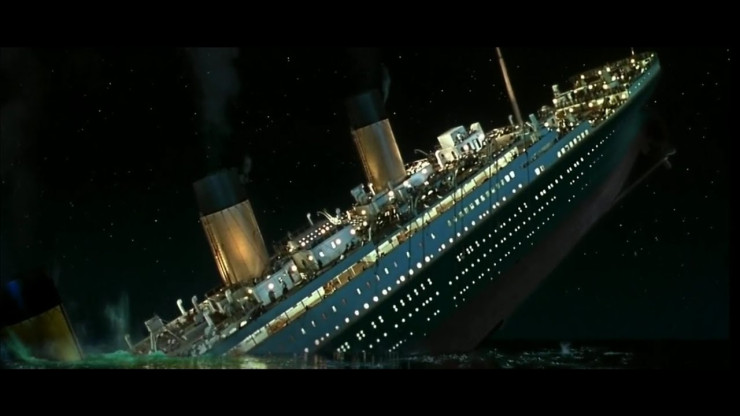
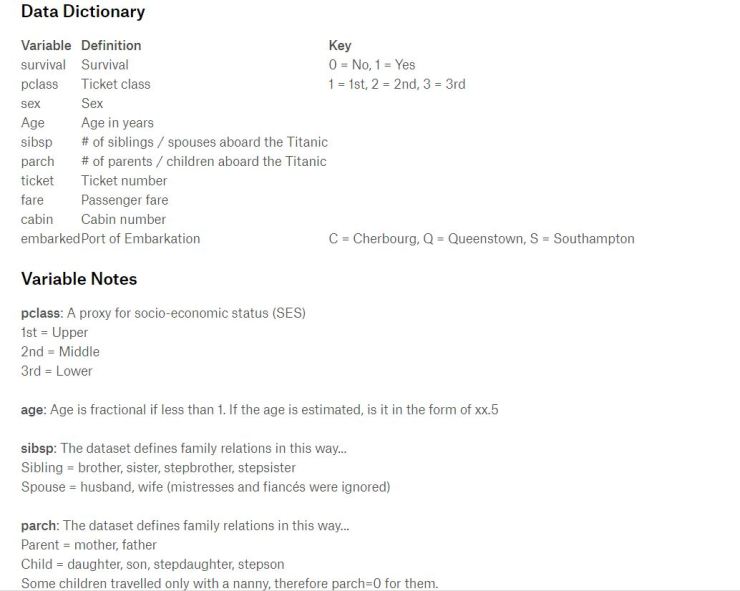
## **ML CLASSIFICATION PROJECT: TITANIC SURVIVOR PREDICTION**



# Background

The Titanic survivor prediction – was part of a Kaggle competition that was held a couple of years back.  The source dataset is [here](https://www.kaggle.com/c/titanic/data) (you may have to be a Kaggle member to access the data).

As usual there are two datasets : the training data and the testing data.  There are many fields in the two datasets :



The key predicted variable here is the Survived/Survival field.  If a person has been marked as “1” against this field – it means that he survived and if he receives a “0” it means that he did not survive.

# Step 1: Data Wrangling

When the training data is reviewed – it was noticed that a few records had a blank ‘Age’ value.  Based on this we carve two sub-datasets.  The first dataset is one without a blank ‘Age’ value.  We use this for the main training purpose.  The second dataset with blank ‘Age’ values is used for validation purposes.

The ‘Ticket’ field is then standardized into numeric – by stripping out all string data from the rows that have any string data in that field.

The ‘Fare’ field is reformatted into a two digit numeric field.  Similarly the ‘Sex’ and ‘Embarked’ fields are converted into categorical numerical values.

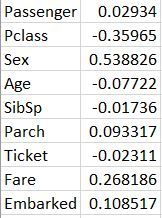
The following fields are also reformatted into numeric fields:

‘PassengerId’, ‘Survived’, ‘Pclass’, ‘Age’, ‘SibSp’, ‘Parch’

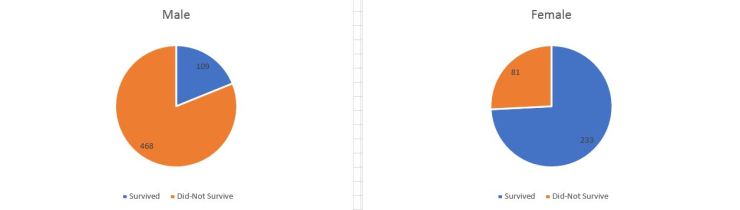
The fields ‘Name’ and ‘Cabin’ are not considered for training as I felt that they don’t add value to the training.

# Step 2: Data Analysis

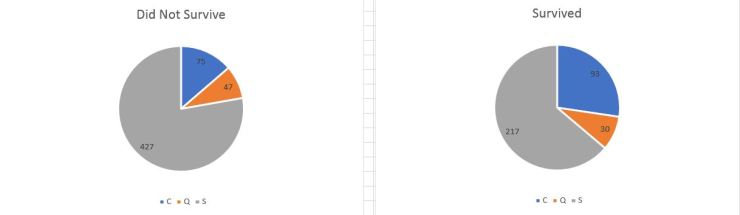
The Data Correlation was derived for all fields against the ‘Survived’ field.  The Excel CORREL formula was used to do this.



As it can be seen – the Sex/Gender of the individual has a very high correlation followed by Fare and Embarked.  Also it was seen in the data that about 74% of the females survived this tragedy.  However only about 19% males survived.



It was also seen that a high number of passengers, who had boarded in Cherbourg survived.  When this was further analysed – it was seen that there was a greater proportion of females who had boarded in Cherbourg (73 females and 95 males) as compared to other locations.



As can be seen in the above charts – Cherbourg was the only location where more passengers survived than those who did not survive (93 against 75).

Since Fare and Embarked also had a significant correlation on the survival – when we deep dived – the following trend emerged:

a. Women were given  higher priority to board a life-boat

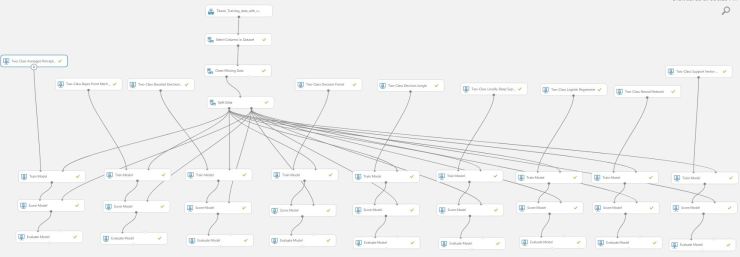
b. Among women – higher priority was given to those who had paid a higher fare

c. Men did not get any such concessions – although those who paid extremely high fares (just two of them) survived as compared to those who just paid high/low fares.

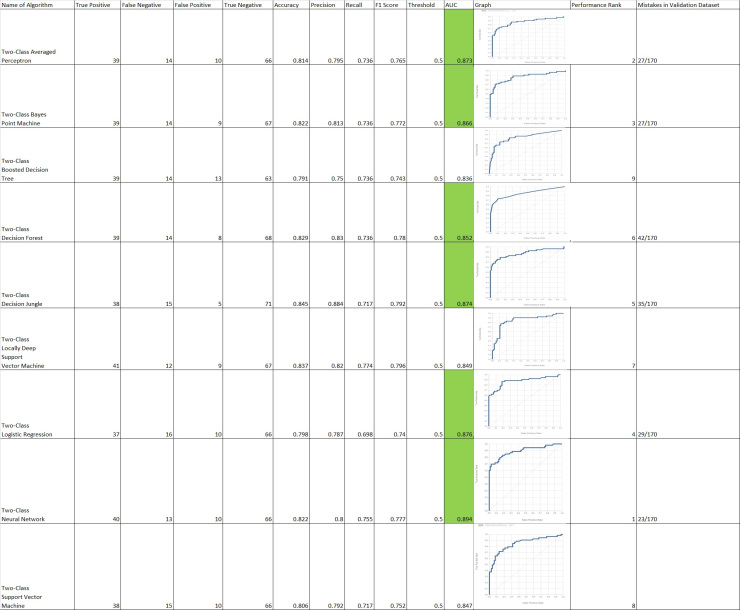
# Step 3: Training the Models

Using Microsoft Azure Machine Learning Studio – a host of classification models were trained against the wrangled training data (without the validation data – where the ‘Age’ field was blank).

The following algorithms are trained against the training data: 1. Two-Class Averaged Perceptron 2. Two-Class Bayes Point Machine 3. Two-Class Boosted Decision Tree 4. Two-Class Decision Forest 5. Two-Class Decision Jungle 6. Two-Class Locally Deep Support Vector Machine 7. Two-Class Logistic Regression 8. Two-Class Neural Network 9. Two-Class Support Vector Machine d. Out of all these – the top three were found to be the Neural Network followed by the Averaged Perceptron and the Bayes Point Machine



The complete Azure project can be found [here](https://gallery.cortanaintelligence.com/Experiment/Titanic-Survival-Prediction-4) on Cortana Gallery.   The metrics for each of the above models was found to be as below:



Based on the above metrics – I downloaded the web service excel for each of the above models.  I ran each model against the validation data (where ‘Age’ was blank).  As mentioned earlier the Two-Class Neural Network came with a good score of only 23 errors out of 170 as compared to the next two (Averaged Perceptron and the Bayes Point Machine) which scored 27 errors out of 170.  The others scored poorly on all metrics. Hence the Two-Class Neural Network was chosen as the algorithm of choice to solve this problem.

# Step 4: Predicting for the test data

[titanic\_test\_predictions](https://rajivsworklife.files.wordpress.com/2018/01/titanic_test_predictions.xlsx)

The above file contains the predictions for the testing data using the web service excel for the Two-Class Neural Network classification algorithm.

# Additional Project

A regression project was done to predict ‘Age’ values in the validation dataset.  The entire project is [here](https://gallery.cortanaintelligence.com/Experiment/Titanic-Age-prediction) on Cortana Gallery.  Liner Regression had a 26% accuracy (the highest relative to other regression algorithms).   However since Age did not have a high correlation to the survival rate – this project was done as a minor addendum.

# Python Code

The entire project was again redone using Python code. To reduce the wrangling effort – the same input files that fed the azure project were fed into the Python code. This time the models that were derived were the Decision Tree and the Random Forest. Again there were two versions of the Decision Tree – with the difference being in the max\_depth parameter.

Between the two Decision Tree and the Random Forest models the best was the Random Forest with a score of 96% accuracy and a confusion matrix of :

array([[424, 0],

[ 22, 268]])

The better of the two Decision Trees had a score of 93% accuracy and a confusion matrix of

array([[415, 9],

[ 36, 254]])

So the Random Forest model was run against the validation data (with 177 records). There were 22 errors out of 177 records – which was better than the 23 out of 170 for the Neural Network model.