AN INVESTIGATION TO INCREASING THE FLYING RATE

AIT 582

VENKATA JAYANDRA KUMAR LADE

G01046700

vlade@gmu.edu

1 TABLE OF CONTENTS

CONTENTS

2 INTRODUCTION

A Data Scientist for an airline A wants to analyze customer database. To find factors that are useful in understanding why some are cancelling. By that, scientist can recommend some factors to advertising team. So that they can create some packages and attract more customers.

3 DATA

Data set contains 6 predictor variables. The target variable is Success. Database consists of 892 customers' records.

Attribute Information

1) **Fare:** A numerical variable describes about total fare paid.
2) **Customer id:** A numerical variable with unique id for each customer.
3) **Description:** Description of each customer including name and age.
4) **Seat Class:** A categorical variable with three seat classes i.e, 1,2,3
5) **Guests:** A numerical variable describes about number of guests accompanying the main customer.
6) **Success:** A categorical variable describes about if the customer successfully flew on the booked trip.

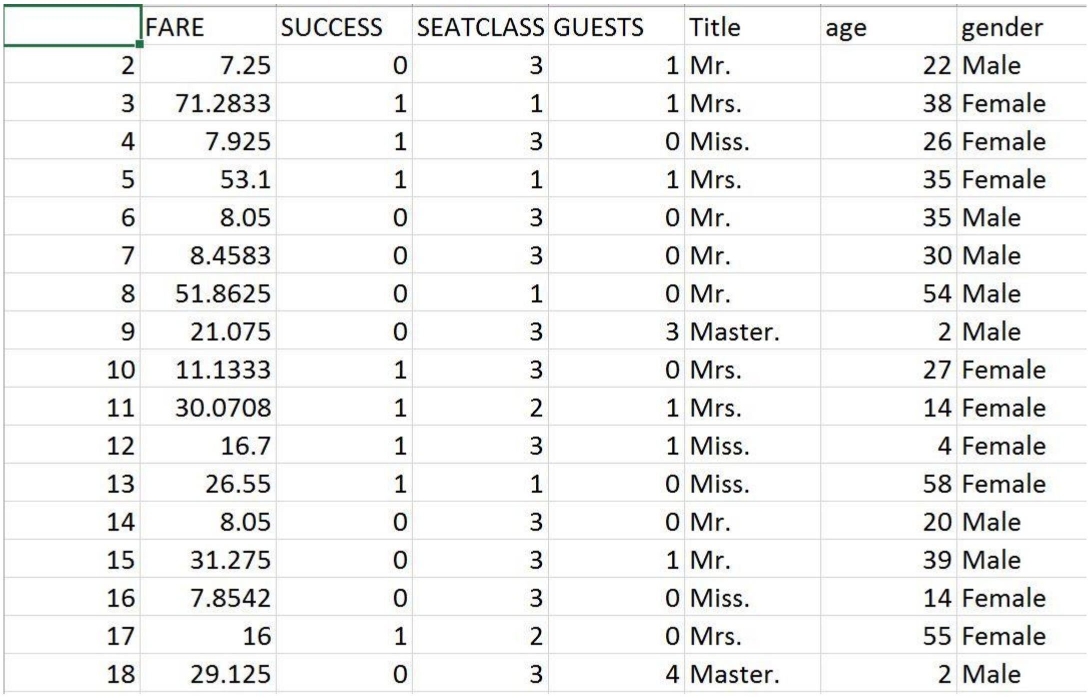
4 Milestone1 – Data Acquisition and Conversion.

- ✓ Data downloaded using R which is in JSON format.
- ✓ Data have been converted from semi-structured data into structured data and saved into CSV file.
- ✓ Converted Csv file is used in next step for metadata extraction.

5 Milestone2- Metadata Extraction and Imputation

- ✓ Meta data type that we observe in the Description field is descriptive metadata. Metadata such as age and gender can be derived from description field.
- ✓ Gender field is created based on the title in the description.
- ✓ The extracted age and gender are appended to original data.

Age have many missing values after extraction. For imputation of age, data is sub-divided with respective their titles. And Missing ages in each sub set are replaced with the median values. Some rows have been removed as gender cannot be determine for some title likes dr., rev. etc. Formatted data is shown below:

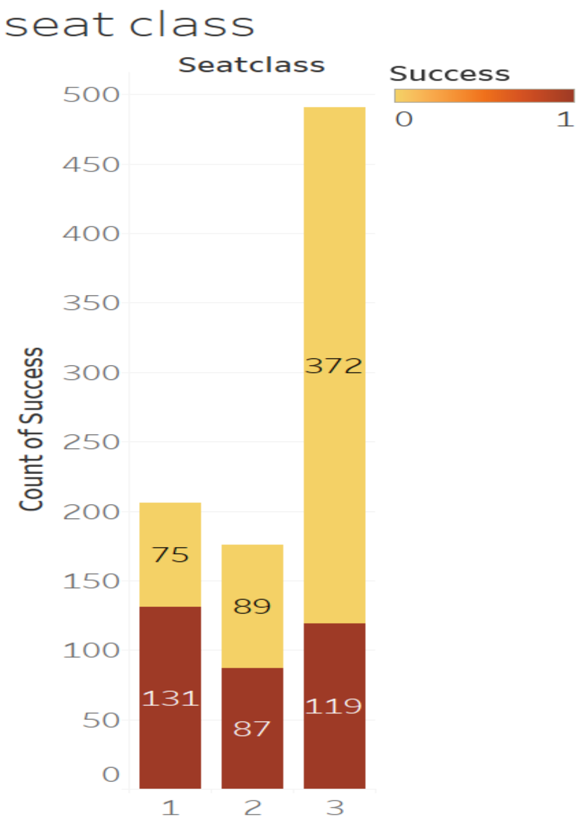


	FARE	SUCCESS	SEATCLASS	GUESTS	Title	age	gender
2	7.25	0	3	1	Mr.	22	Male
3	71.2833	1	1	1	Mrs.	38	Female
4	7.925	1	3	0	Miss.	26	Female
5	53.1	1	1	1	Mrs.	35	Female
6	8.05	0	3	0	Mr.	35	Male
7	8.4583	0	3	0	Mr.	30	Male
8	51.8625	0	1	0	Mr.	54	Male
9	21.075	0	3	3	Master.	2	Male
10	11.1333	1	3	0	Mrs.	27	Female
11	30.0708	1	2	1	Mrs.	14	Female
12	16.7	1	3	1	Miss.	4	Female
13	26.55	1	1	0	Miss.	58	Female
14	8.05	0	3	0	Mr.	20	Male
15	31.275	0	3	1	Mr.	39	Male
16	7.8542	0	3	0	Miss.	14	Female
17	16	1	2	0	Mrs.	55	Female
18	29.125	0	3	4	Master.	2	Male

6 Milestone 3- Metadata Exploration

Seat Class:

- ✓ Seat class 1 had highest number of success and seat class 3 have more number of failures.

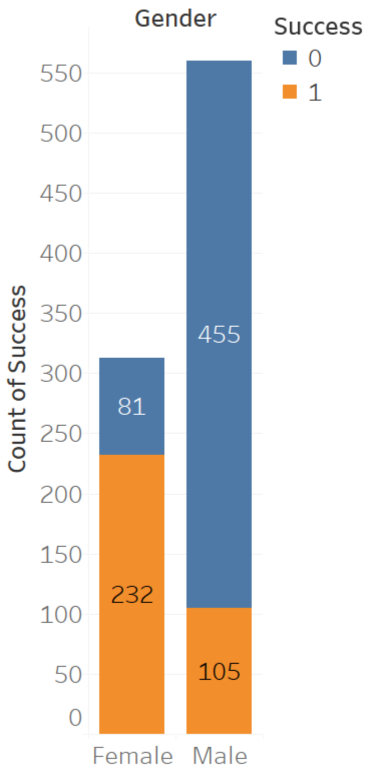


Count of Success for each Seatclass. Color shows details about Success. The marks are labeled by count of Success. The data is filtered on Success, which keeps 0 and 1.

Gender:

- ✓ Comparing male and females, females have more success rate than males.

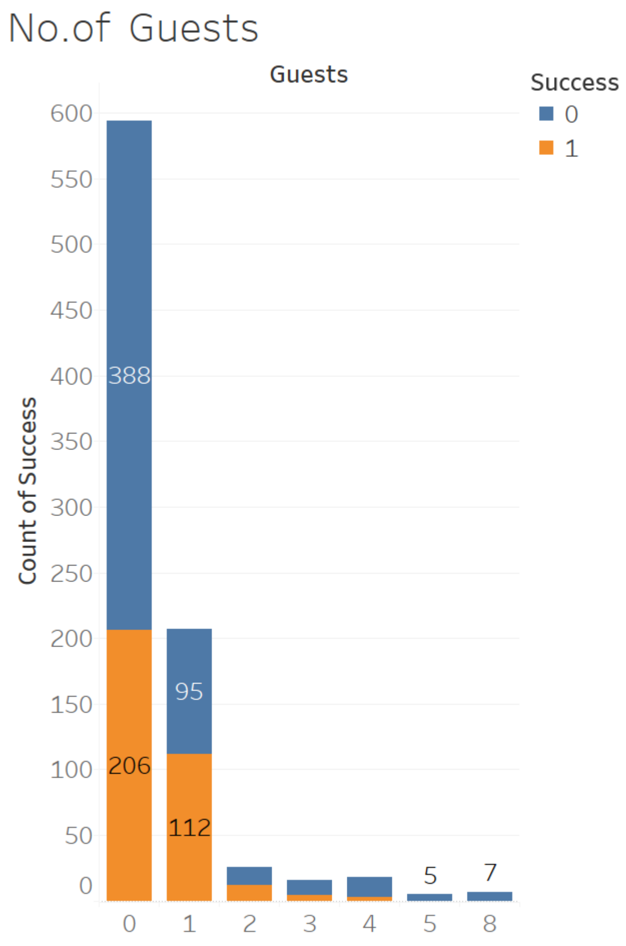
Gender



Count of Success for each Gender. Color shows details about Success. The marks are labeled by count of Success. The view is filtered on Success, which keeps 0 and 1.

Number of Guests:

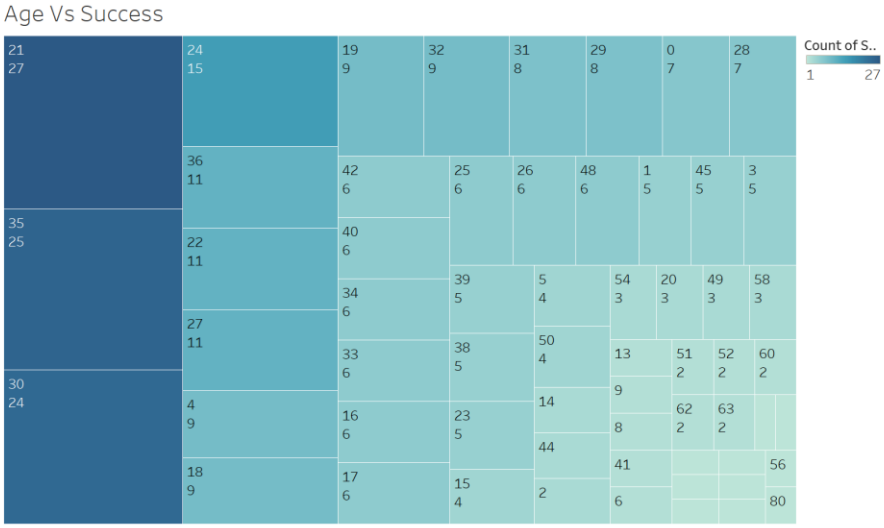
- ✓ Count of Number of successes are more in number of guests 0.
- ✓ But by considering percentages we can observe that number of guest 1 has high success rate.
- ✓ By clear observation, there is no success for the number of guests 5 and 8.



Count of Success for each Guests. Color shows details about Success. The marks are labeled by count of Success. The view is filtered on Success, which keeps 0 and 1.

Age V/s Success:

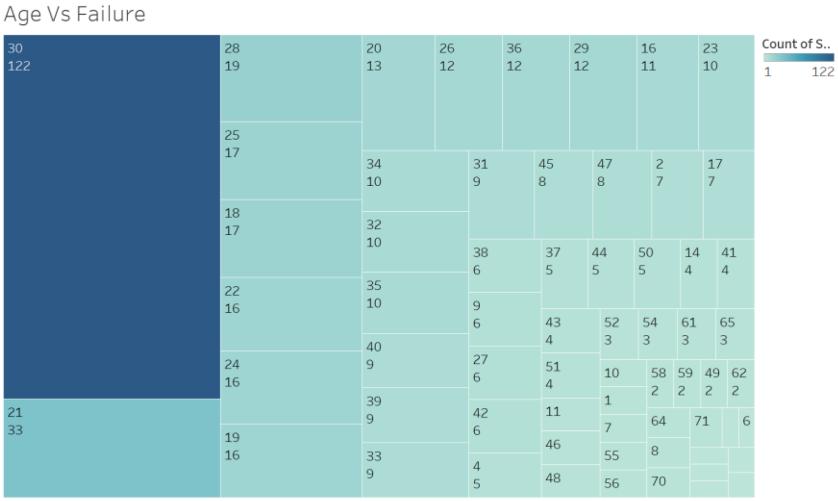
- ✓ Age 21 has high success in flying.


Age Vs Success

Age and count of Success. Color shows count of Success. Size shows count of Success. The marks are labeled by Age and count of Success. The data is filtered on Success, which keeps 1.

Age Vs Failure:

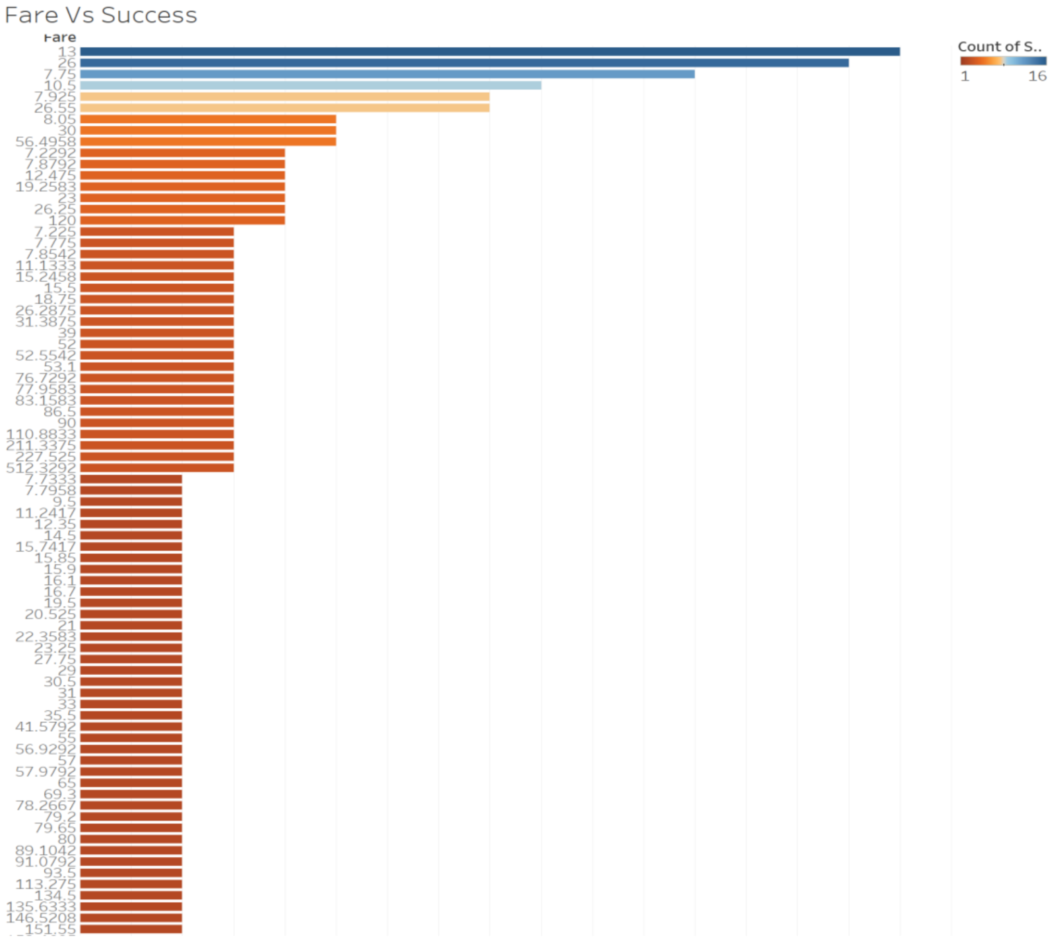
- ✓ Age 30 have more number of cancellations.


Age Vs Failure

Age and count of Success. Color shows count of Success. Size shows count of Success. The marks are labeled by Age and count of Success. The data is filtered on Success, which keeps 0.

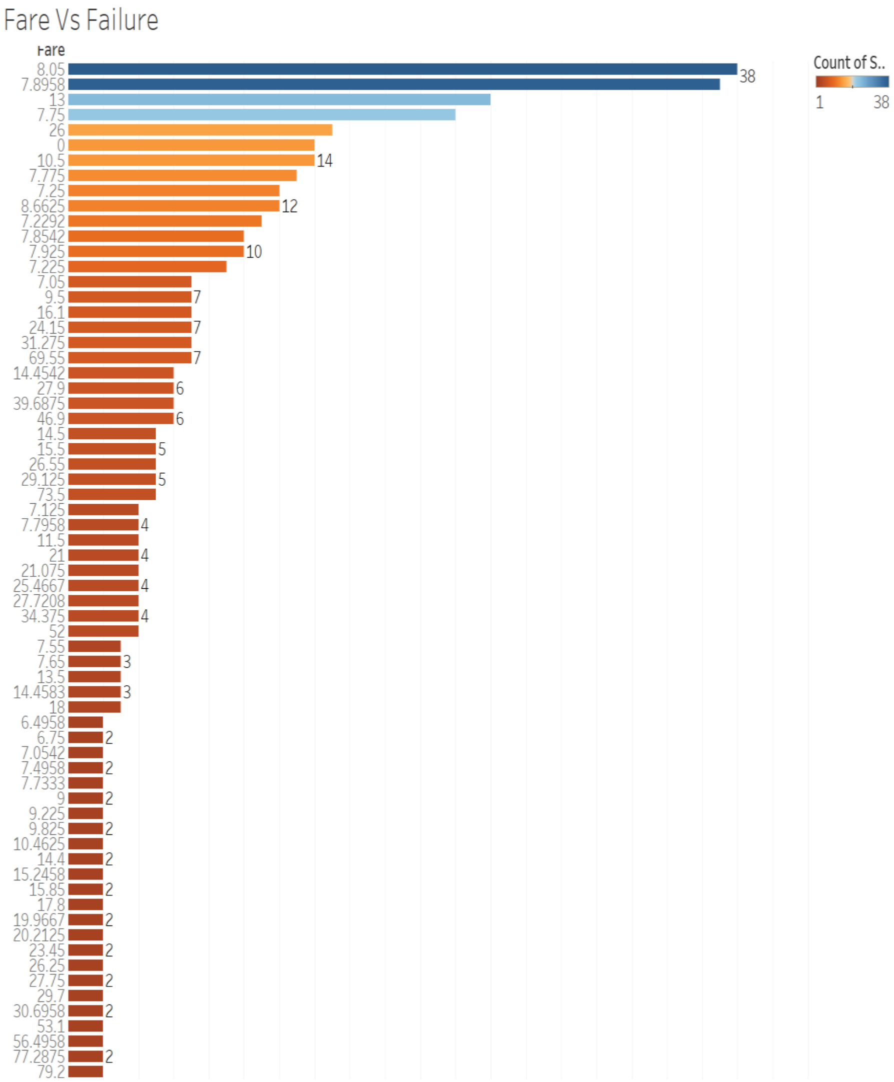
Fare Vs Success:

- ✓ Fare value 13 has highest success.



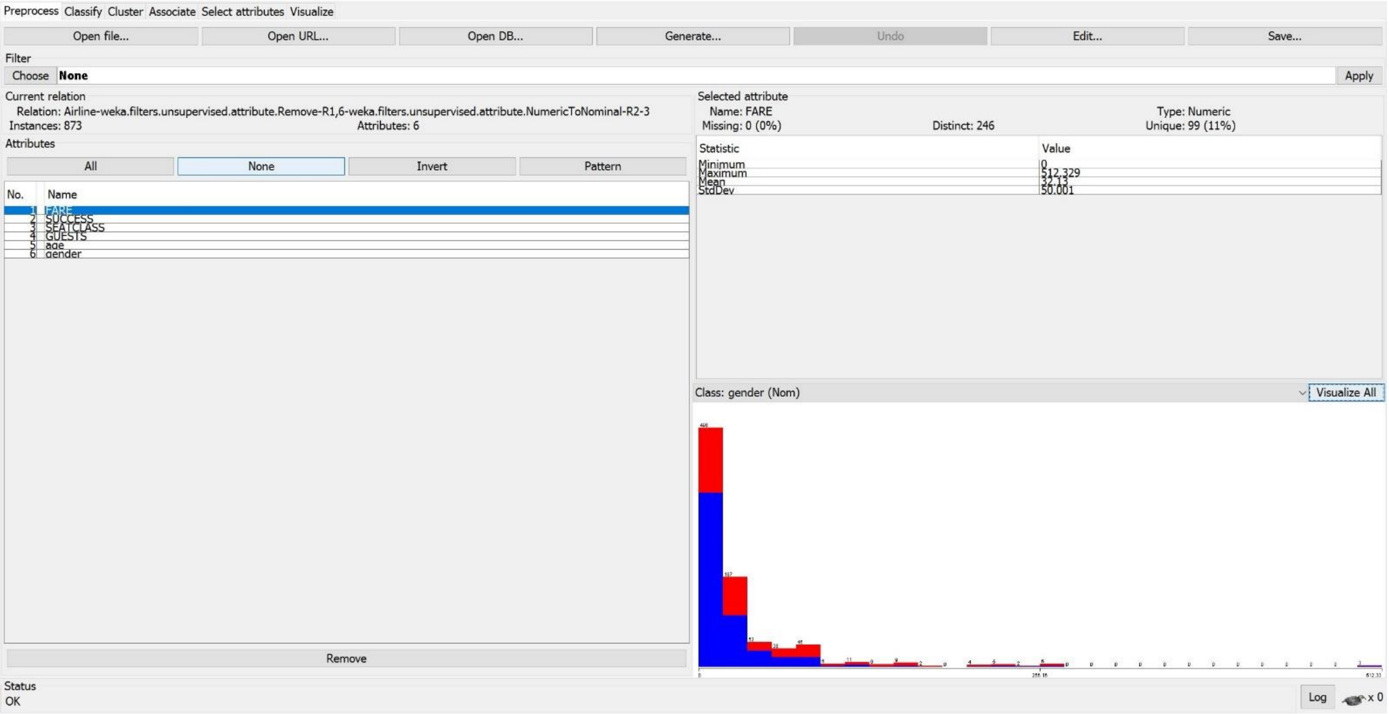
Fare Vs Failure:

- ✓ Fare value 8.05 has highest failure.



7 Milestone4 – Attribute Preparation

In this milestone, csv file was directly imported with all fields. Before saving a csv file from R, unnecessary fields like description and customer id were removed. Under preprocess tab success and seat class have been converted from numerical to nominal. The processed data is stored in arff file format.



Filter
Choose **None**

Current relation
Relation: Airline-weka.filters.unsupervised.attribute.Remove-R1,6-weka.filters.unsupervised.attribute.NumericToNominal-R2-3
Instances: 873
Attributes: 6

Attributes

No.	Name
1	FARE
2	SUCCESS
3	SEATCLASS
4	GUESTS
5	age
6	gender

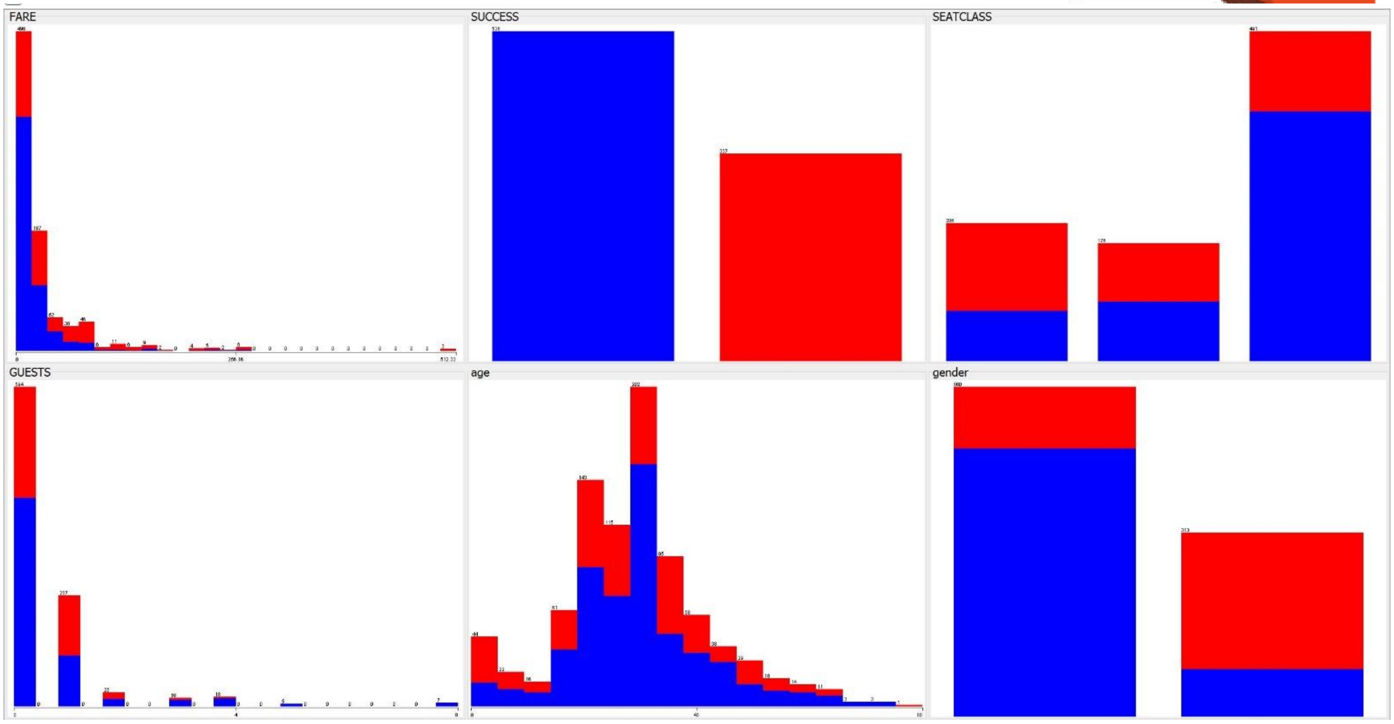
Selected attribute
Name: FARE
Type: Numeric
Missing: 0 (0%)
Distinct: 246
Unique: 99 (11%)

Statistic	Value
Minimum	0
Maximum	512.329
Mean	32.13
StdDev	50.001

Class: gender (Nom)

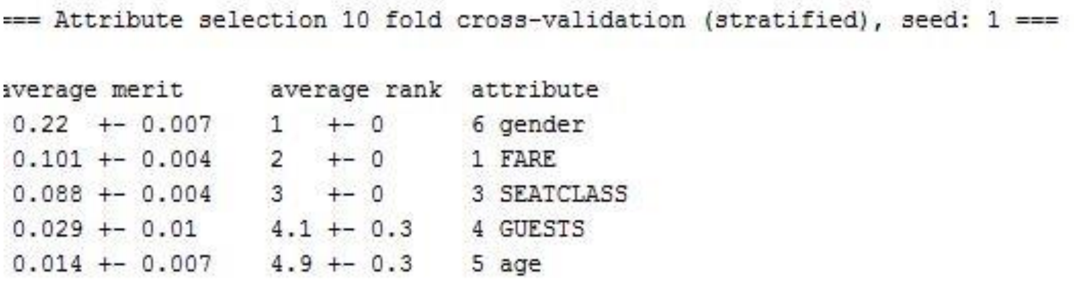
Status
OK

After conversion, in weka all attributes can be visualize.



In the above visualization, every attribute is visualized based on success. First one is fare where the number of success and failure for each fare can be interpreted. Second is success where total number of success and failure can be interpreted. Third one is seat class where the number of success and failure for each seat class can be interpreted. Fourth one is Guests where total number of success and failure for each number of guests from 0 to 8 can be interpreted. Fifth one is age where total number of success and failure for each age can be interpreted. Sixth one is Gender where total number of success and failure for male and female can be interpreted.

Attribute selection is selected based is performed under select attributes tab with attribute evaluator as infogainattreval and ranker method with 10-fold cross-validation. The rank for every attribute can be determined.



```
=== Attribute selection 10 fold cross-validation (stratified), seed: 1 ===

average merit       average rank  attribute
 0.22  +- 0.007     1   +- 0      6 gender
 0.101 +- 0.004     2   +- 0      1 FARE
 0.088 +- 0.004     3   +- 0      3 SEATCLASS
 0.029 +- 0.01      4.1 +- 0.3    4 GUESTS
 0.014 +- 0.007     4.9 +- 0.3    5 age
```

Top two attributes are gender and fare.

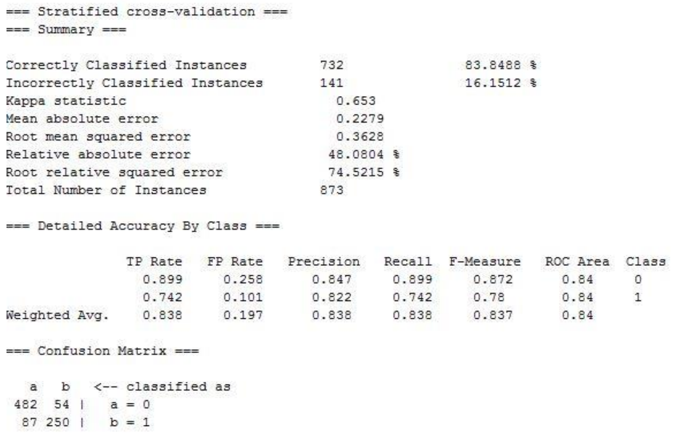
8 Milestone5-Prediction Modelling

1) Simple cart Tree

CART Tree algorithm performs a recursive procedure by which the target variable is divided into various groups based on a decision rule that is designed to optimize the impurity measurement as each one of the successive groups are formed.

Simple cart tree can be performed under classify tab, by choosing classifier as simple cart.

10-fold cross-validation is performed on the data. The confusion matrix and error rates will be displayed.

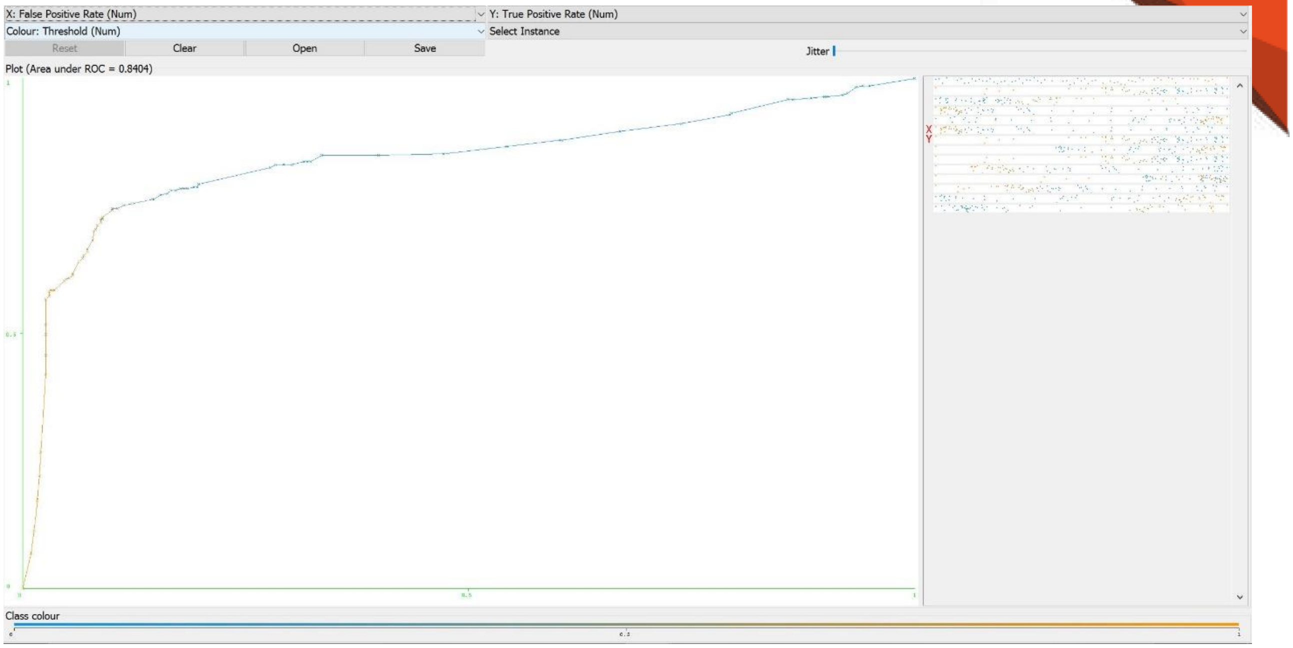


```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         732               83.8488 %
Incorrectly Classified Instances       141               16.1512 %
Kappa statistic                          0.653
Mean absolute error                      0.2279
Root mean squared error                  0.3628
Relative absolute error                 48.0804 %
Root relative squared error             74.5215 %
Total Number of Instances              873

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.899     0.258      0.847     0.899     0.872      0.84     0
                 0.742     0.101      0.822     0.742     0.78       0.84     1
Weighted Avg.    0.838     0.197      0.838     0.838     0.837      0.84

=== Confusion Matrix ===

   a   b   <-- classified as
 482  54 |   a = 0
  87 250 |   b = 1
```

Accuracy is 83.48 % and error rate is 16.15

Roc curve can also be generated.

Area under curve is 0.84, as it greater than 0.8 model it is a good model.



X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

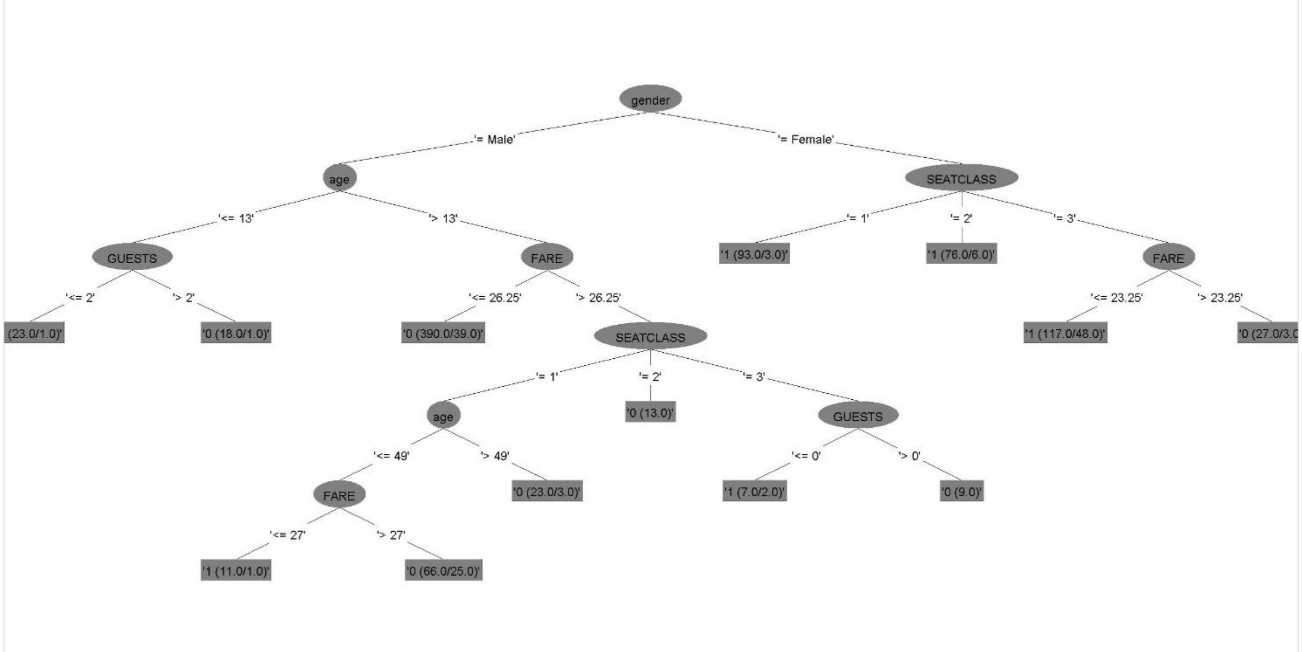
Plot (Area under ROC = 0.8404)

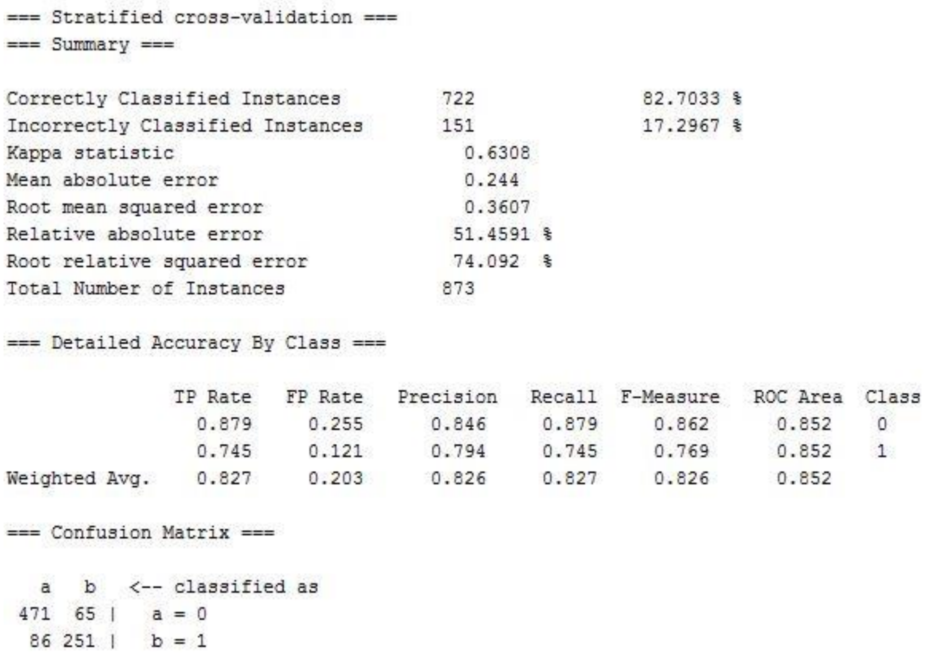
Class colour

2) J48

J48 is an algorithm implement using ID3 algorithm, by which it builds a tree with best predictors.

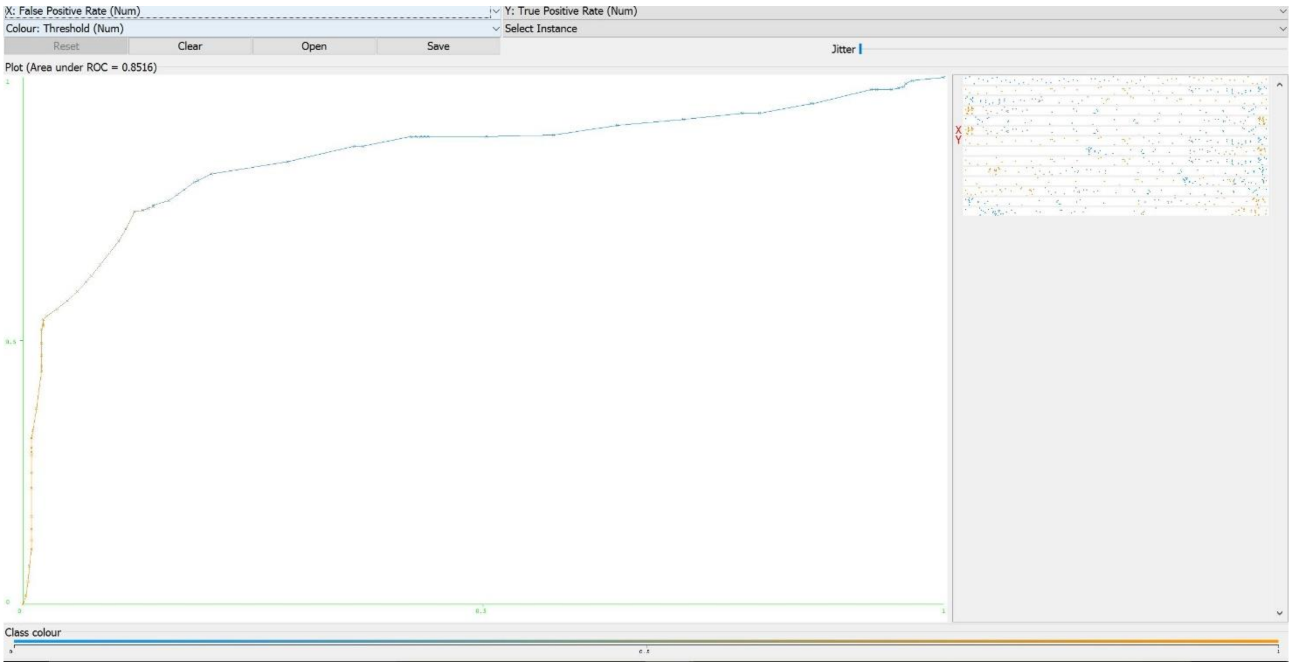
J48can be performed under classify tab, by choosing classifier as J48. 10-fold cross-validation is performed on the data. The confusion matrix and error rates and tree will be displayed.





```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         722               82.7033 %
Incorrectly Classified Instances       151               17.2967 %
Kappa statistic                          0.6308
Mean absolute error                      0.244
Root mean squared error                  0.3607
Relative absolute error                 51.4591 %
Root relative squared error             74.092  %
Total Number of Instances              873

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.879     0.255     0.846     0.879     0.862      0.852    0
                 0.745     0.121     0.794     0.745     0.769      0.852    1
Weighted Avg.    0.827     0.203     0.826     0.827     0.826      0.852

=== Confusion Matrix ===

   a   b   <-- classified as
 471  65 |   a = 0
  86 251 |   b = 1
```

From the above two pictures, gender is the important predictor to classify. Accuracy is 82.7% and error rate is 17.3%

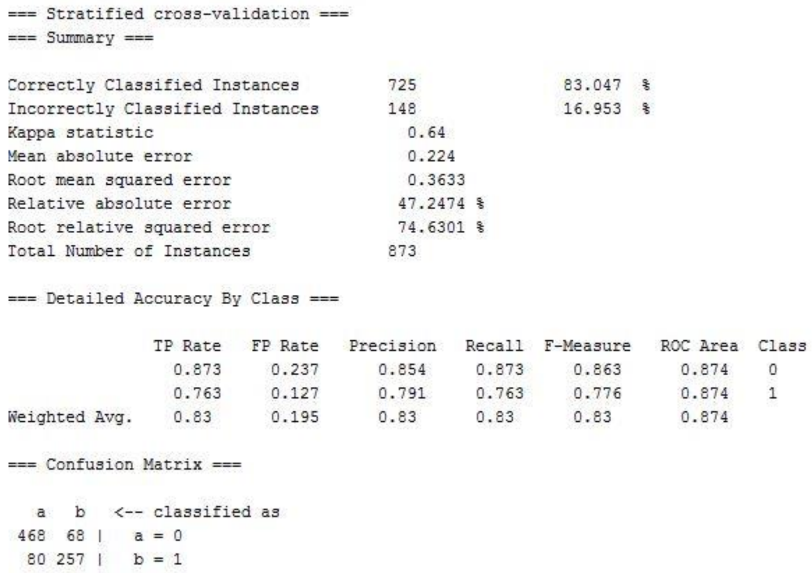


Area under curve is 0.85, as it greater than 0.8 model it is a good model.

3) Random Forest

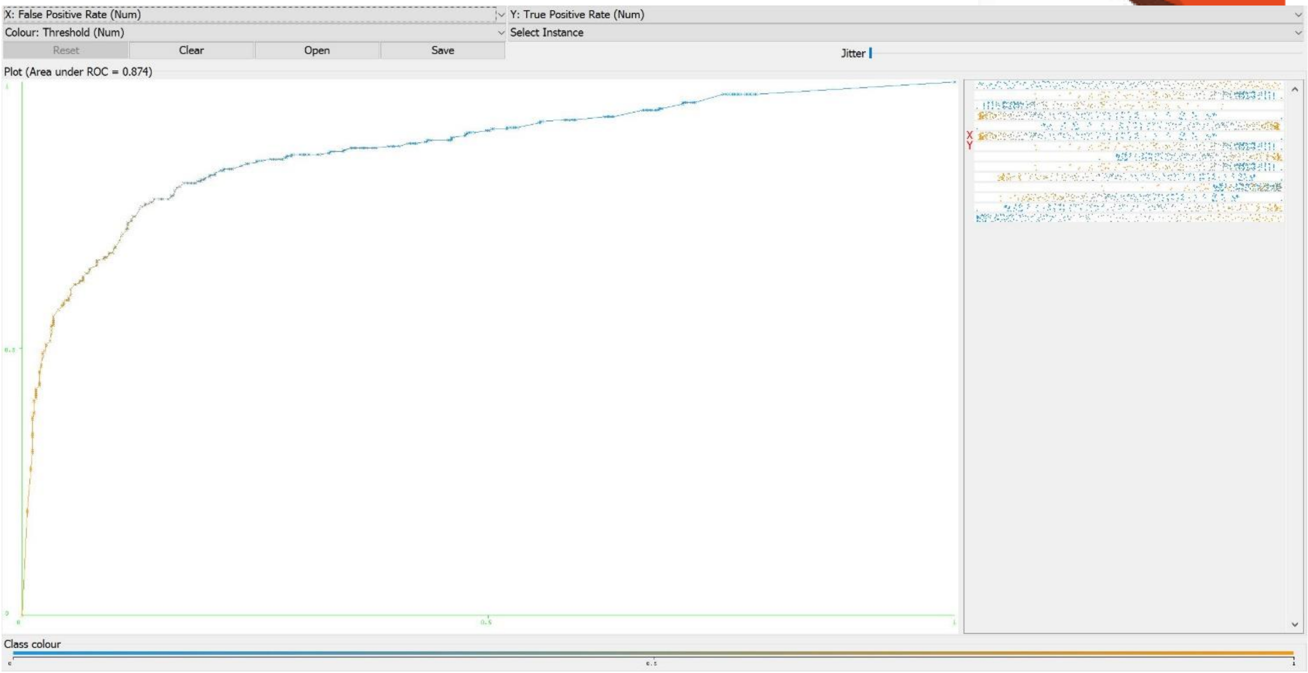
Random Forest Build several decision trees on bootstrapped training sample, but when building these trees, each time a split in a tree is considered, a random sample of m predictors is chosen as split candidates from the full set of $p$predictors (Usually $m=p$ for classification tree and $m=p/3$for regression tree; if $m=p$, random forests amounts to bagging.)

Random forest can be performed under classify tab, by choosing classifier as Random Forest. 10-fold cross-validation is performed on the data. The confusion matrix and error rates will be displayed.



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         725               83.047  %
Incorrectly Classified Instances       148               16.953  %
Kappa statistic                          0.64
Mean absolute error                      0.224
Root mean squared error                  0.3633
Relative absolute error                 47.2474 %
Root relative squared error             74.6301 %
Total Number of Instances              873

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.873     0.237      0.854     0.873     0.863      0.874    0
                 0.763     0.127      0.791     0.763     0.776      0.874    1
Weighted Avg.    0.83      0.195      0.83      0.83      0.83       0.874

=== Confusion Matrix ===

   a   b   <-- classified as
 468  68 |   a = 0
  80 257 |   b = 1
```

Accuracy is 83.1% and error rate is 16.9%



X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Select Instance

Plot (Area under ROC = 0.874)

Area under curve is 0.87, as it greater than 0.8 model it is a good model.

9 Conclusion

- Based on Accuracy, we can say that simple cart tree is the best model.
- According to simple cart tree, Gender is the main classifier to increase success.
- In female, factors to be considered for success are
  - ✓ seat class 3 with age >36.5
  - ✓ seat class 3 with fare>23.35
- In male, factors to be considered for success are
  - ✓ no.of guest >2 with 13<age<49.6
  - ✓ seat class 1 or 3 with age>49.5
  - ✓ age >13 with fare >26.75 with seat class 2