Gantayat (22BLC1005)**

2.**Niket Girdhar (22BLC1044)**

3.**Yuganshu Kumar (22BLC1370)**

4.**Deeptanshu Bhattacharya (22BLC1244)**

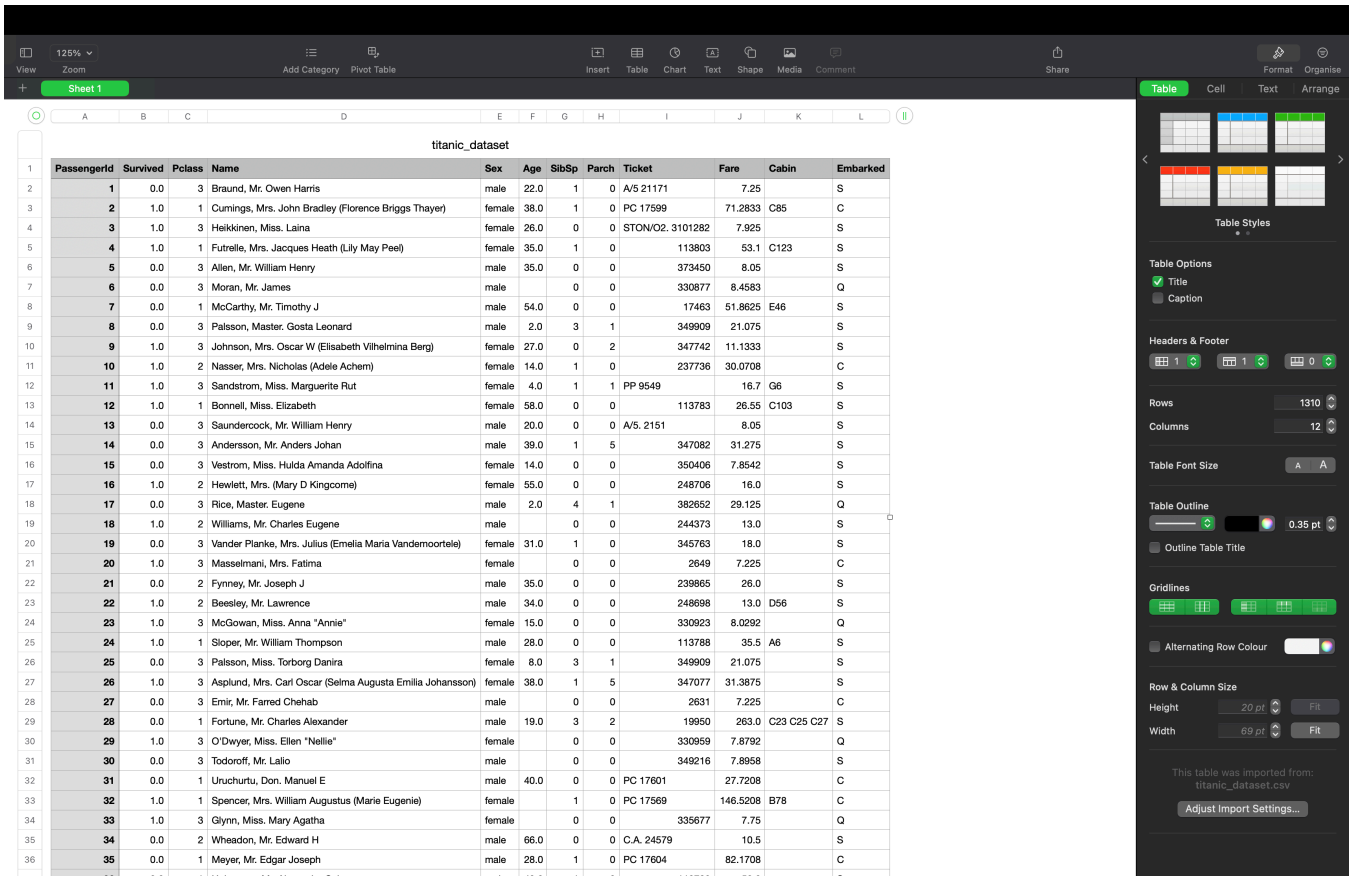
SUBMITTED TO:

Dr. DURGAPRASAD P
durgaprasad.p@vit.ac.in

SOFTWARE USED : RSTUDIO

DATA SOURCE : KAGGLE.COM

This dataset includes various information such as passenger details, survival status, class, and boarding information. The dataset contains 1,309 entries.



titanic_dataset

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0.0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.25		S
2	1.0	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38.0	1	0	PC 17599	71.2833	C85	C
3	1.0	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.925		S
4	1.0	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1	C123	S
5	0.0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.05		S
6	0.0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0.0	1	McCarthy, Mr. Timothy J	male	54.0	0	0	17463	51.8625	E46	S
8	0.0	3	Palsson, Master. Gosta Leonard	male	2.0	3	1	349909	21.075		S
9	1.0	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27.0	0	2	347742	11.1333		S
10	1.0	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14.0	1	0	237736	30.0708		C
11	1.0	3	Sandstrom, Miss. Marguerite Rut	female	4.0	1	1	PP 9549	16.7	G6	S
12	1.0	1	Bonnell, Miss. Elizabeth	female	58.0	0	0	113783	26.55	C103	S
13	0.0	3	Saundercock, Mr. William Henry	male	20.0	0	0	A/5. 2151	8.05		S
14	0.0	3	Andersson, Mr. Anders Johan	male	39.0	1	5	347082	31.275		S
15	0.0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	14.0	0	0	350406	7.8542		S
16	1.0	2	Hewlett, Mrs. (Mary D Kingcome)	female	55.0	0	0	248706	16.0		S
17	0.0	3	Rice, Master. Eugene	male	2.0	4	1	382652	29.125		Q
18	1.0	2	Williams, Mr. Charles Eugene	male		0	0	244373	13.0		S
19	0.0	3	Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)	female	31.0	1	0	345763	18.0		S
20	1.0	3	Masselmani, Mrs. Fatima	female		0	0	2649	7.225		C
21	0.0	2	Fynney, Mr. Joseph J	male	35.0	0	0	239865	26.0		S
22	1.0	2	Beesley, Mr. Lawrence	male	34.0	0	0	248698	13.0	D56	S
23	1.0	3	McGowan, Miss. Anna "Annie"	female	15.0	0	0	330923	8.0292		Q
24	1.0	1	Sloper, Mr. William Thompson	male	28.0	0	0	113788	35.5	A6	S
25	0.0	3	Palsson, Miss. Torborg Danira	female	8.0	3	1	349909	21.075		S
26	1.0	3	Asplund, Mrs. Carl Oscar (Selma Augusta Emilia Johansson)	female	38.0	1	5	347077	31.3875		S
27	0.0	3	Emir, Mr. Farred Chehab	male		0	0	2631	7.225		C
28	0.0	1	Fortune, Mr. Charles Alexander	male	19.0	3	2	19950	263.0	C23 C25 C27	S
29	1.0	3	O'Dwyer, Miss. Ellen "Nellie"	female		0	0	330959	7.8792		Q
30	0.0	3	Todoroff, Mr. Lalio	male		0	0	349216	7.8958		S
31	0.0	1	Uruchurtu, Don. Manuel E	male	40.0	0	0	PC 17601	27.7208		C
32	1.0	1	Spencer, Mrs. William Augustus (Marie Eugenie)	female		1	0	PC 17569	146.5208	B78	C
33	1.0	3	Glynn, Miss. Mary Agatha	female		0	0	335677	7.75		Q
34	0.0	2	Wheadon, Mr. Edward H	male	66.0	0	0	C.A. 24579	10.5		S
35	0.0	1	Meyer, Mr. Edgar Joseph	male	28.0	1	0	PC 17604	82.1708		C

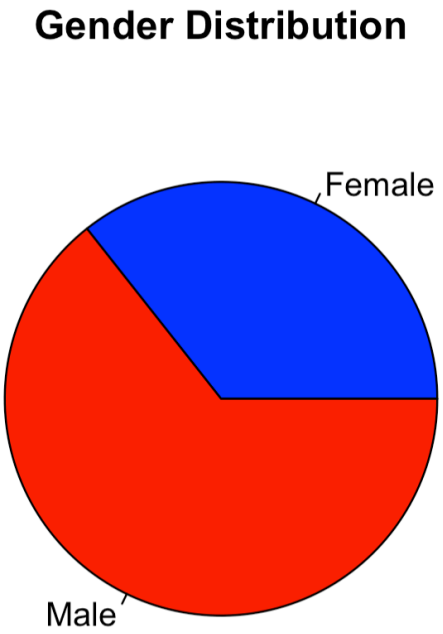
NOTE:

V1 = Passenger ID
V2 = Survival Status (1 = Survived, 0 = Did not survive)
V3 = Passenger Class (1st, 2nd, or 3rd class)
V4 = Name
V5 = Sex (Gender of the passenger)
V6 = Age
V7 = Number of Siblings/Spouses Aboard (SibSp)
V8 = Number of Parents/Children Aboard (Parch)
V9 = Ticket Number
V10 = Fare (Ticket fare)
V11 = Cabin
V12 = Embarked (Port of Embarkation: C = Cherbourg, Q = Queenstown, S = Southampton)

Insights using the Dataset:

Gender Distribution:

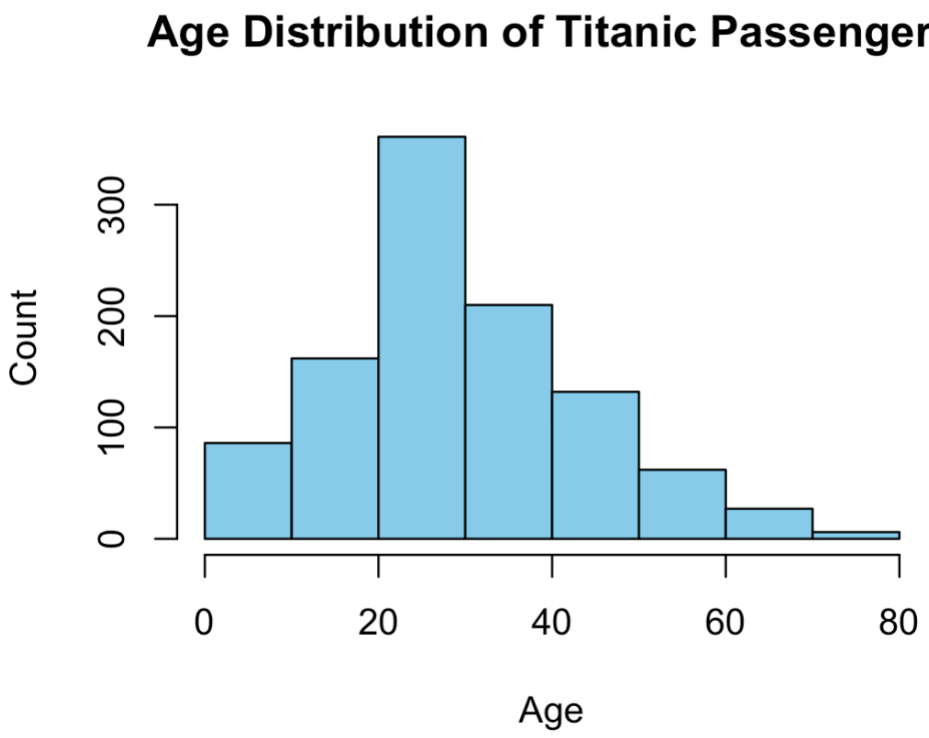
```
> sex = titanic_dataset$...5
> sex <- ifelse(sex == 'male', 1, 0)
> counts <- table(sex)
> labels <- c("Female","Male")
> colors <- c("blue","red")
> pie(counts,labels = labels,main = "Gender Distribution", col = colors)
>
```



Age Distribution:

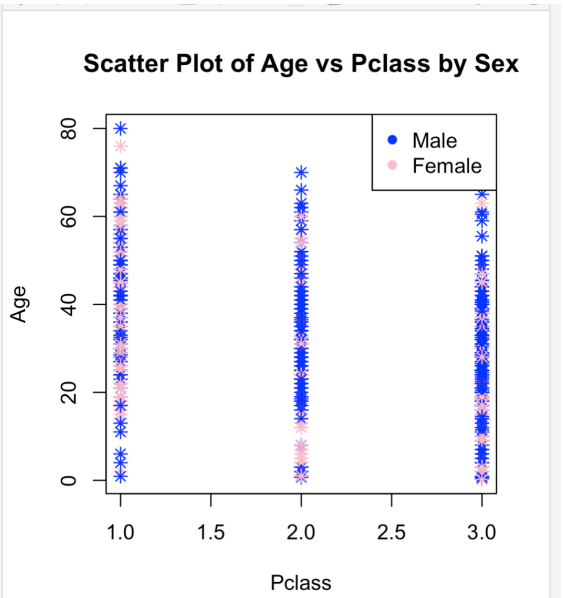
```
> age = titanic_dataset$...6
> summary(age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.17   21.00   28.00   29.88   39.00   80.00     263
>
```

```
> age = titanic_dataset$...6
> hist(age, breaks = 10,col = "skyblue", main = "Age Distribution
of Titanic Passengers", xlab = "Age",ylab = "Count")
>
```



Age vs Passenger_Class by Sex distribution:

```r
colors <- ifelse(titanic_dataset$...5 == "male", "blue", "pink")

plot(titanic_dataset$...3, titanic_dataset$...6, col = colors,pch = 8, xlab = "Pclass", ylab = "Age",
     main = "Scatter Plot of Age vs Pclass by Sex", xaxt = "n")

axis(1, at = 1:3, labels = c("1st Class", "2nd Class", "3rd Class"))

legend("topright", legend = c("Male", "Female"), col = c("blue", "pink"), pch = 16)

```



About our Project:

The Titanic Survival Prediction app, an interactive tool that allows users to explore survival predictions based on key factors that impacted the passengers' fates on the Titanic. Using real historical data, this shiny-based app helps illustrate the relationship between a passenger's characteristics and their likelihood of survival.

How it works?

The app is powered by a logistic regression model, trained on the famous Titanic dataset, to estimate the probability of survival based on:

- **Ticket Class**: First, Second, or Third class ticket
- **Gender**: Male or Female
- **Age**: Passenger's age at the time of travel
- **Port of Embarkation**: Cherbourg, Queenstown, or Southampton

Users can input these details and press "Check Survival" to get a predicted probability and survival outcome (Yes/No) based on historical patterns.

Visual Insights:

In addition to prediction, this app offers graphical insights into survival trends:

- **Survival Rate by Ticket Class**: A bar chart displaying survival rates across ticket classes.
- **Survival Rate by Gender and Class**: A faceted plot comparing survival rates based on gender within each ticket class.
- **Age Distribution by Class and Survival**: A boxplot showing the age distribution among survivors and non-survivors across classes.

Probability Concepts applied:

Probability Concepts Applied

This project applies several probability concepts, including:

- **Conditional Probability**: The probability of survival given certain characteristics (e.g., being in first class).
- **Logistic Regression**: An application of conditional probability where the likelihood of an outcome is modeled using a log-odds function.
- **Binary Classification Thresholding**: Probabilities are mapped to binary outcomes using a 0.5 threshold, which allows for interpreting results within the framework of survival likelihood.

Methodology:

Data Cleaning and Preprocessing: The Titanic dataset is first preprocessed by removing any missing data, and categorical variables (like ticket class and port of embarkation) are converted into factors.

Logistic Regression Model: The logistic regression model is chosen for its suitability in estimating binary outcomes (survival or not). The model's response variable is survival, while the predictor variables include:

- **Ticket Class** (Pclass)
- **Gender** (Sex)
- **Age**
- **Port of Embarkation** (Embarked)

Probability Calculation: The model calculates a survival probability between 0 and 1 for each input scenario. The threshold is set at 0.5, meaning predictions with a probability greater than 0.5 are classified as “Yes” (survived) and those below as “No” (did not survive).

Visualization: The app provides graphical insights into survival rates, leveraging the probabilistic outcomes to explore historical survival trends. Visualization tools used include:

- **Survival Rate by Ticket Class**: Bar chart showing class-wise survival probability.
- **Survival Rate by Gender and Class**: Faceted bar chart to observe class and gender interactions in survival.
- **Age Distribution by Class and Survival**: Boxplot to illustrate how age varied across classes and between survivors and non-survivors.

Code:

```
library(shiny)
library(dplyr)
library(ggplot2)

titanic_data <- read.csv("titanic_dataset.csv")
titanic_data <- na.omit(titanic_data)
titanic_data$Pclass <- factor(titanic_data$Pclass, levels = c(1, 2, 3))
titanic_data$Sex <- factor(titanic_data$Sex, levels = c("male", "female"))
titanic_data$Embarked <- factor(titanic_data$Embarked, levels = c("C", "Q", "S"))
titanic_data$Survived <- factor(titanic_data$Survived, levels = c(0, 1), labels = c("No", "Yes"))

model <- glm(Survived ~ Pclass + Sex + Age + Embarked, data = titanic_data, family = binomial)

ui <- fluidPage(
  titlePanel("Titanic Survival Prediction"),
  sidebarLayout(
    sidebarPanel(
      selectInput("pclass", "Ticket Class", choices = c("1st" = 1, "2nd" = 2, "3rd" = 3)),
      selectInput("sex", "Sex", choices = c("Male" = "male", "Female" = "female")),
      numericInput("age", "Age", value = 30, min = 0, max = 100),
      selectInput("embarked", "Port of Embarkation", choices = c("Cherbourg" = "C",
"Queenstown" = "Q", "Southampton" = "S")),
      actionButton("predict", "Check Survival")
    ),
    mainPanel(
      textOutput("result"),
      verbatimTextOutput("debug_info"),
      fluidRow(
        column(width = 6, plotOutput("gender_class_plot")),
        column(width = 6, plotOutput("age_class_plot"))
```

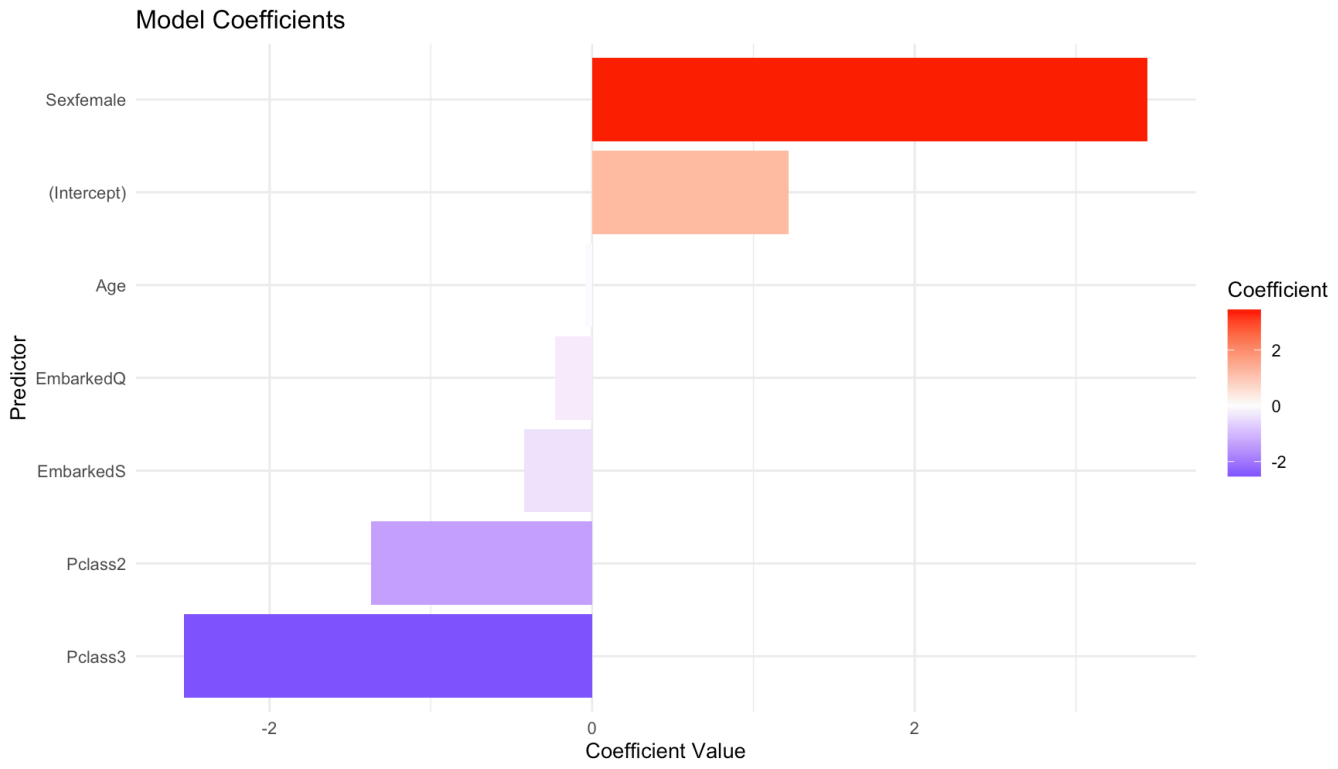
```
    ),
    plotOutput("ticket_class_plot")
   )
  )
)

server <- function(input, output) {

  observeEvent(input$predict, {
   user_data <- data.frame(
    Pclass = factor(as.integer(input$pclass), levels = levels(titanic_data$Pclass)),
    Sex = factor(input$sex, levels = levels(titanic_data$Sex)),
    Age = as.numeric(input$age),
    Embarked = factor(input$embarked, levels = levels(titanic_data$Embarked))
   )
   prediction <- predict(model, newdata = user_data, type = "response")
   survival <- ifelse(prediction > 0.5, "Yes", "No")
   survival_num <- ifelse(survival == "Yes", 1, 0)

   output$result <- renderText({
    paste("Predicted Survival:", survival, "(Probability:", round(prediction * 100, 2), "%)")
   })

   output$ticket_class_plot <- renderPlot({
    user_point <- data.frame(
     Pclass = factor(as.integer(input$pclass), levels = levels(titanic_data$Pclass)),
     Survived = factor(survival, levels = c("No", "Yes"))
    )
    ggplot(titanic_data, aes(x = Pclass, fill = Survived)) +
     geom_bar(position = "fill") +
     labs(title = "Survival Rate by Ticket Class", x = "Ticket Class", y = "Proportion") +
     scale_fill_manual(values = c("No" = "red", "Yes" = "green")) +
```

```
      theme_minimal() +
      geom_point(data = user_point, aes(x = Pclass, y = survival_num), color = "black", size = 3, shape = 4)
    })
  })

  output$gender_class_plot <- renderPlot({
    ggplot(titanic_data, aes(x = Sex, fill = Survived)) +
      geom_bar(position = "fill") +
      facet_wrap(~ Pclass) +
      labs(title = "Survival Rate by Gender and Class", x = "Gender", y = "Proportion") +
      scale_fill_manual(values = c("No" = "red", "Yes" = "green")) +
      theme_minimal()
  })

  output$age_class_plot <- renderPlot({
    ggplot(titanic_data, aes(x = Pclass, y = Age, fill = Survived)) +
      geom_boxplot(alpha = 0.6) +
      labs(title = "Age Distribution by Ticket Class and Survival", x = "Ticket Class", y = "Age") +
      scale_fill_manual(values = c("No" = "red", "Yes" = "green")) +
      theme_minimal()
  })

  output$debug_info <- renderPrint({
    user_data <- data.frame(
      Pclass = factor(as.integer(input$pclass), levels = levels(titanic_data$Pclass)),
      Sex = factor(input$sex, levels = levels(titanic_data$Sex)),
      Age = as.numeric(input$age),
      Embarked = factor(input$embarked, levels = levels(titanic_data$Embarked))
    )
    print(user_data)
  })
```

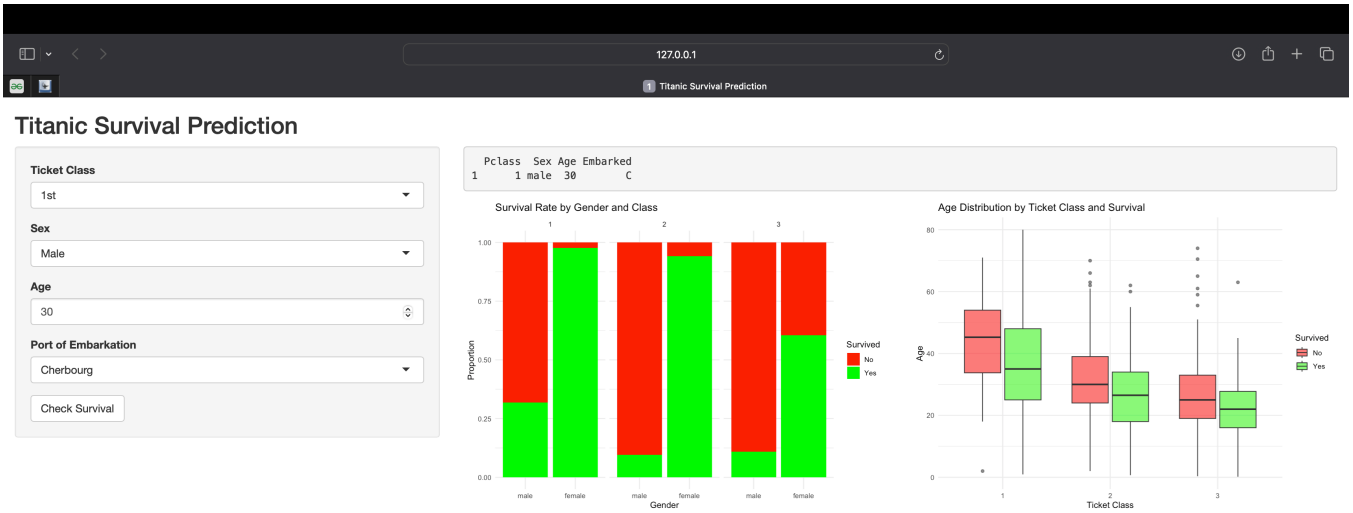
}

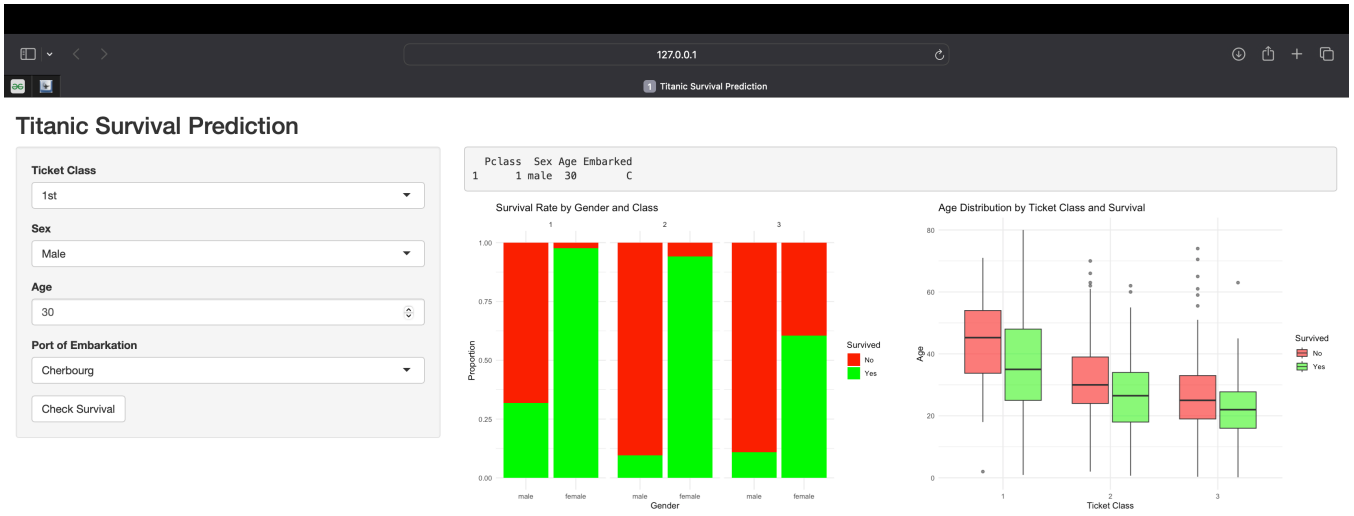
shinyApp(ui = ui, server = server)

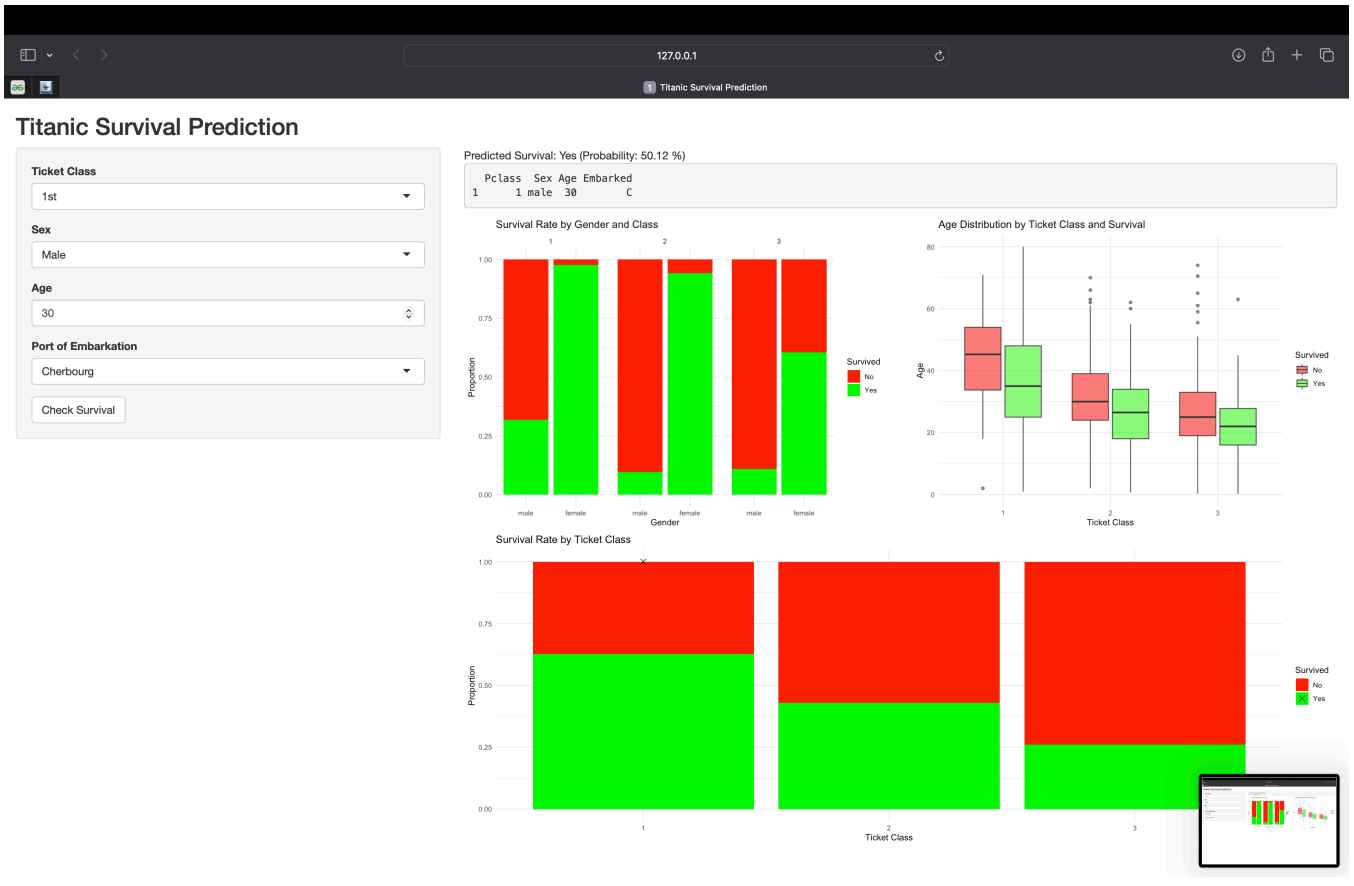
Model Coeffient relationship:



Images of the APP:




127.0.0.1
Titanic Survival Prediction
Titanic Survival Prediction
Ticket Class
1st
Sex
Male
Age
30
Port of Embarkation
Cherbourg
Check Survival
Pclass Sex Age Embarked
1 1 male 30 C
Survival Rate by Gender and Class
Proportion
Gender
male
female
Survived
No
Yes
Age Distribution by Ticket Class and Survival
Age
Ticket Class


127.0.0.1
Titanic Survival Prediction
Titanic Survival Prediction
Ticket Class
1st
Sex
Male
Age
30
Port of Embarkation
Cherbourg
Check Survival
Predicted Survival: Yes (Probability: 50.12 %)
Pclass Sex Age Embarked
1 1 male 30 C
Survival Rate by Gender and Class
Proportion
Gender
Survived
No
Yes
Age Distribution by Ticket Class and Survival
Age
Ticket Class
Survival Rate by Ticket Class

127.0.0.1

Titanic Survival Prediction

Titanic Survival Prediction

Ticket Class

2nd

Sex

Male

Age

30

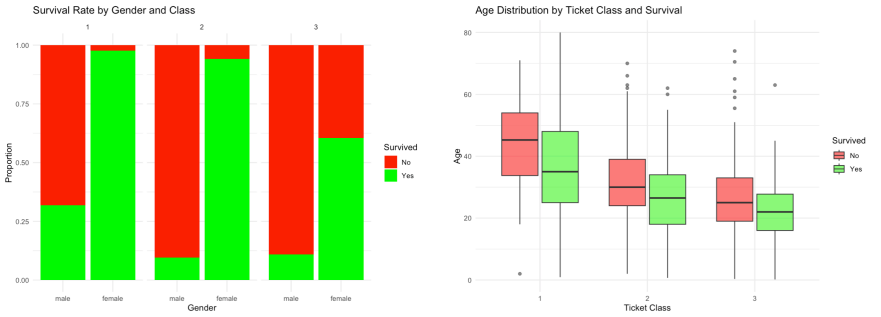
Port of Embarkation

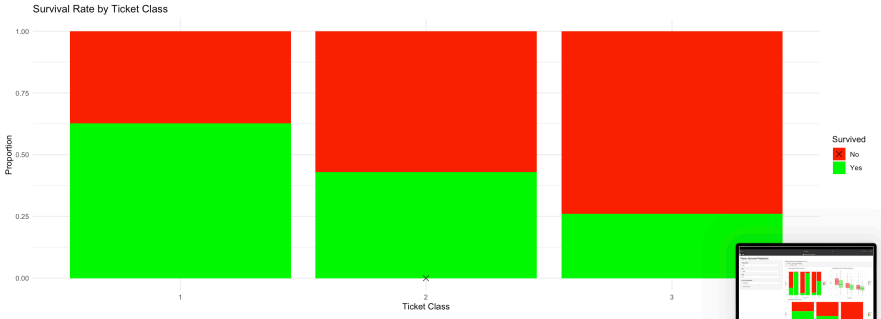
Cherbourg

Check Survival

Predicted Survival: No (Probability: 20.29 %)

```
  Pclass  Sex Age Embarked
1      2 male  30        C
```





VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

Thank You

SUBMITTED BY:

1. **Shreya Gantayat (22BLC1005)**
2. **Niket Girdhar (22BLC1044)**
3. **Yuganshu Kumar (22BLC1370)**
4. **Deeptanshu Bhattacharya (22BLC1244)**

Project