June, 2024

REPORT

STATISTICS

Group Members

COHNDSE233F -084

COHNDSE233F-085

COHNDSE233F-086

COHNDSE233F-O87

Statistics for Computing Assessment - 04

Higher National Diploma in Software Engineering 23.3F

Risk Prediction of Survival of a Shipwreck

Table of Contents

1. Introduction

The purpose of this report is to investigate and forecast the passenger survival in a shipwreck using the famous incident Titanic shipwreck dataset.

By grasping the correlations between different factors, we aim to uncover insights that could prove passenger survival rates.

In this research,

i. **what sorts of people were more likely to survive?**
ii. **What factors are associated with passenger survival?**
iii. **How does passenger class affect survival rates?**
iv. **Does age influence the likelihood of survival?**
v. **Are there any differences in survival based on gender?**

Objectives,

i. **To determine the relationship between passenger demographics (age, gender) and survival.**
ii. **To assess the impact of socio-economic status (represented by passenger class) on survival rates.**
iii. **To evaluate the influence of family presence (sibsp, parch) on survival chances.**
iv. **To identify key predictors of survival using a logistic regression model.**

Variable Description,

I. **Nominal – Survival / Name / sex / Ticket Number / cabin / port of Embarkation / Lifeboat / Address**
II. **Ordinal – Pclass**
III. **Continuous – Age / Passenger Fare**
IV. **Discrete – Amount of siblings / Number of Parents / Body Number**

2. Data Analysis

- Bar plots provided insights into the distribution of categorical variables within the dataset, such as gender or age, enabling us to understand their prevalence and composition
- Histograms offered detailed views of the distribution of individual variables, highlighting their skewness, central tendency, and variability.
- Pie charts succinctly summarized the proportions of categorical variables, offering a quick overview of factors.
- Lastly, box plots enabled comparisons of the distribution of numerical variables across different categories, helping us discern variations and trends among various factors.

3. Methodology

❖ Data Exploration and Cleaning

- Handling missing Values

  Identify and manage missing data by using imputation techniques or by removing rows and columns with excessive missing values

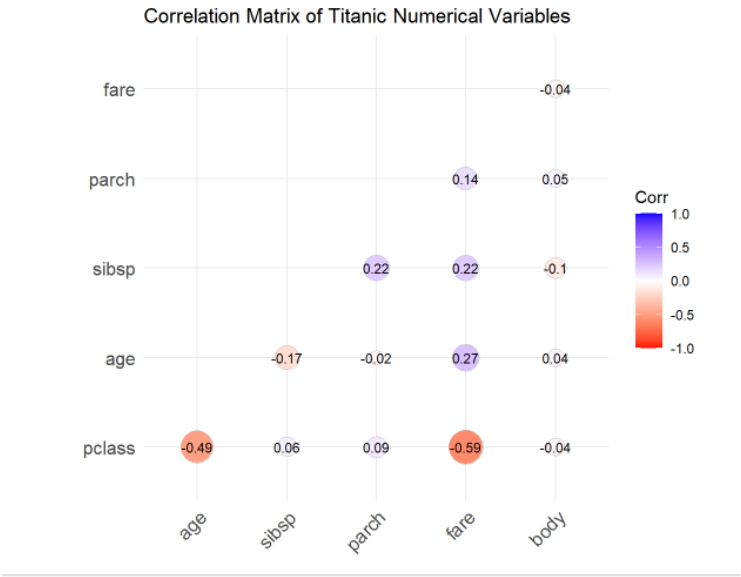
- Data Transformation

  Convert categorical variables into numerical ones if necessary (e.g., encoding gender as 0 and 1).

- Feature Engineering

  Create new features from existing ones (e.g., family size from `sibsp` and `parch`).

❖ Identifying Correlations

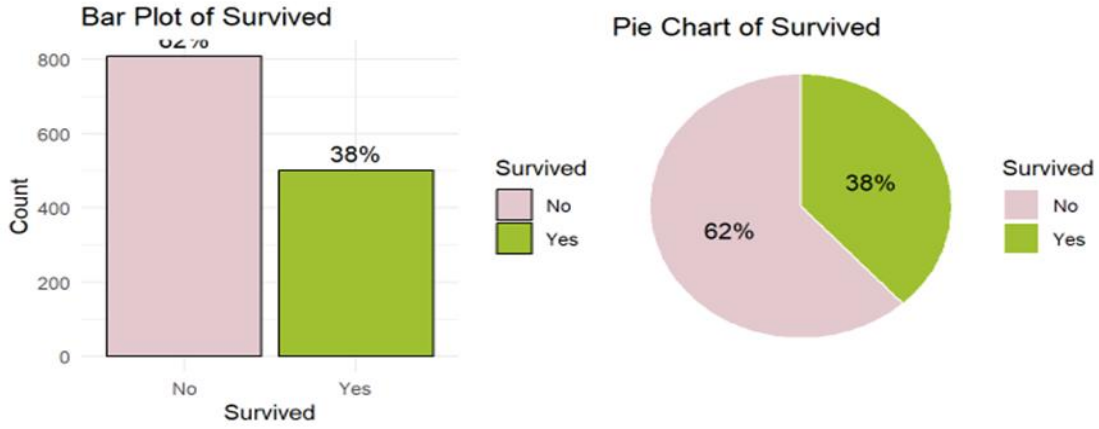


Now I need to know the strength of the correlations in between various variables and after i am able to visualize the relationships with graphs to get the idea of how they are related with each other. I will select the significant strong relationships according to the correlation coefficient.

4.Data Visualizations

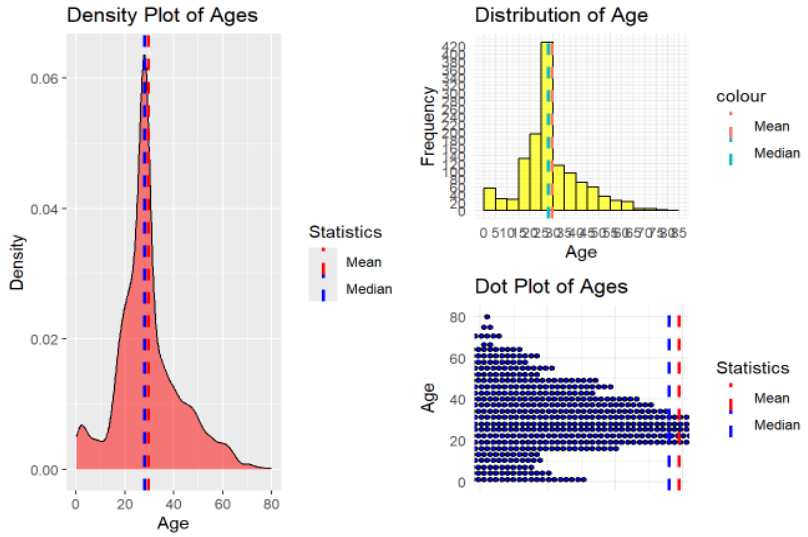
Through various visualization techniques such as bar plots, scatter plots, histograms, pie charts, stem-and-leaf plots, and box plots, we aim to gain insights into the relationships between different variables and the passenger survival rate

01.



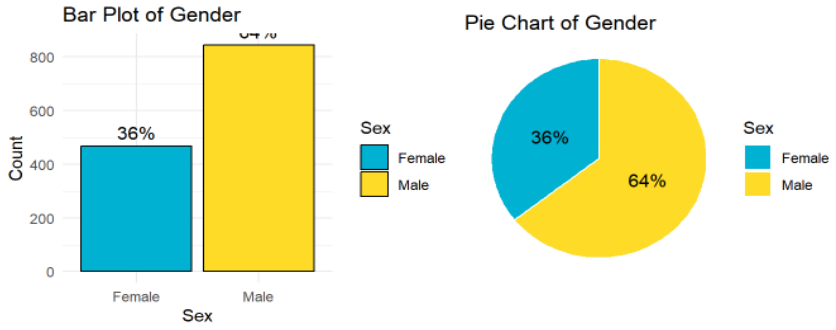
Comparing the bar graph and pie chart allows for different interpretations of the distribution of passenger survival count. The bar graph provides a clear visual comparison of counts, while the pie chart emphasizes the proportions of individuals who survived and not survived relative to the total population. From these visualizations, it is evident that a significant number of passengers did not survive.

02.



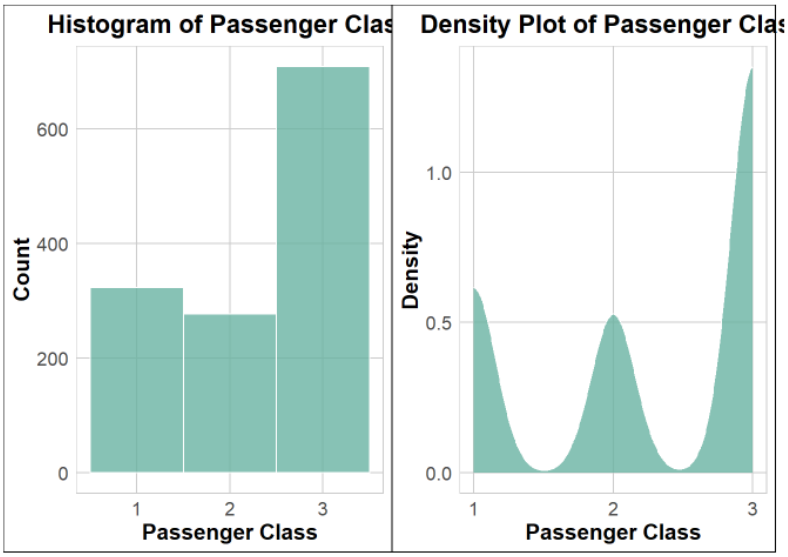
Visualization of age in the context of passenger survival analysis serves as a fundamental aspect of data exploration and understanding. Through various visualization techniques such as histograms, dot plots, and density plots, we can gain valuable insights into the distribution of age and its relationship with survival outcomes. Our analysis reveals that a significant number of passengers are between the ages of 20 and 40, highlighting a key demographic in the dataset.

03.



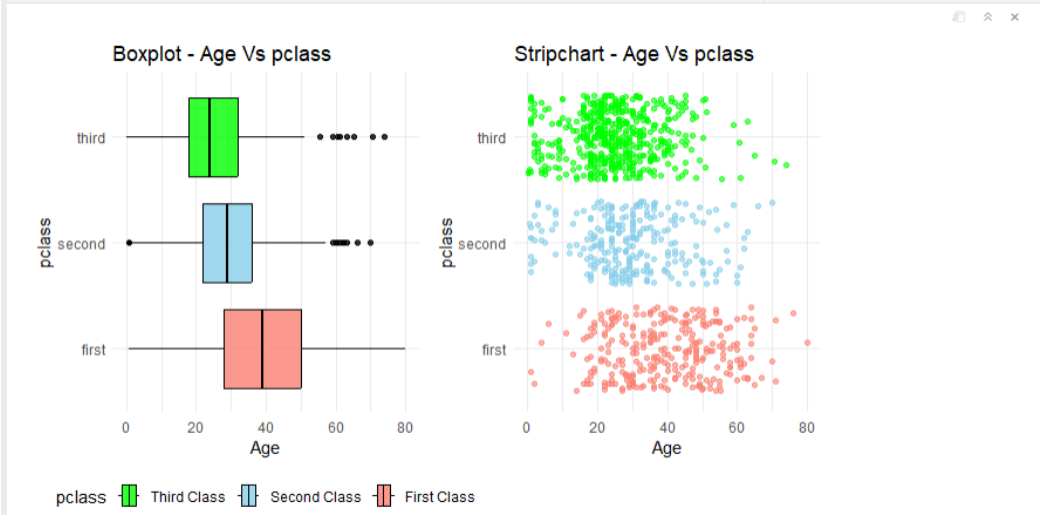
A pie chart & bar chart provides a simple and intuitive representation of the proportion of male and female passengers. We can take gender status and their count, offering a better understanding of its impact on passenger survival. From these visualizations, it is clear that there were many male passengers on the boat.

04.



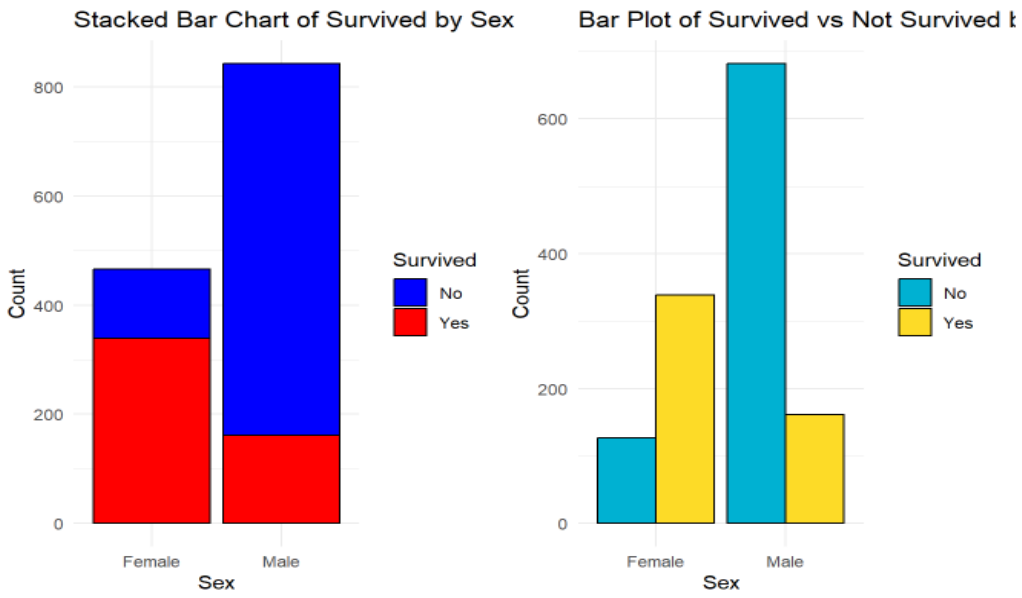
This histogram and the density plot show the passenger count and density in each passenger class. According to the graph we can see that there are approximately 300 passengers in the 1st class, 250 passengers in the 2nd class and 700 passengers in the 3rd class. The largest number of passengers are in the 3rd class. In the density plot we can see the highest peak is in the 3rd class indicating the high concentration of passengers.

05.



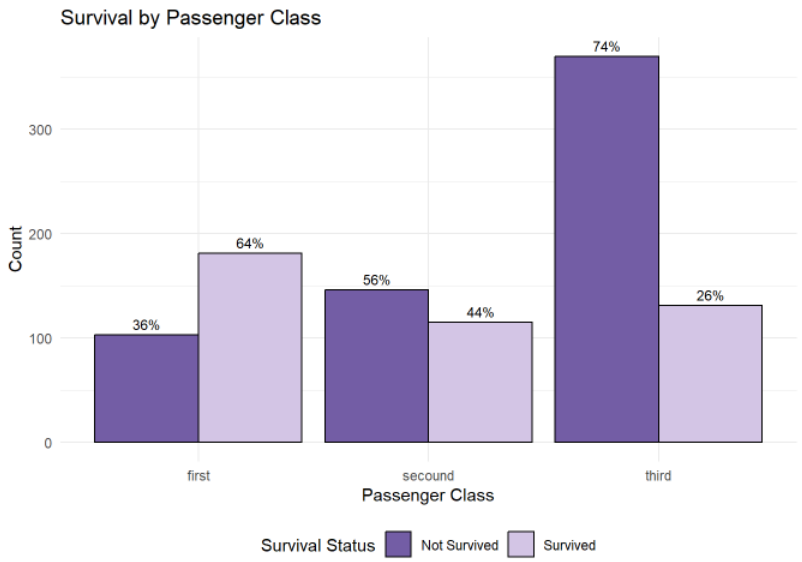
This box plot and the strip chart compare the passenger class and the age variables. By analyzing the box plot we can interpret that the 1st class median age is higher compared to other classes. This shows the socio-economic difference meaning wealthier passengers are older and occupying the 1st class. The strip chart shows the distribution of the age across a wider range to show more diversity. And show more younger passengers were occupying the 3rd class.

06.



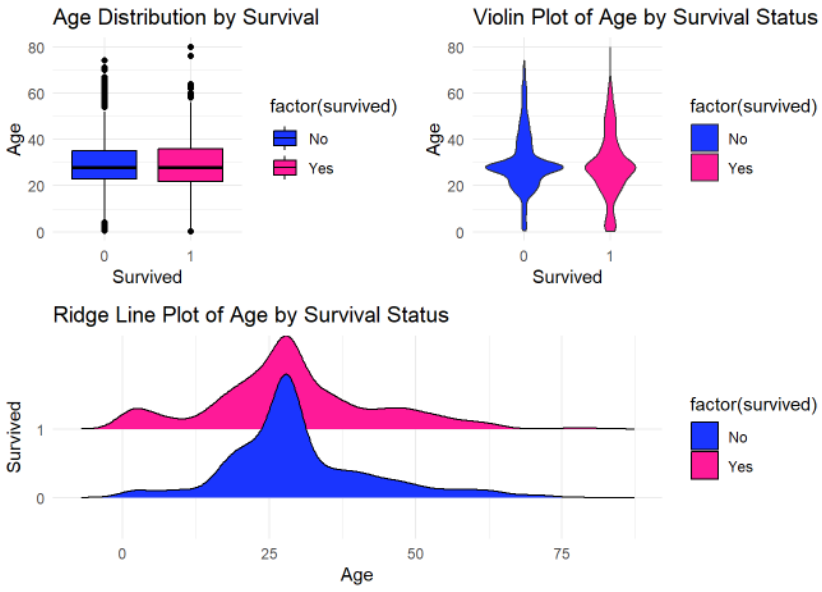
Each bar is divided into segments, with each segment representing the count of male and female count. By examining the stacked bar chart, we can easily compare the distribution of male and female count. Analyzing the stacked bar chart helps identify any disparities or patterns in the prevalence of male and female count who survived and not survived. It provides a clear visual representation of how gender impact the survival rate, thereby aiding in understanding the relationship between gender and passenger survival. The conclusion drawn from this analysis is that there were many more males overall, but a notable proportion of the male passengers didn't survive compared to the female passengers.

07.



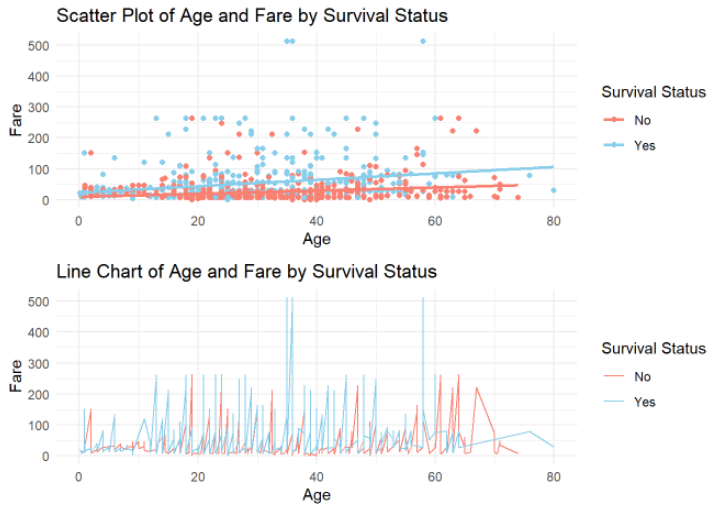
In this context of passenger class, the bar graph will display the proportions of passengers who survived and not survived based on passenger class. By using this bar graph, we can get a visual idea about the effectiveness in passenger class in passenger survival. The conclusion drawn from this analysis is that a lot of passengers in third class did not survive.

08.



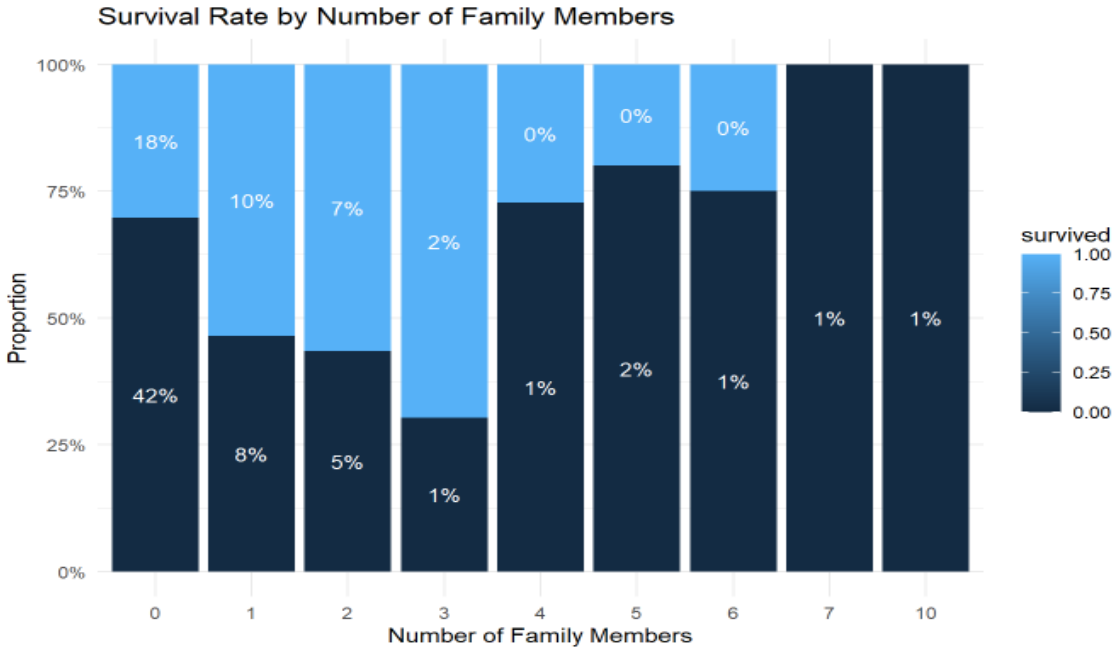
These graphs show that the age was a huge factor when considering about the passenger survival. In these graphs we can see the younger passengers were able to survive more than the older passengers showing that the age is an important factor in this kind of an incident.

09.



In this chart we compare between age and fare variables. According to this chart we can see that higher the fare higher the survival rate of the passengers.

10.



This stacked bar graph shows the survival rate based on the number of family members aboard. By analyzing this graph, we can see that passengers who has zero family members aboard survived more than the other categories.

5. Model Evaluation

```
cat("Mean Squared Error (MSE):", mse, "\n")
```

```
## Mean Squared Error (MSE): 0.2280496
```

```
cat("Root Mean Squared Error (RMSE):", rmse, "\n")
```

```
## Root Mean Squared Error (RMSE): 0.4775454
```

```
cat("R-Squared:", rsquared, "\n")
```

```
## R-Squared: -0.0177446
```

When evaluating a regression model, metrics such as R-squared (coefficient of determination), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) are commonly used to assess its performance.

R-Squared (R^2): The R-squared value measures the proportion of the variance in the dependent variable that is explained by the independent variables in the model. A higher R-squared value indicates a better fit of the model to the data. However, in my case, the R-squared value is negative (-0.0177446), which suggests that the model performs worse than a model that simply predicts the mean of the dependent variable.

Mean Squared Error (MSE): MSE measures the average squared difference between the actual and predicted values of the dependent variable. A lower MSE indicates better model performance. In my case, the MSE value is 0.228.

Root Mean Squared Error (RMSE): RMSE is the square root of the MSE and represents the average magnitude of the errors in the predicted values. Like MSE, a lower RMSE indicates better model performance. In my case, the RMSE value is 0.4775454.

Based on these evaluation metrics, it appears that the regression model has poor performance, as indicated by the negative R-squared value and relatively high MSE and RMSE values. This suggests that the model does not accurately predict the dependent variable and may require further refinement or the inclusion of additional features to improve its predictive capability

5.Prediction

Get the actual values of the target variable from t

```
# Get the actual values of the target variable from the test dataset
actual_values <- testing_set$diabetes

# Compare the predicted values with the actual values
comparison <- data.frame(Actual = actual_values, Predicted = predictions)
print(comparison)
```

```
##     Actual Predicted
## 2        0 0.5877725
## 3        0 0.4398513
## 12       0 0.3216994
## 15       0 0.5735277
## 19       0 0.3913103
## 38       1 0.2966691
## 44       1 0.4004125
## 47       0 0.5983897
## 49       0 0.2507610
```

These predictions are the model's estimates of the dependent variable (diabetes risk) for specific test data points. The "Actual" column represents the true values of the dependent variable, while the "Predicted" column shows the values predicted by the regression model.

For example:

- In the first row, the actual value of diabetes risk is 0, and the model predicts a value of approximately 0.588.
- In the second row, the actual value is 0, and the predicted value is approximately 0.440.

Similarly, for each row, there is a comparison between the actual and predicted values of diabetes risk.

These predictions can be evaluated further using metrics such as accuracy, precision, recall, and the F1 score to assess the model's performance on the test data. If the predicted values closely match the actual values, it indicates that the model is making accurate predictions. However, if there are large discrepancies between the actual and predicted values, it suggests that the model may need refinement or additional data to improve its predictive capability.