AIRLINE PROJECT

FINAL PROJECT REPORT

AIT 582: APPLICATIONS OF METADATA IN COMPLEX BIG DATA PROBLEMS

JINU KINGCY SEBASTIN

jinusebastin10@gmail.com

G01106302

GEORGE MASON UNIVERSITY

INTRODUCTION

The role of a Data Scientist in an organization is essential in analyzing massive amounts of data that are either structured or unstructured, leading to better decision-making in order to meet the organization's business requirements and goals. Data Scientists must possess strong analytic, statistical and mathematical skills to reveal patterns from unstructured data and make reasonable future predictions.

GOAL OF THE PROJECT

The essence of objective for the Airline project is to analyze the customer database and identify those factors that are crucial in determining the reason for certain customers to fly the airlines while others cancel their reservations. The key task is to come up with a classification model to predict if customers will fly the Airline or not in the future. Recommendations based on the prediction are given to the advertising team of the organization to attract more customers thereby increasing the revenue of the Airline.

TOOLS USED

	R Studio	Tableau	IBM Watson Analytics	Weka
Data Acquisition & Conversion	✓			
Metadata Extraction & Imputation	✓			
Metadata Exploration	✓	✓	✓	
Feature Selection				✓
Prediction Modeling				✓

DATASET DESCRIPTION

The database consists of 891 customer records with 6 major attributes. A detailed description of those attributes is mentioned as follows.

- **Customer ID**: A unique ID assigned to each customer
- **Success**: Determines if the customer successfully flew his/her itinerary
- **Description**: Contains customer's details such as name, age and title
- **Seat class**: First class, Business class, Economy class (1,2,3)

- **Guests**: Number of guests flying with the primary customer
- **Fare**: Total fare paid by the customer

The dataset is programmatically downloaded from the file that is in JSON format http://ist.gmu.edu/~hpurohit/courses/ait582-proj-data-spring16.json. It is converted to a CSV (Comma Separated Values) file for straightforward data manipulation and interpretation. In R studio, the CSV file is converted to a data frame as shown below.

	CUSTOMERID	SUCCESS	DESCRIPTION	SEATCLASS	GUESTS	FARE
1	1	0	Braund, Mr. Owen Harris;22	3	1	7.2500
2	2	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer);38	1	1	71.2833
3	3	1	Heikkinen, Miss. Laina;26	3	0	7.9250
4	4	1	Futrelle, Mrs. Jacques Heath (Lily May Peel);35	1	1	53.1000
5	5	0	Allen, Mr. William Henry;35	3	0	8.0500
6	6	0	Moran, Mr. James;	3	0	8.4583
7	7	0	McCarthy, Mr. Timothy J;54	1	0	51.8625
8	8	0	Palsson, Master. Gosta Leonard;2	3	3	21.0750
9	9	1	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg);27	3	0	11.1333
10	10	1	Nasser, Mrs. Nicholas (Adele Achem);14	2	1	30.0708
11	11	1	Sandstrom, Miss. Marguerite Rut;4	3	1	16.7000
12	12	1	Bonnell, Miss. Elizabeth;58	1	0	26.5500

Figure 1. Data frame showing the output fields of the customer database in R Studio

For the purpose of training and test set, the airline dataset is partitioned in to 80% - 20% (80%-training set, 20%-test set). A comparatively larger portion has been used in the training set to achieve better results from our data when used in data mining algorithms.

PROJECT MILESTONE-1

METADATA EXTRACTION AND IMPUTATION

The data field of Description has different metadata information such as age, title, first name and last name in a single field. Substantial metadata from Description can be extracted and appended as additional fields for each of the data record. Age and Title are useful metadata that are added to our data frame. Gender is derived for each customer depending on their Title. Therefore, we have 9 major attributes in our data frame as shown below.

	CUSTOMERID	SUCCESS	DESCRIPTION	SEATCLASS	GUESTS	FARE	TITLE	GENDER	AGE
1	1	0	Braund, Mr. Owen Harris;22	3	1	7.2500	Mr.	Male	22
2	2	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer);38	1	1	71.2833	Mrs.	Female	38
3	3	1	Heikkinen, Miss. Laina;26	3	0	7.9250	Miss.	Female	26
4	4	1	Futrelle, Mrs. Jacques Heath (Lily May Peel);35	1	1	53.1000	Mrs.	Female	35
5	5	0	Allen, Mr. William Henry;35	3	0	8.0500	Mr.	Male	35
6	6	0	Moran, Mr. James;	3	0	8.4583	Mr.	Male	0
7	7	0	McCarthy, Mr. Timothy J;54	1	0	51.8625	Mr.	Male	54
8	8	0	Palsson, Master. Gosta Leonard;2	3	3	21.0750	Master.	Male	2
9	9	1	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg);27	3	0	11.1333	Mrs.	Female	27
10	10	1	Nasser, Mrs. Nicholas (Adele Achem);14	2	1	30.0708	Mrs.	Female	14
11	11	1	Sandstrom, Miss. Marguerite Rut;4	3	1	16.7000	Miss.	Female	4
12	12	1	Bonnell, Miss. Elizabeth;58	1	0	26.5500	Miss.	Female	58

Figure 2. Data frame after metadata extraction and appending three new fields

The dataset needs to undergo Preprocessing such as the cleaning of data because it has missing values and outliers, after which it can be made useful for data mining processes. The data fields Age has missing values which are imputed with its median value grouping by the Title and Fare has missing value which are imputed with their mean value grouping by the Seat class. A histogram representation of customer's Age with outliers is depicted below.

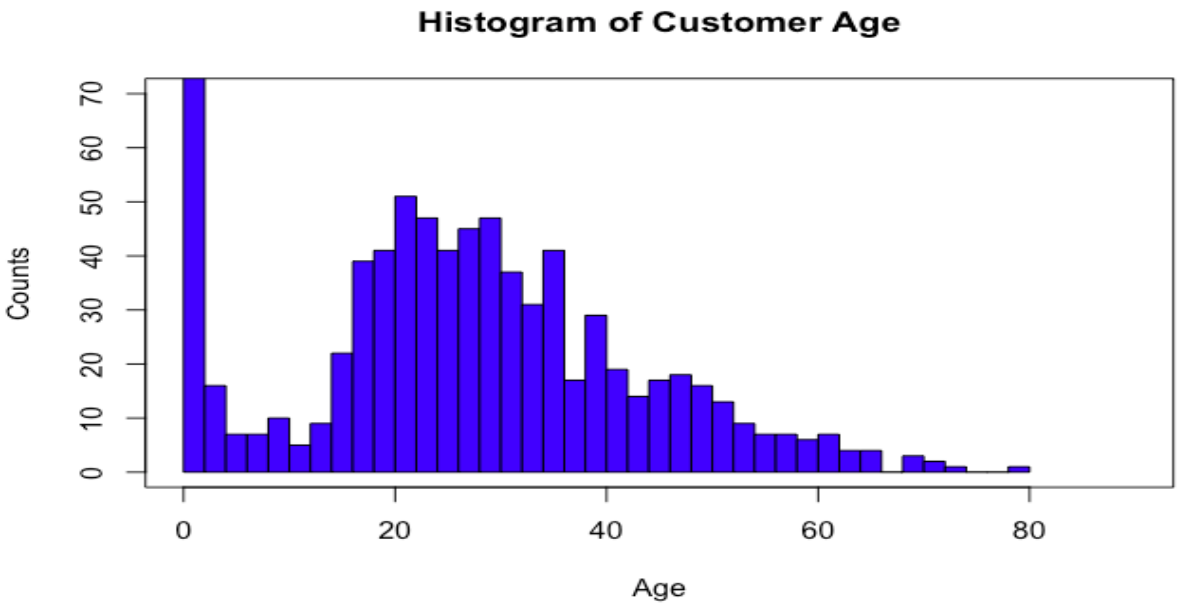

Figure 3. Histogram depiction of Customer's Age with outliers and before imputation

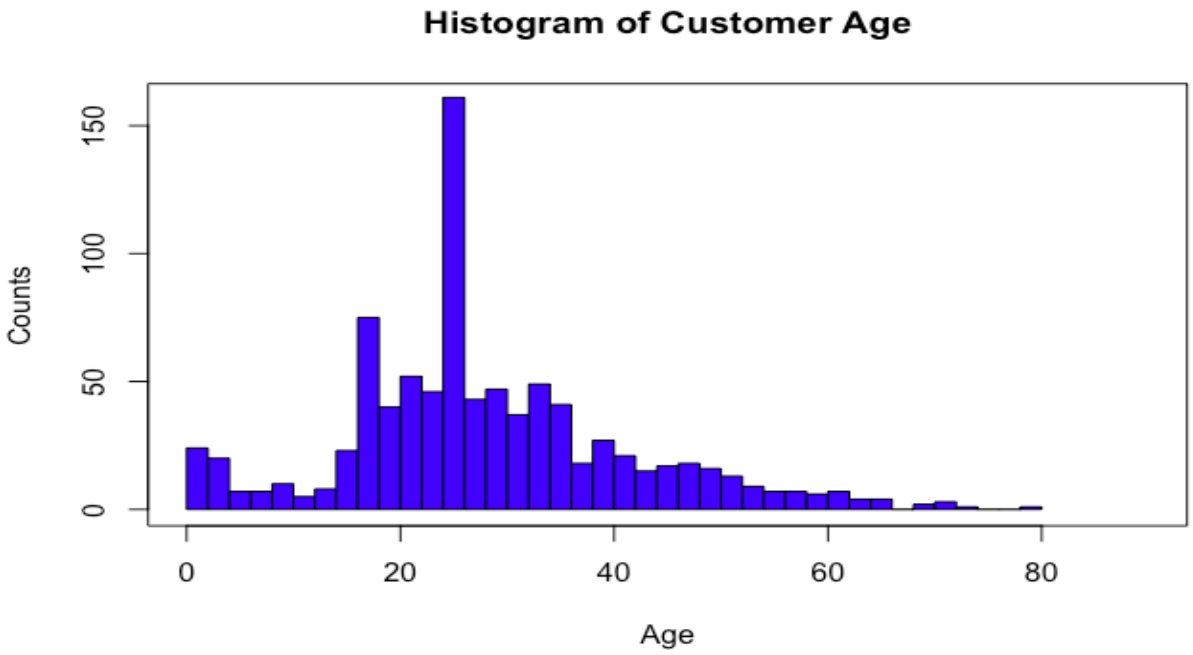

Figure 4. Histogram depiction of Customer's Age after imputation and removing outliers

In order to deal with the outliers, the dataset has to be normalized. There are two main normalization techniques namely, Min-Max normalization and Z-Score Standardization. Using, min-max normalization the skewness from the data can be reduced, thus preparing our data for data mining process and feature selection.

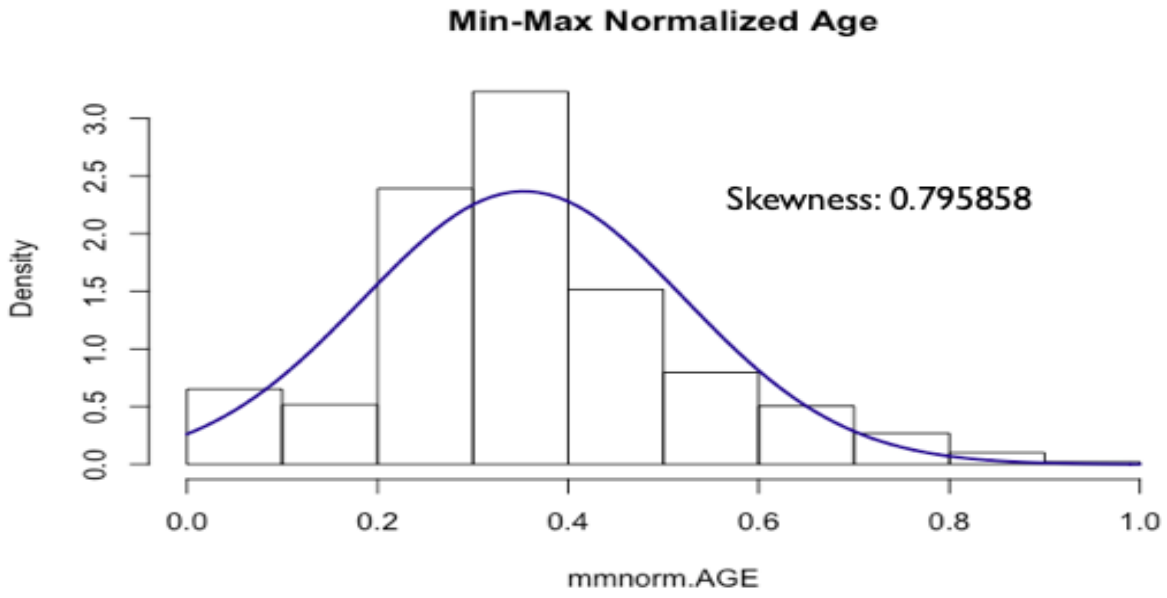

Figure 5. Customer's Age after normalizing the data and reducing the skewness

PROJECT MILESTONE-2

SUMMARIZATION AND VISUALIZATION

A Statistical summary for the customer's Age and Fare is depicted with the mean, median and standard deviation values.

```
> print(data_q.statsum)
   var    min    q25 median    q75      max     mean       sd
1  AGE 0.4200 21.000   25.0 35.000  80.0000 28.55799 13.41191
2 FARE 4.0125  7.925   14.5 31.275 512.3292 32.87692 49.69010
```

Figure 6. Statistical summary of Age and Fare

Various visualizations are obtained using R studio, Tableau and IBM Watson Analytics tools to relate and analyze different variables.

From the bar graph shown below, it is evident that majority of the passengers fly in the Economy class. A possible reason is that third class is affordable by most of the people.

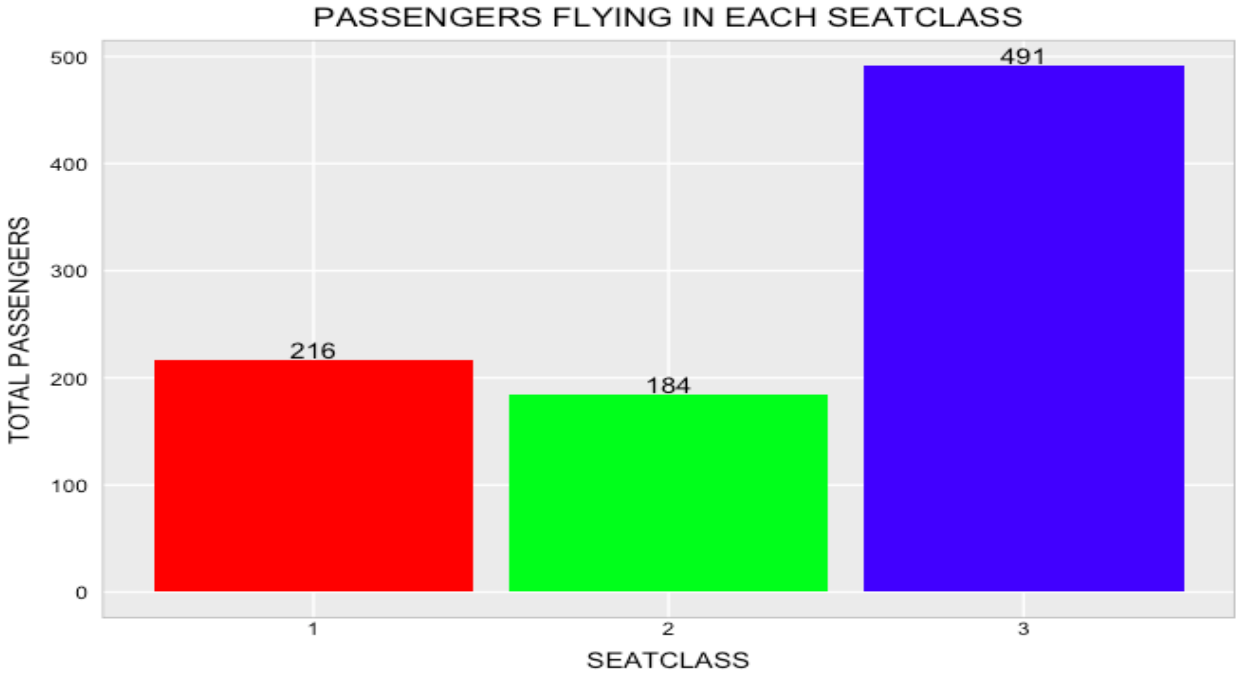

Figure 7. Bar Plot representation of passengers flying in all three classes

From the chart below, it is clearly seen that among the passengers traveling in all three classes, male passengers surpass the female passengers.

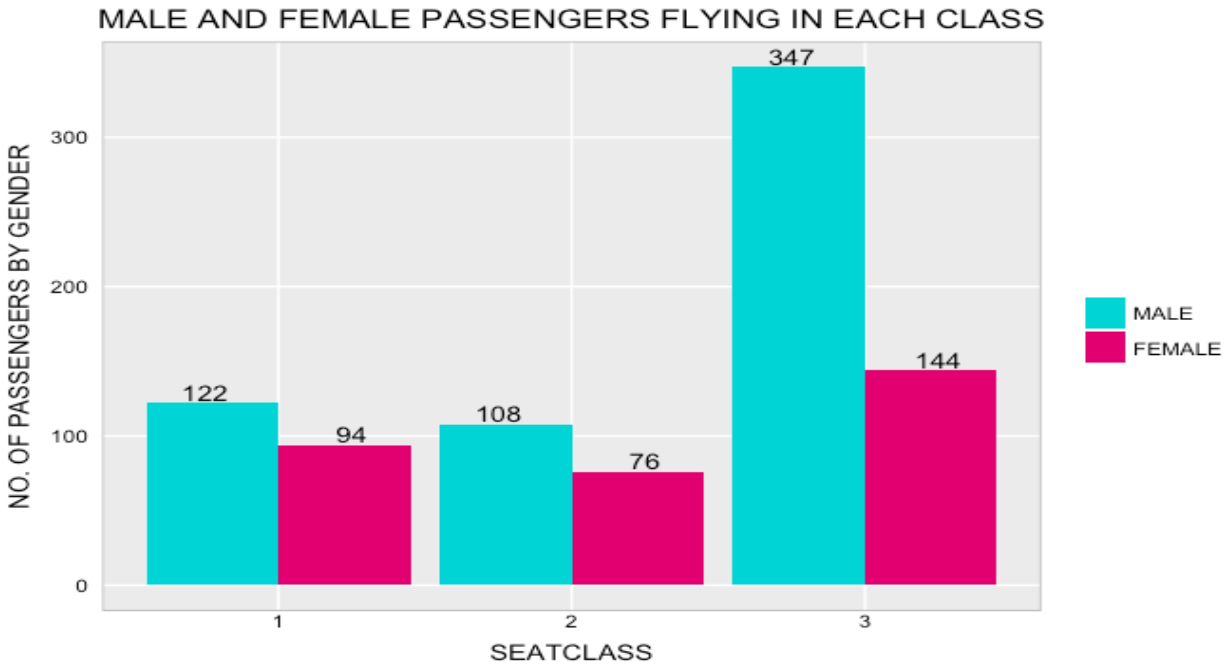

Figure 8. Grouped bar plot representation of passengers flying in each class by gender

The below visualization depicts the number of customers flying with guests. Majority of the customers prefer to fly alone followed by those flying with one guest.

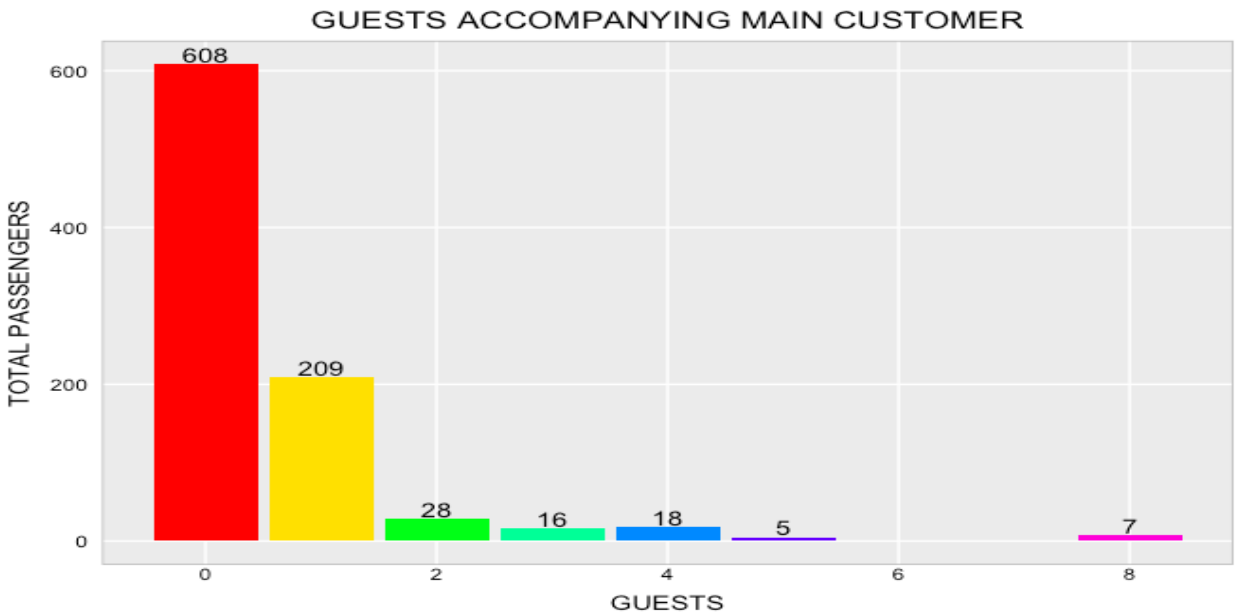

Figure 9. Bar plot depiction of guests accompanying the main customer

The plot below portrays the fact that customers traveling alone pay more on an average and customers traveling with more guests pay a discounted price.

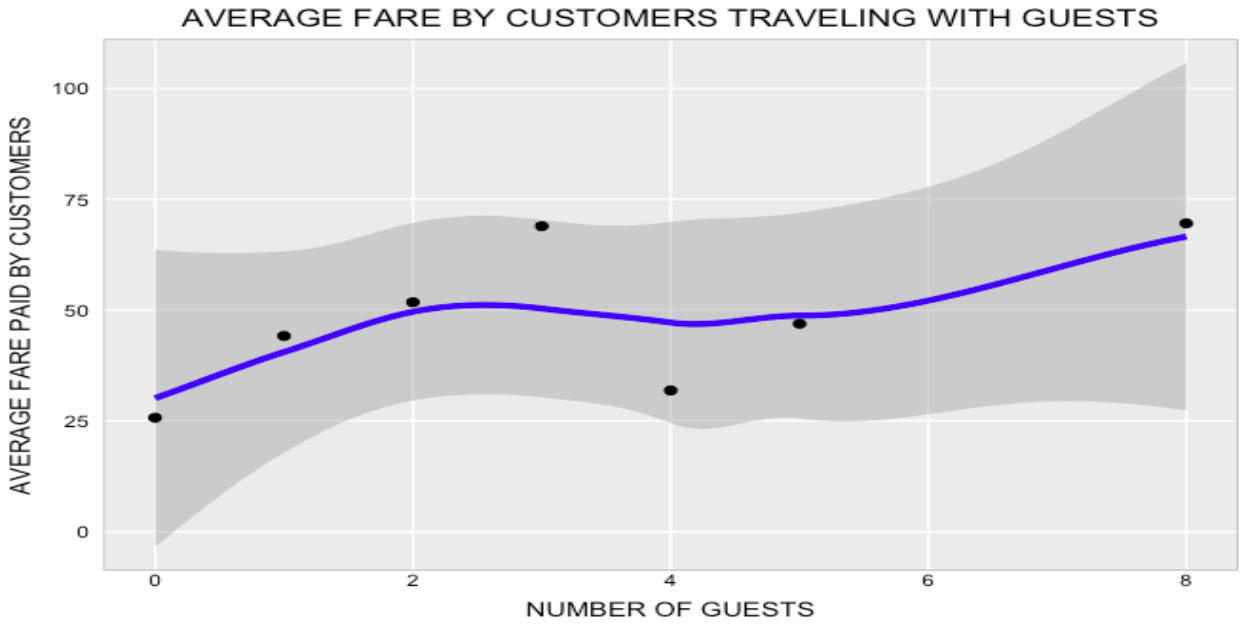

Figure 10. Scatterplot representing the average fare paid by customers traveling with guests

The bar chart shows the average fare paid by customers traveling in each class. Economy class is cheaper which justifies the fact that majority of the customers fly in 3rd class.

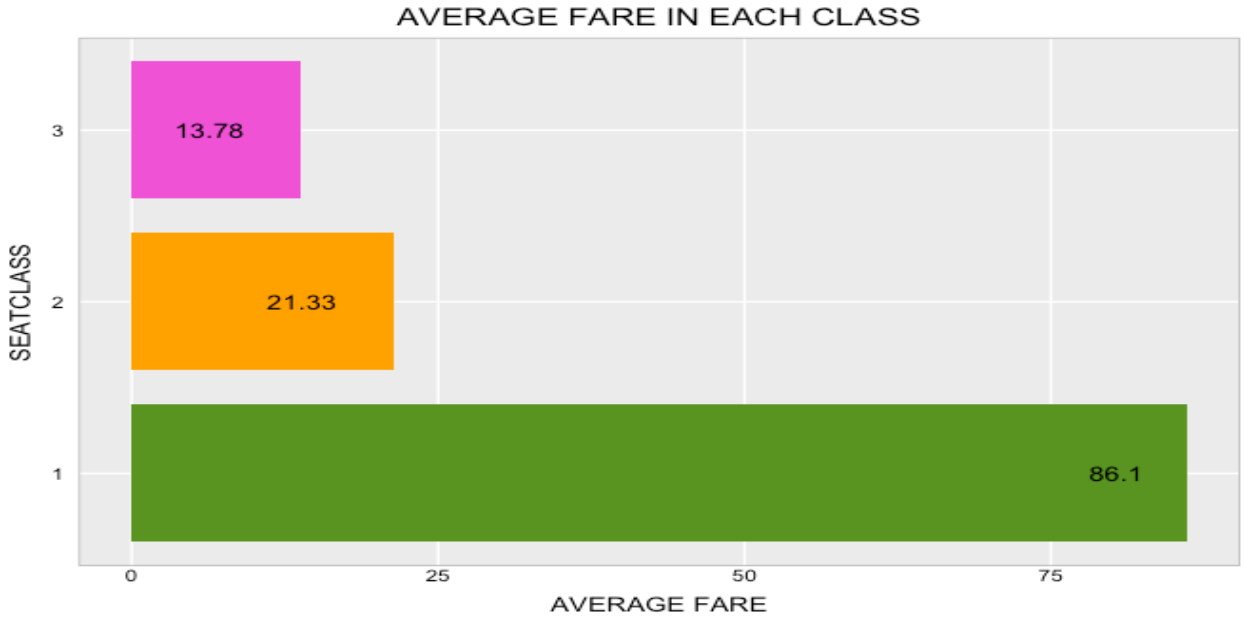

Figure 11. Bar plot showing average fare in each class

The pie charts interpret the number of guests traveling in each class by gender. Most of them fly in the 3rd class, preferably the reason being cost as its cheaper. Female guests fly more in First class compared to men.

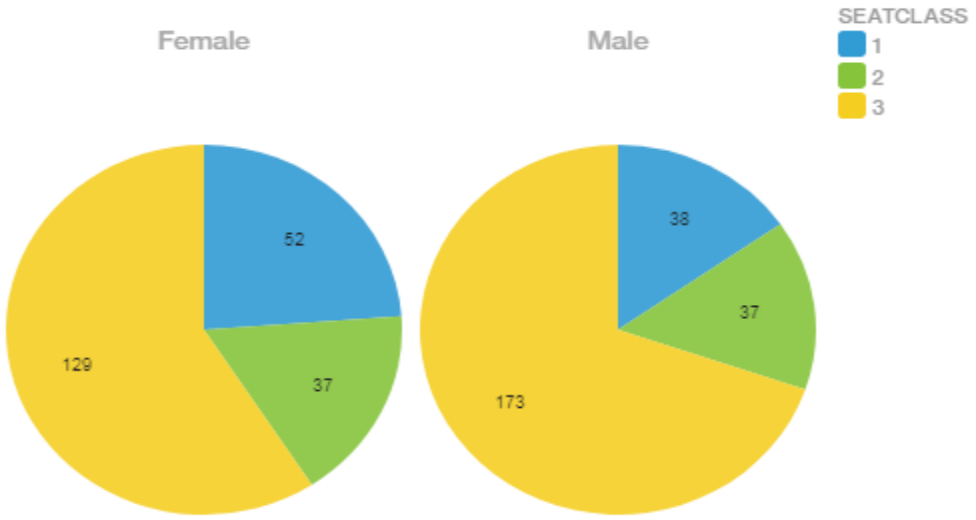

Figure 12. Pie chart representation of Guests traveling in each class

The density plot graph characterizes that customers pay $32.90 on an overall average for flying the Airline.

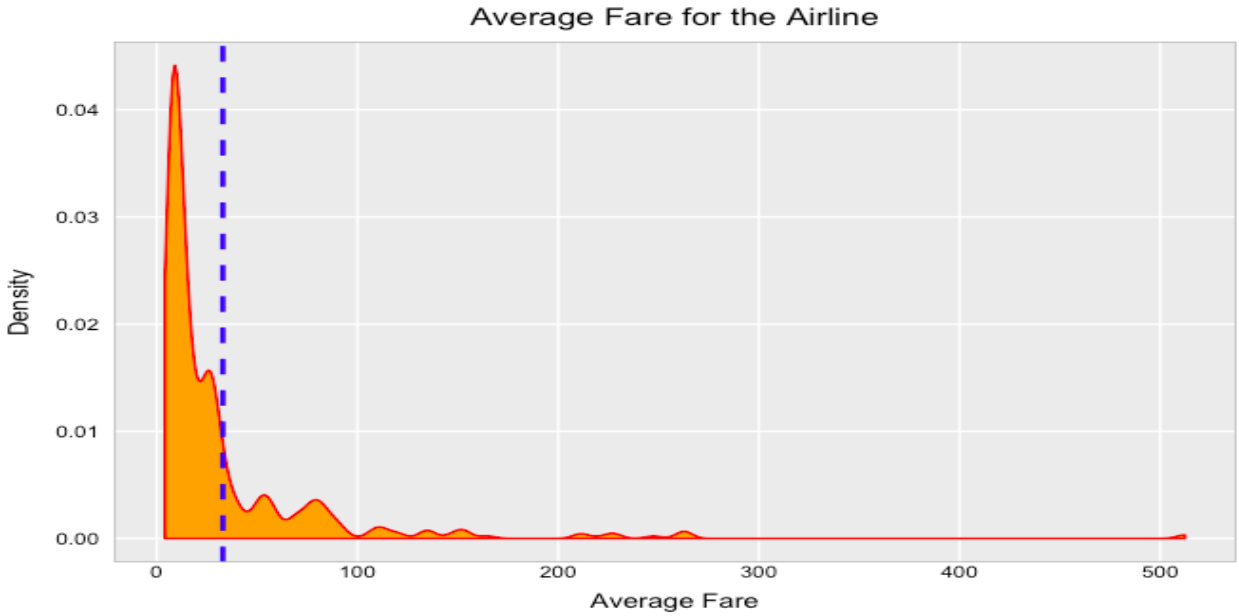

Figure 13. Density plot representing the average cost a customer pays for the Airline

This chart illustrates the total Fare received by the airline grouping customers by their age. It is clear that the Airline makes most of its revenue from its customers in the age group of 25.

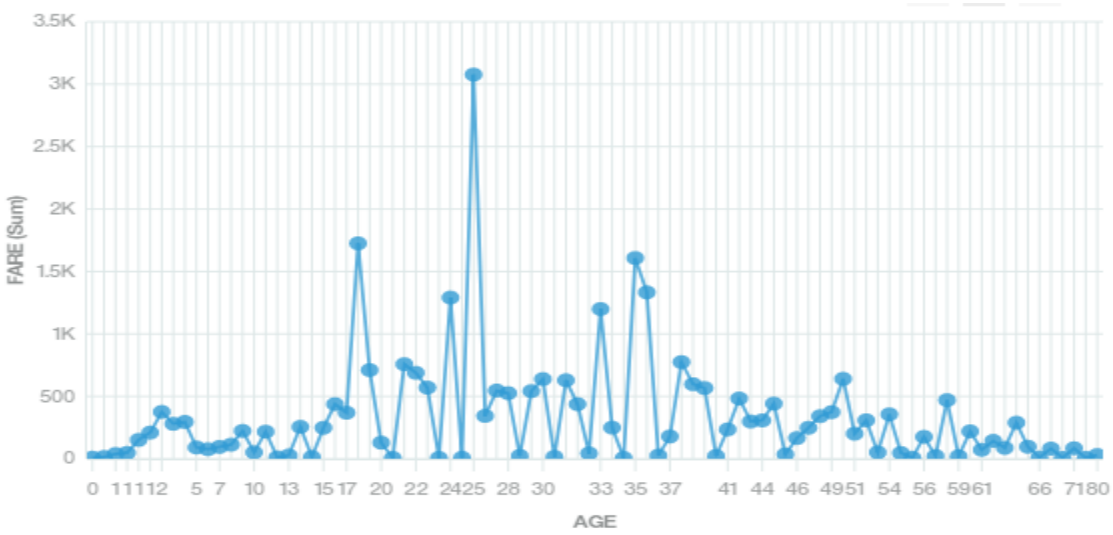

Figure 14. Dot plot depicting the highest fare earned by the airline grouping customers by age

The visualizations below explain the fact that customers traveling in the age group of 18 has the most success rate of flying the Airline and those traveling in the age group of 25 has the maximum cancellation rate.

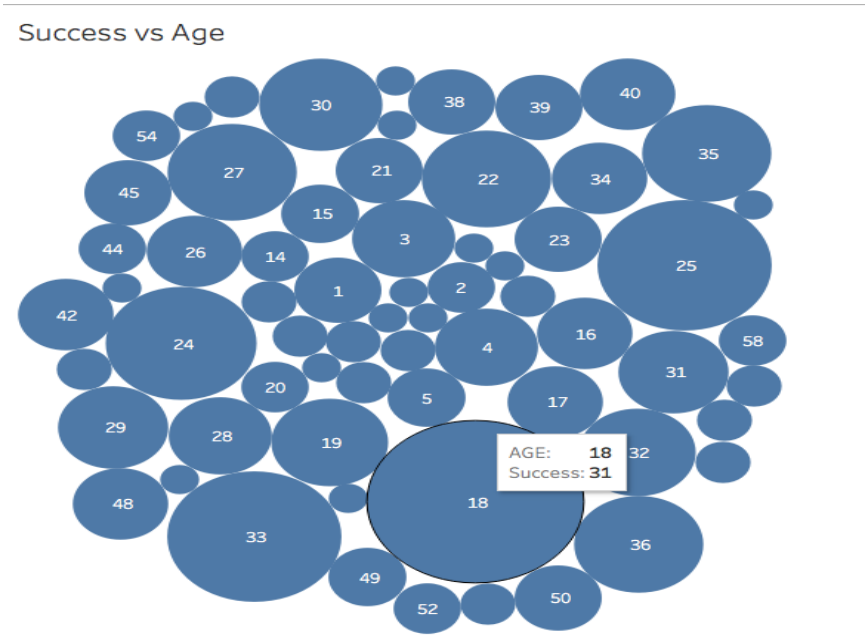

Figure 15. Packed bubbles chart depicting highest success by age

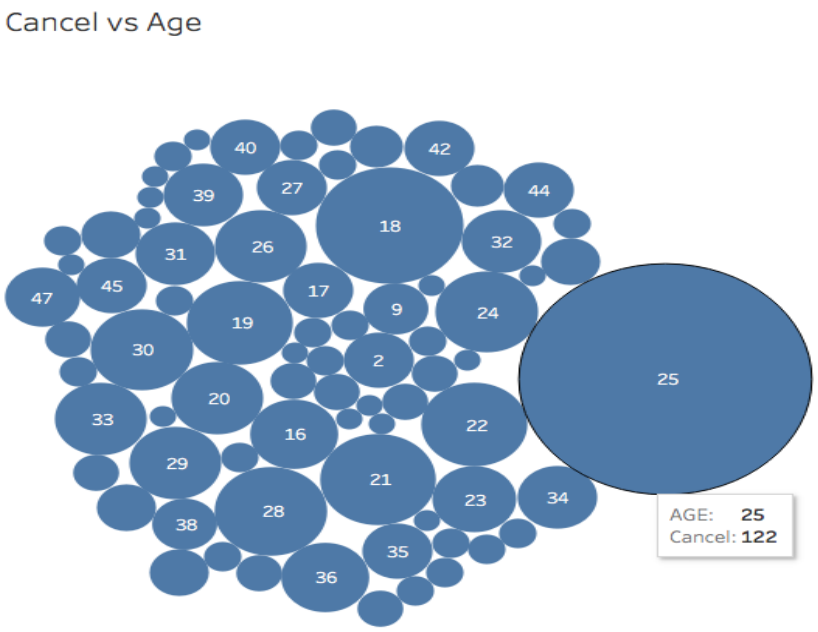

Figure 16. Packed bubbles chart depicting highest cancellation by age

Based on the graph below, conclusions are made that highest cancellations are made in the 3rd class.

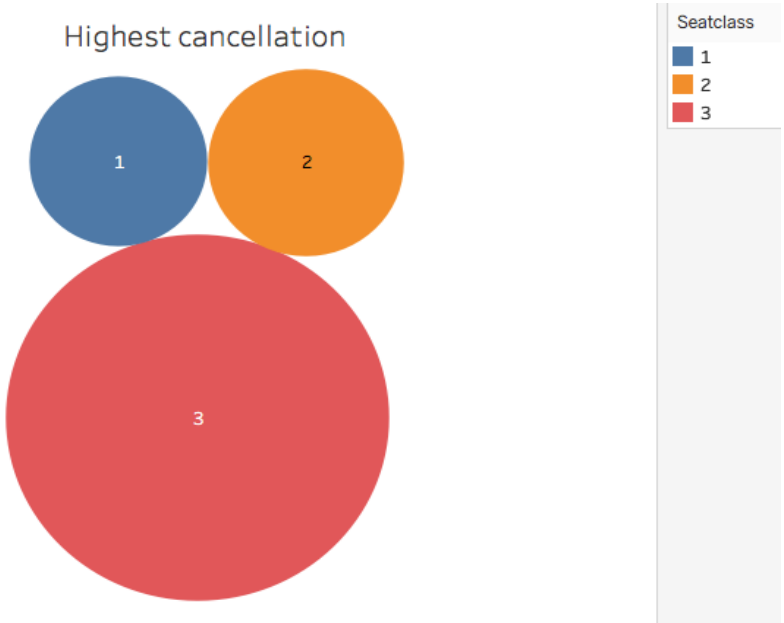

Figure 17. Packed bubble showing cancellation rate in each class

The chart explains the success rate of passengers flying the airline by gender. It is vivid that greater number of men cancel the flight and the success rate of women flying the airline is more than men.

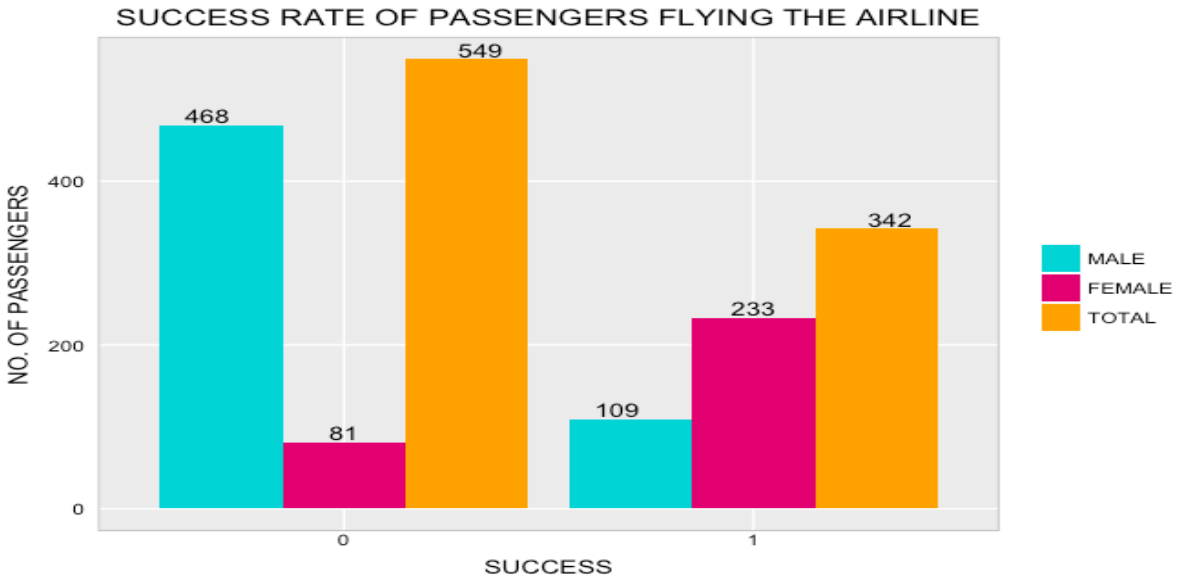

Figure 18. Grouped bar chart representing success rate of passengers by gender

The below visualization proves that a predominant part of customers flying the airline are men.

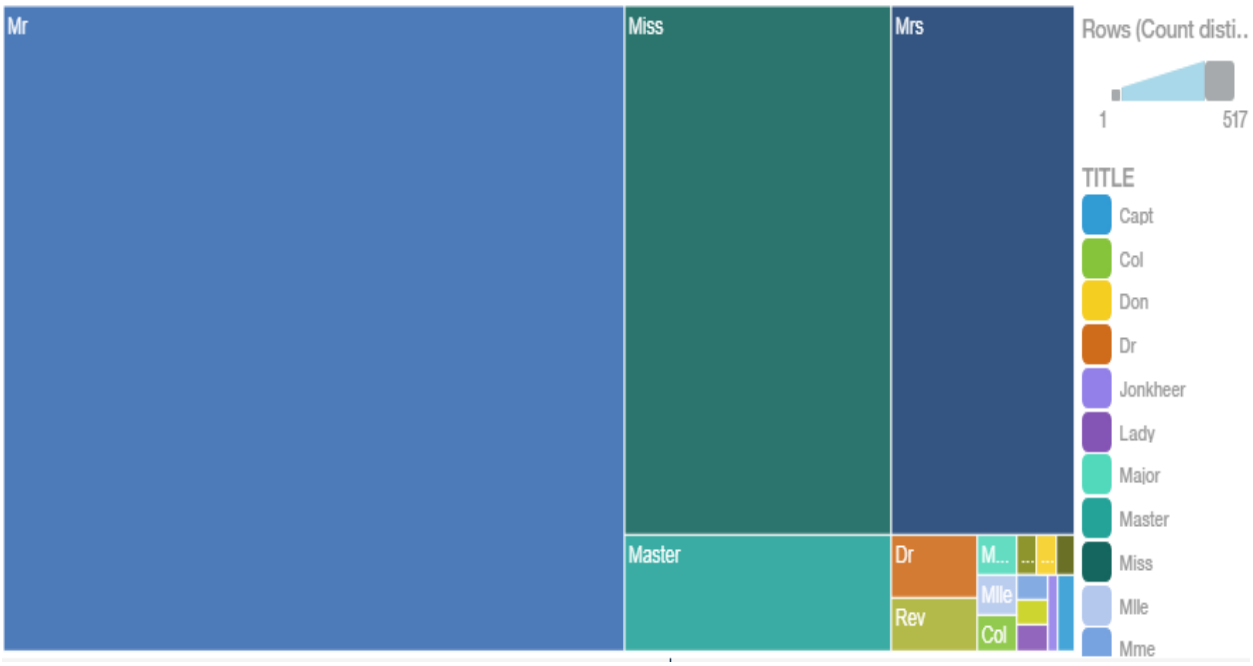

Figure 19. Tree map showing the distribution of gender in the airline by Title

PROJECT MILESTONE-3

ANALYTICS

The CSV file is then converted to ARFF (Attribute Relation File Format) file containing only selected variables for attributes. The target variable is Success and the predictor variables are Seat class, Gender, Fare, Guests and Age. Success and Seat class are set to nominal while the remaining attributes are numeric. The ARFF file is now in the right format to prepare for mining algorithms. Using Weka tool, the attribute data file is loaded, and a comprehensive visualization is obtained.

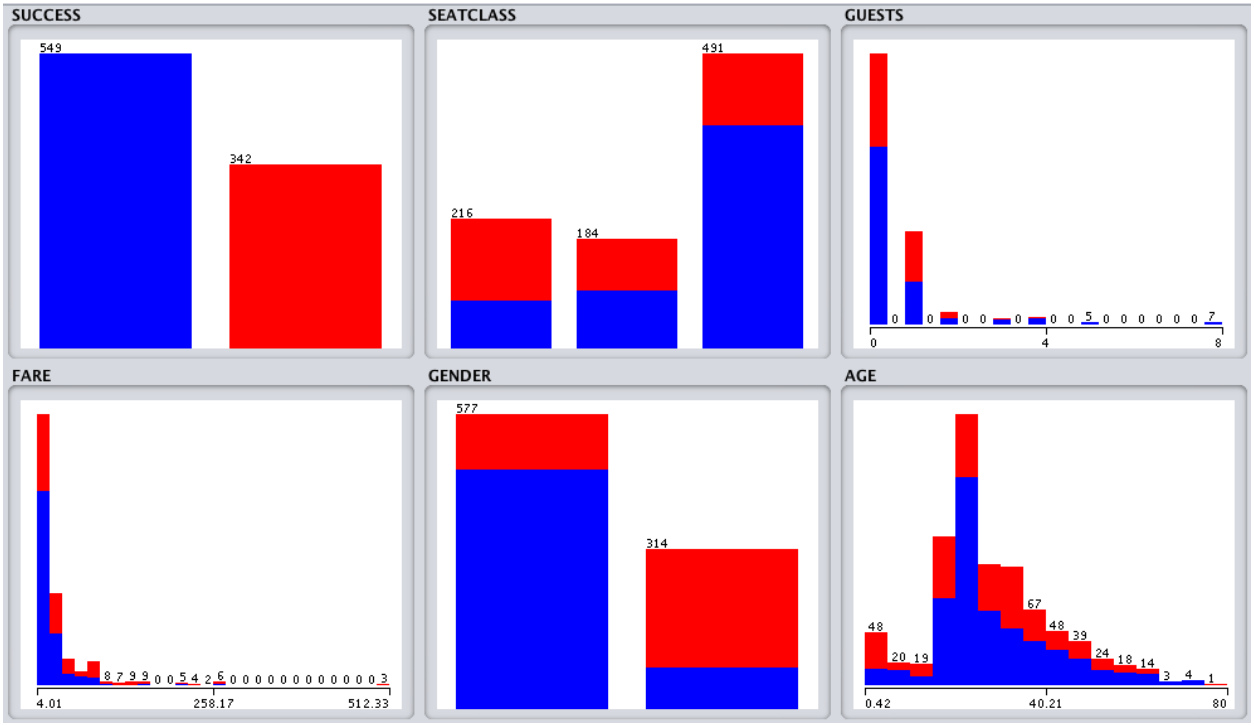

Figure 20. Comprehensive visualization of the selected attributes obtained from Weka

FEATURE SELECTION

Feature selection is the process of choosing subset of relevant variables and predictors for our model. There are many techniques for feature selection available in Weka but the technique we have used is “Information Gain Based Feature Selection”. It chooses the best attributes by measuring the information gain. Attributes that contribute more will have higher information gain value and vice versa. Using the feature selection technique and 10-fold cross validation the top 2 attributes are identified. They are Gender and Seat class, closely followed by Fare.

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

average merit      average rank  attribute
 0.218 +- 0.006    1    +- 0     5 GENDER
 0.084 +- 0.006    2.5  +- 0.5   2 SEATCLASS
 0.085 +- 0.006    2.5  +- 0.5   4 FARE
 0.029 +- 0.003    4.4  +- 0.49  3 GUESTS
 0.03  +- 0.016    4.6  +- 0.49  6 AGE
```

Figure 21. Attribute selection using Information Gain Based method and 10-fold cross validation

PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis is a powerful technique more useful for finding new variables, also known as principal components, from the observed variables that are informative and have uncorrelated features. PCA allows dimension reduction thus rejecting the features with low variance. The correlation between Seat class and Fare can be obtained using PCA technique in Weka. It is evident that most of the passengers travel in the third class (marked in red) where the fare is relatively low than the first and second class.

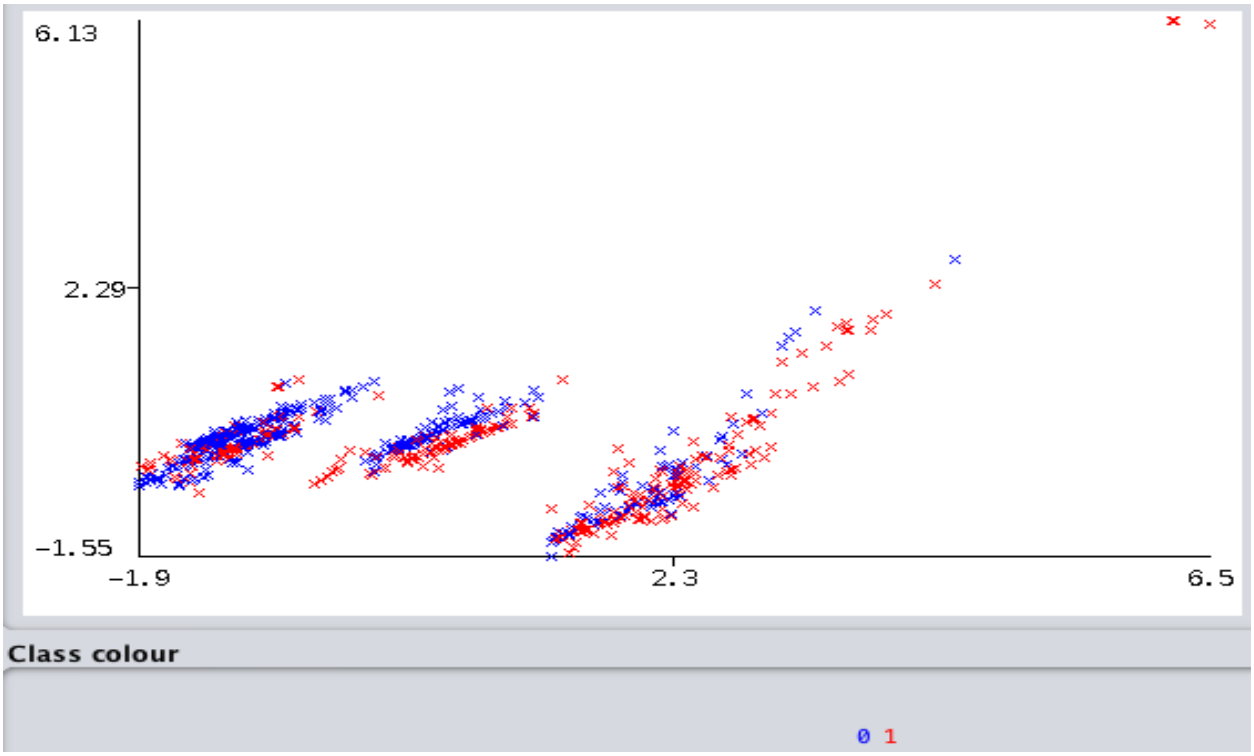

Figure 22. Correlation between Seat class and Fare using PCA technique in Weka

The correlation between Seat class and Age can be obtained in a similar manner in Weka. It is inferred that passengers travelling in the third class are mostly young adults around the age of 18-25 with fairly middle-aged and elderly customers. Whereas the first class has most of its customers that are middle-aged with very few young adults. Similarly, new information can be interpreted by the correlation between selected attributes using PCA.

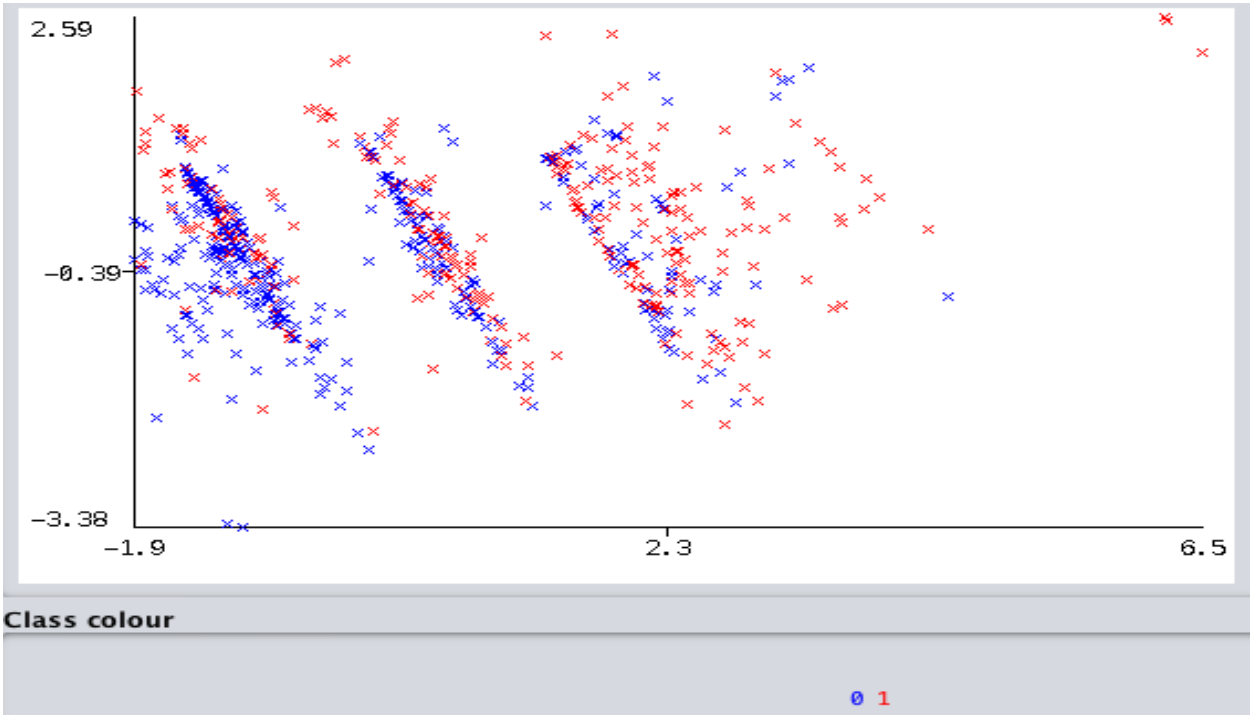

Figure 23. Correlation between Seat class and Age using PCA technique in Weka

PROJECT MILESTONE-4

PREDICTION MODELING

There are various prediction modeling methods we could choose to apply for the model. In our scenario, we use Decision Tree and Random Forest algorithms.

J48

Decision tree J48 was developed by Weka and it is the implementation of the ID3 algorithm. By applying decision tree J48 on the selected attributes we could predict the target variable for a new attribute. The classifier output denotes the Accuracy as 81.59% with correctly classified instances as 727 out of the 891 instances. The Area under ROC is 0.8421 which signifies that the model is good.

Classifier output

```
|   |   FARE > 25.125: 0 (270.0/51.0)

Number of Leaves  :     13

Size of the tree :      23


Time taken to build model: 0.02 seconds

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

Correctly Classified Instances         727               81.5937 %
Incorrectly Classified Instances       164               18.4063 %
Kappa statistic                          0.6057
Mean absolute error                      0.2544
Root mean squared error                  0.3705
Relative absolute error                 53.7733 %
Root relative squared error             76.1874 %
Total Number of Instances              891

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.872     0.275     0.836      0.872    0.854      0.607    0.842     0.846     0
                 0.725     0.128     0.780      0.725    0.752      0.607    0.842     0.801     1
Weighted Avg.    0.816     0.218     0.814      0.816    0.815      0.607    0.842     0.829

=== Confusion Matrix ===

   a   b   <-- classified as
 479  70 |   a = 0
  94 248 |   b = 1
```

Figure 24. Classifier output of J48 showing Accuracy of 81.59%

Tree View

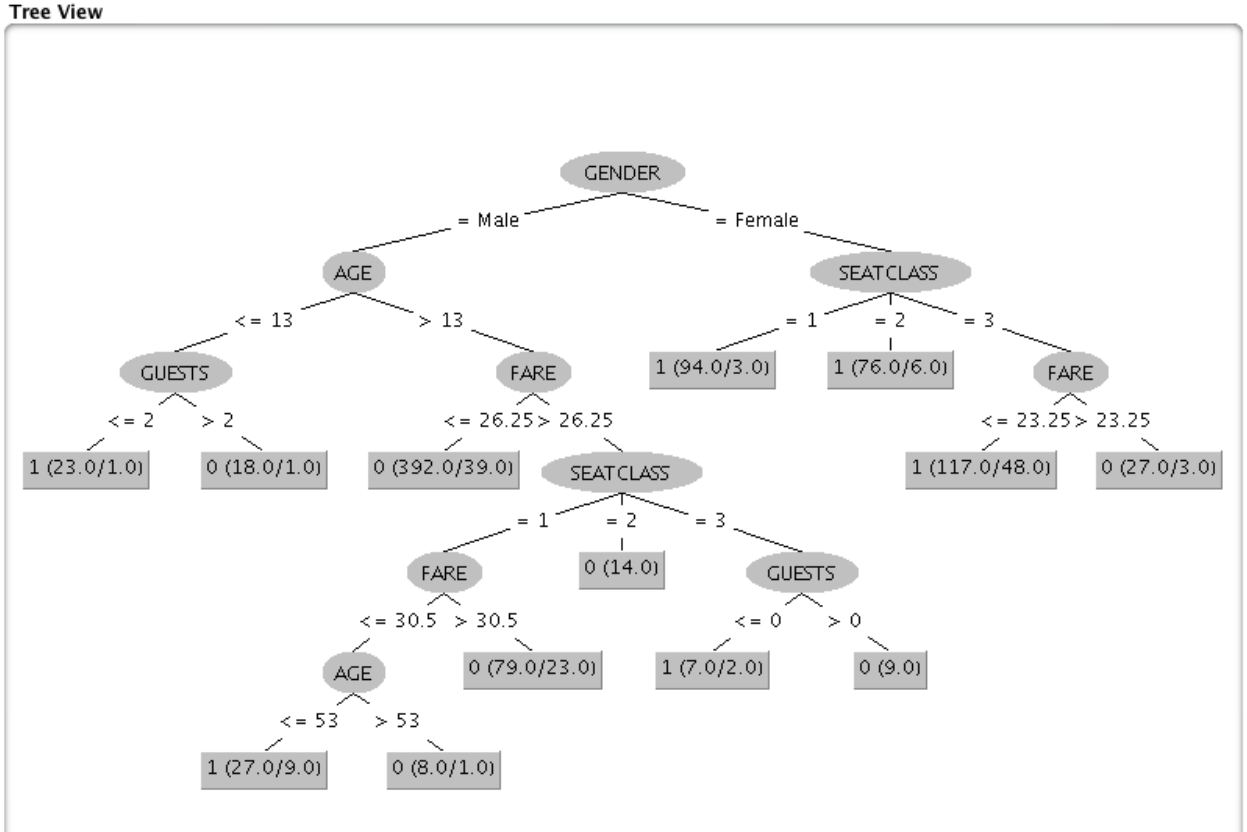

Figure 25. J48 Tree view

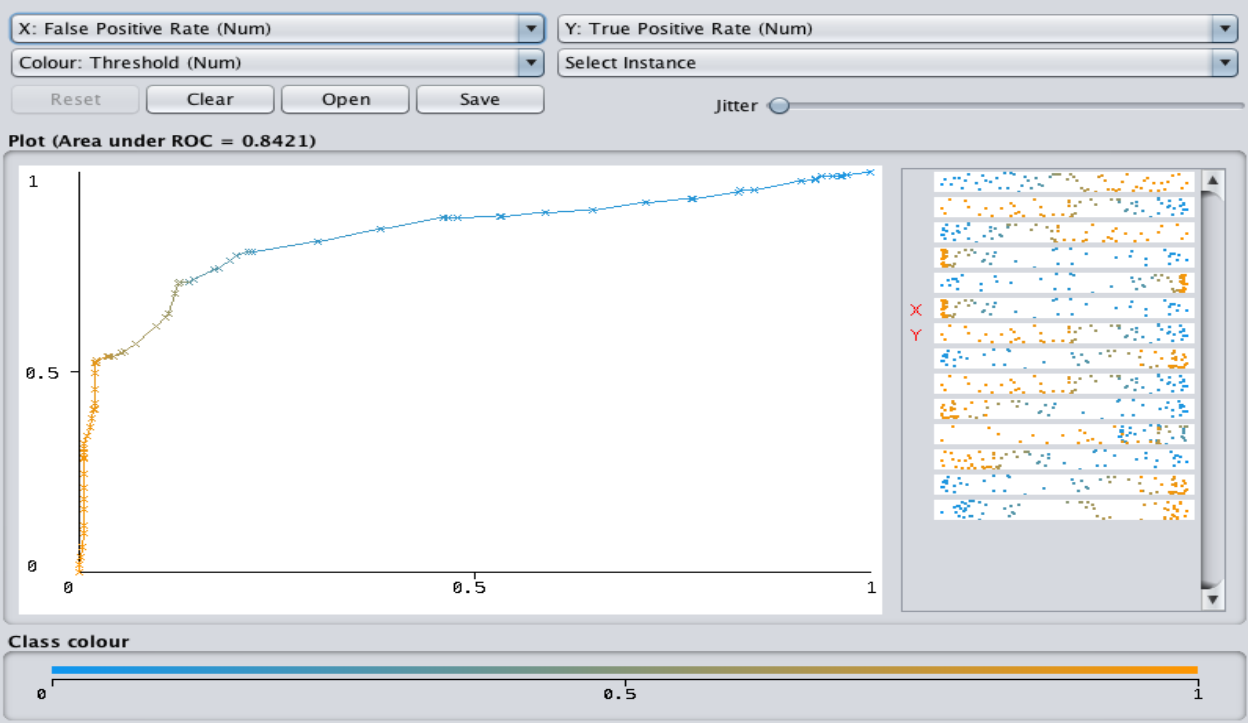

Figure 26. The Area under ROC curve is 0.8421, predicting a good model

RANDOM FOREST

Random Forest is fast and an efficient algorithm for modeling as it comprises of multiple decision trees. The random forest output denotes the Accuracy as 80.47% with correctly classified instances as 717 out of the 891 instances. The Area under ROC is 0.8558 which signifies that the model is better.

```
RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 0.17 seconds

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

Correctly Classified Instances         717               80.4714 %
Incorrectly Classified Instances       174               19.5286 %
Kappa statistic                          0.5825
Mean absolute error                      0.2393
Root mean squared error                  0.3811
Relative absolute error                 50.5837 %
Root relative squared error             78.3724 %
Total Number of Instances              891

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0.860     0.284     0.830       0.860   0.844      0.583    0.856     0.876     0
                 0.716     0.140     0.761       0.716   0.738      0.583    0.856     0.806     1
Weighted Avg.    0.805     0.229     0.803       0.805   0.804      0.583    0.856     0.849

=== Confusion Matrix ===

   a   b   <-- classified as
 472  77 |   a = 0
  97 245 |   b = 1
```

Figure 27. Classifier output of Random Forest algorithm showing Accuracy of 80.47%

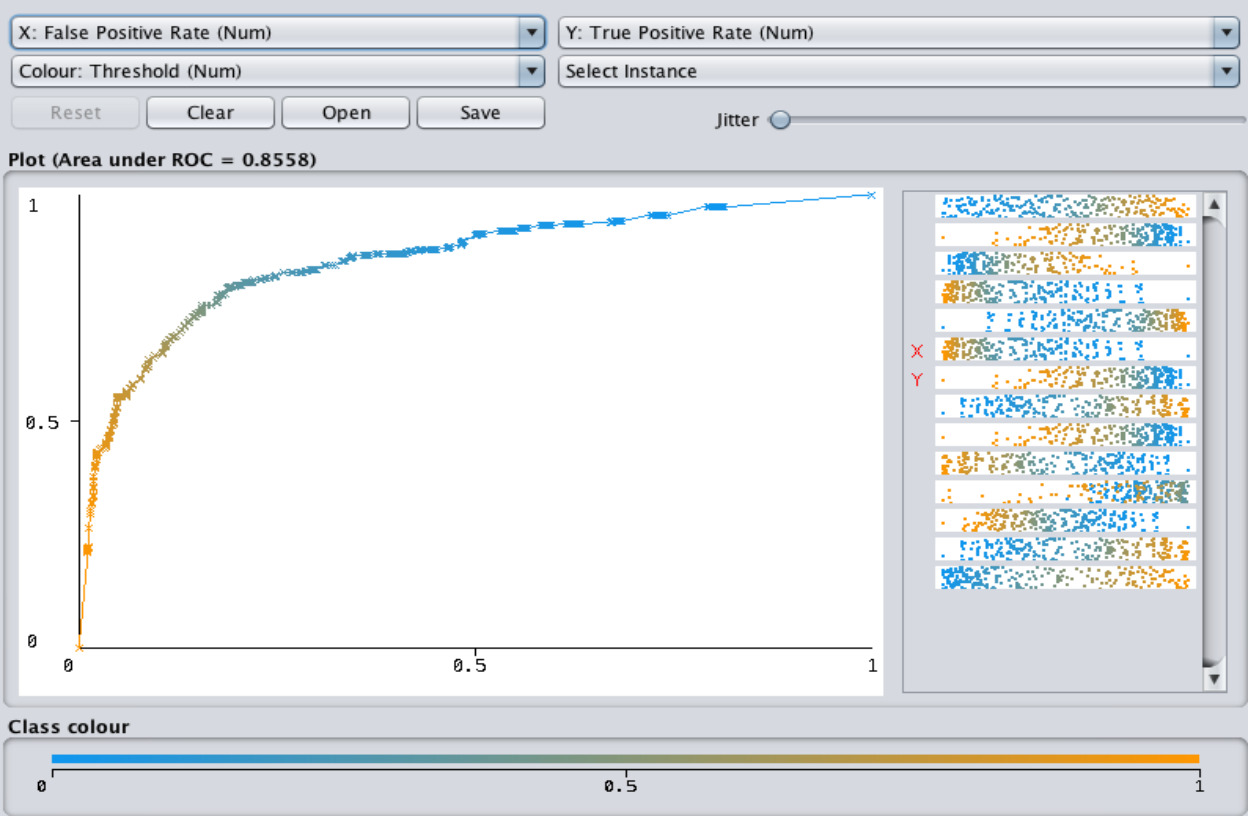

Figure 28. The Area under ROC curve is 0.8558, predicting a better model

We perceive that to determine the performance of a classifier within its total working range we have to analyze the ROC value. In general, a value above 0.8 is considered to be a good model. In our case, the ROC value for Random Forest was higher than for J48 signifying that Random Forest is the best option for classification/prediction model.

RECOMMENDATION

It is inferred that maximum cancellation happens in

- Seat class 3, the reason may be price is higher for certain customers, that they choose other airlines providing cheaper fare for the same class. Provide competitive deals.
- Among men, as it clearly signifies that the Airline is male dominant. More offers can be provided for women to attract female customers.
- Among the age group of 25, which is also the age bracket that provides the Airline with high revenue. Attract more customers within age group 25 by adding sky miles benefits and free upgrades.

RDF TRIPLES OUTPUT

RDF triples of random customer records.

```
<?xml version="1.0" encoding="utf-8" ?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
           xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
             xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
         xmlns="http://example.org/data/Airproject.csv#">
<rdf:Description rdf:about="http://example.org/data/Airproject.csv#row=6">
    <CUSTOMERID rdf:datatype="&xsd;integer">5</CUSTOMERID>
    <SUCCESS rdf:datatype="&xsd;integer">0</SUCCESS>
    <SEATCLASS rdf:datatype="&xsd;integer">3</SEATCLASS>
    <GUESTS rdf:datatype="&xsd;integer">0</GUESTS>
    <FARE rdf:datatype="&xsd;decimal">8.05</FARE>
    <TITLE xml:lang="en-ca"> Mr</TITLE>
    <GENDER xml:lang="en-ca">Male</GENDER>
    <AGE rdf:datatype="&xsd;integer">35</AGE>
</rdf:Description>
<rdf:Description rdf:about="http://example.org/data/Airproject.csv#row=90">
    <CUSTOMERID rdf:datatype="&xsd;integer">89</CUSTOMERID>
    <SUCCESS rdf:datatype="&xsd;integer">1</SUCCESS>
    <SEATCLASS rdf:datatype="&xsd;integer">1</SEATCLASS>
    <GUESTS rdf:datatype="&xsd;integer">3</GUESTS>
    <FARE rdf:datatype="&xsd;decimal">263</FARE>
    <TITLE xml:lang="en-ca"> Miss</TITLE>
    <GENDER xml:lang="en-ca">Female</GENDER>
    <AGE rdf:datatype="&xsd;integer">23</AGE>
</rdf:Description>
<rdf:Description rdf:about="http://example.org/data/Airproject.csv#row=192">
    <CUSTOMERID rdf:datatype="&xsd;integer">191</CUSTOMERID>
    <SUCCESS rdf:datatype="&xsd;integer">1</SUCCESS>
    <SEATCLASS rdf:datatype="&xsd;integer">2</SEATCLASS>
    <GUESTS rdf:datatype="&xsd;integer">0</GUESTS>
    <FARE rdf:datatype="&xsd;decimal">13</FARE>
    <TITLE xml:lang="en-ca"> Mrs</TITLE>
    <GENDER xml:lang="en-ca">Female</GENDER>
    <AGE rdf:datatype="&xsd;integer">32</AGE>
</rdf:Description>
<rdf:Description rdf:about="http://example.org/data/Airproject.csv#row=269">
    <CUSTOMERID rdf:datatype="&xsd;integer">268</CUSTOMERID>
    <SUCCESS rdf:datatype="&xsd;integer">1</SUCCESS>
    <SEATCLASS rdf:datatype="&xsd;integer">3</SEATCLASS>
    <GUESTS rdf:datatype="&xsd;integer">1</GUESTS>
    <FARE rdf:datatype="&xsd;decimal">7.775</FARE>
    <TITLE xml:lang="en-ca"> Mr</TITLE>
    <GENDER xml:lang="en-ca">Male</GENDER>
    <AGE rdf:datatype="&xsd;integer">25</AGE>
</rdf:Description>

</rdf:RDF>
```

REFERENCES

- White, S.K. (2018). What is a data scientist? A key data analytics role and a lucrative career. Referenced from https://www.cio.com/article/3217026/data-science/what-is-a-data-scientist-a-key-data-analytics-role-and-a-lucrative-career.html
- Jordi Girones. Data Mining with R. J48. Referenced from http://data-mining.business-intelligence.uoc.edu/home/j48-decision-tree
- Brownlee, J. (2016). How to perform feature selection with Machine learning data in Weka. Referenced from https://machinelearningmastery.com/perform-feature-selection-machine-learning-data-weka/
- Leo Breiman. Random Forests. Referenced from https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm
- Analytics Vidhya. (2016). Tutorial on 5 powerful R packages used for imputing missing values. Referenced from https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/
- Sanchez, G. (2012) 5 functions to do Principal components Analysis in R. Referenced from http://www.gastonsanchez.com/visually-enforced/how-to/2012/06/17/PCA-in-R/
- https://dvcs.w3.org/hg/rdf/raw-file/default/rdf-xml/index.html
- Larose, D.T (2014). Discovering Knowledge in Data -An Introduction to Data Mining