1. In the Data Set we have 15 attributes and 1309 examples.

Dataset:

Graphical user interface, application, Excel

Description automatically generated

From the statistics of the data as shown in below image. we can see 263 missing values in age, 564 missing values in Home/Destination,263 missing values in Midpoint age,1 missing value in fare,1014 missing values cabin,2 missing values in port,823 missing values in lifeboat,1188 missing values in body.

A screenshot of a computer

Description automatically generated

There are no duplicate values in the dataset. Below process will help us to identify duplicate values. When we run the below process, we can see that there are no duplicates in data set

Graphical user interface, text, application

Description automatically generated

2. From the Below graph we can see that there are 809 not survived passengers and 500 survived passengers.

Graphical user interface, chart

Description automatically generated

From the distribution below we can say that age range of Titanic passengers is as follows

Chart, histogram

Description automatically generated

Age range is from Minimum 0.167 to Maximum 80 Hence Range = 80-0.167=79.833 . From the distribution above Majority of titanic passengers are in the range of 15 to 25.

Age Distribution:

Chart, histogram

Description automatically generated

Survived Distribution:

Graphical user interface

Description automatically generated

1. From the image below we can say that there are 323 1st class passengers, 277 2nd class passengers, 709 3rd class passengers

Graphical user interface, chart, pie chart

Description automatically generated

Most common home destination of Titanic passengers is New York. Below process will be used to calculate mod of Home/Destination.

Graphical user interface, text, application

Description automatically generated

Output:

Graphical user interface, text, application, email

Description automatically generated

1. Below process is used for creation of logistic regression model

Graphical user interface, diagram

Description automatically generated

Step 1: we need to provide Titanic data as Input to the process

Step2: As described in question we need create model for only subset of attributes. So, I have used select Attribute operator to select only age, passenger class, sex and survived

Step3: From previous questions we get that there are no duplicates in datasets and there are missing values. I replaced missing values with average using replace missing values operator

Step3: To apply logistic regression model I have used Nominal to Binomial Type conversion operator for survived attribute

Step4: I have defined survived as label for prediction to the model using set role operator.

Step5: I have applied logistic regression model operator to apply Logistic regression on data set

Step6: Model simulator operator is used for model simulation

Another Approach: we can select the required attributes in turbo prep and create logistic regression model using Auto model

Graphical user interface

Description automatically generated

1. AUC score of the model is 82.4% as follows

Graphical user interface

Description automatically generated

1. From the simulator we can say that there is 96% chances that a female baby below 10 years old travelling as 1st class passenger will survive

Graphical user interface

Description automatically generated

1. From simulator we can say that there is 71% probability that a female baby below 10 years travelling as 3rd class passenger will survive

Graphical user interface

Description automatically generated

1. From simulator we can say that there is 38%chance that a male of 40 years old first-class passenger will get survived

Graphical user interface, application

Description automatically generated

1. From the simulator we can say that there is 7%chance that a male of 40 years old Third-class passenger will get survived

Graphical user interface, application

Description automatically generated

1. From this analysis I can say that passengers of younger age will get to most likely survived than the passengers of older ages.
2. Mosaic graph:

Graphical user interface, chart

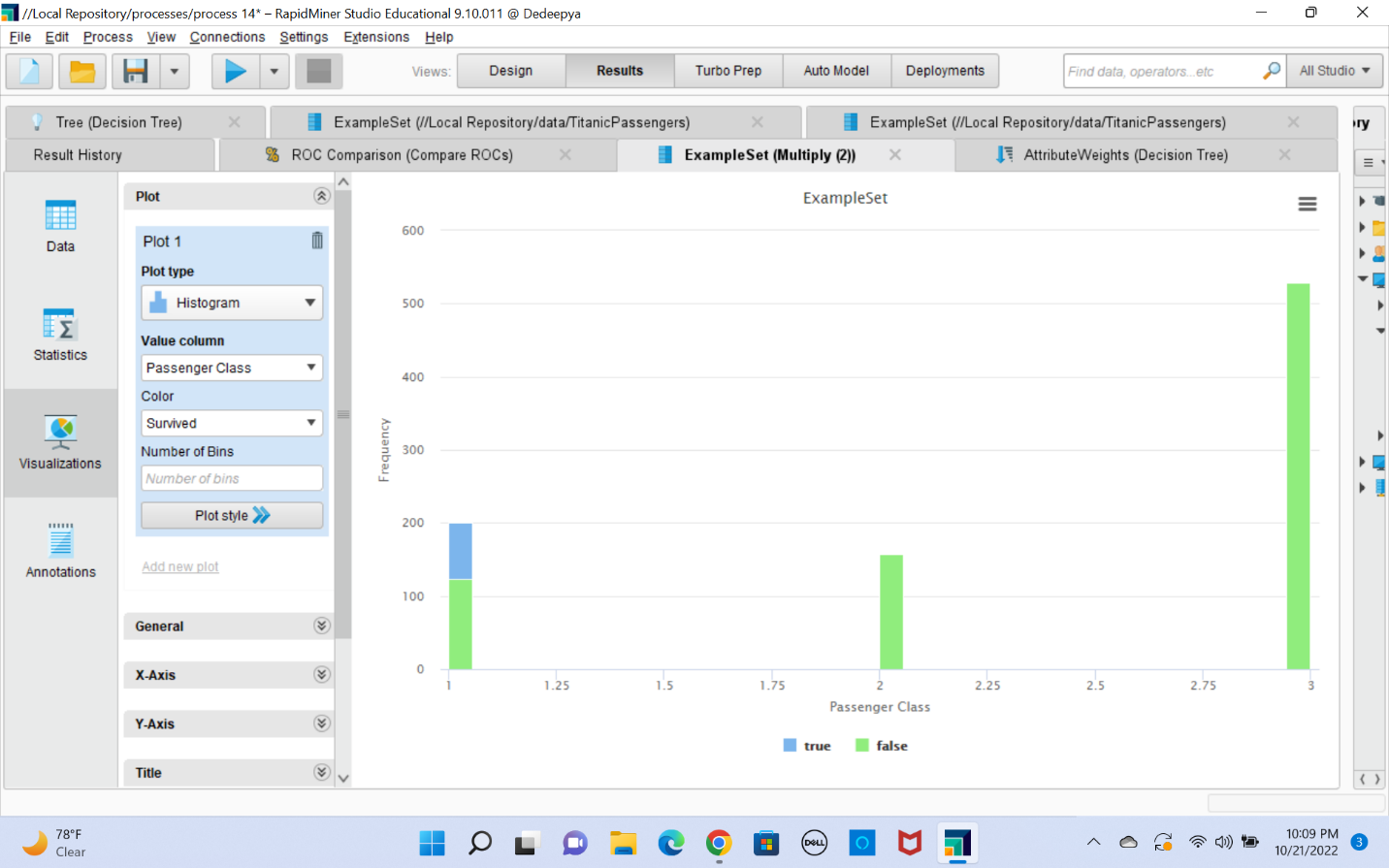
Description automatically generated

From the below graphs we can say that most people of age range 20-50 are not likely to survive.

Chart

Description automatically generated

From the below graph we can say that people from passenger class 2nd and 3rd are not likely to get survived



1. RapidMiner process create a decision-tree with 3 variables as below:

Diagram

Description automatically generated

Step 1: we need to provide Titanic data as Input to the process

Step2: As described in question we need create model for only subset of attributes. So, I have used select Attribute operator to select only age, passenger class, sex and survived

Step3: To apply decision tree model I have used Nominal to Binomial Type conversion operator for survived attribute

Step4: I have defined survived as label for prediction to the model using set role operator.

Step5: I have applied Decision tree model operator to apply Decision Tree on data set

Output:

Graphical user interface, text, Word

Description automatically generatedGraphical user interface, text, application, email

Description automatically generated

Another approach Decision tree using auto model is as follows:

Graphical user interface, text

Description automatically generated

1. To get AUC score of at least 0.80 as shown in below image. Total 7 splits are needed Graphical user interface, text

   Description automatically generated

Graphical user interface, text

Description automatically generated

B. Root node has sex and hence sex contributes the most to predicting “Survived” target class.

From the correlation matrix we can say that sex has highest correlation with Survived.

Correlation Matrix:Graphical user interface, text, application

Description automatically generated