STUDENT NAME: NESTOR ROMERO LEON

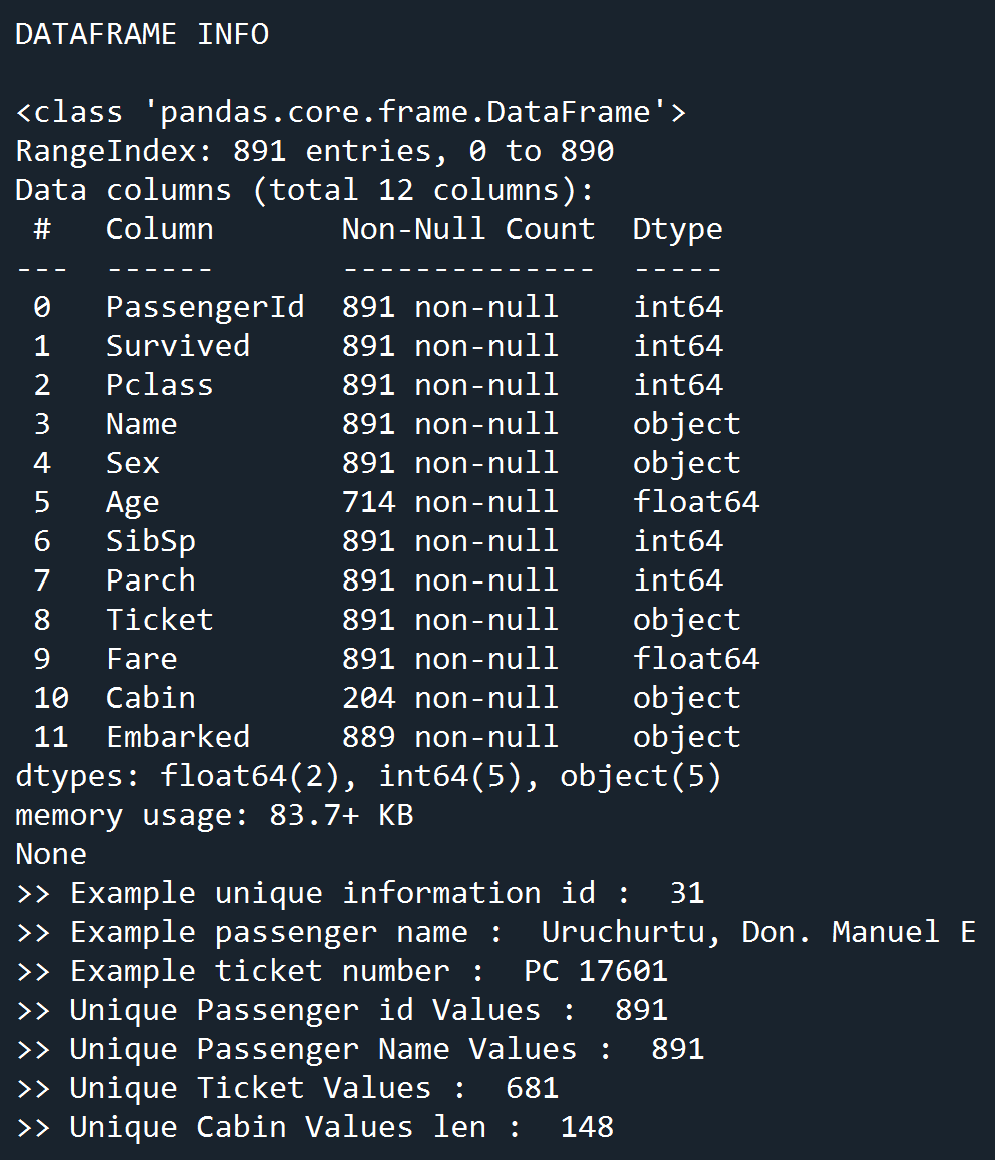
STUDENT ID: 301133331

**INTRODUCTION TO ARTIFICIAL INTELLIGENCE**

**LOGISTIC REGRESSION – ASSIGNMENT**

**EXERCISE 1**

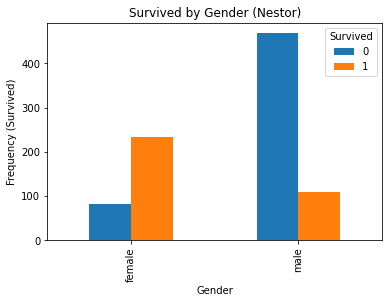
**PART B. NUMERAL 4 – COLUMNS TO REMOVE**

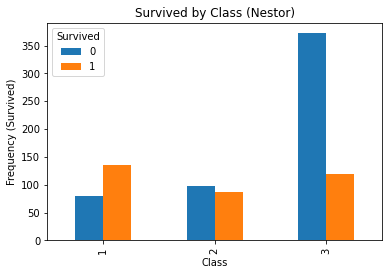


According to the data frame information, the columns to remove will be:

PassengerId, Name, Ticket, and Cabin

The first 3 variables have mostly unique values for each record and the Cabin variable has too many missing values.

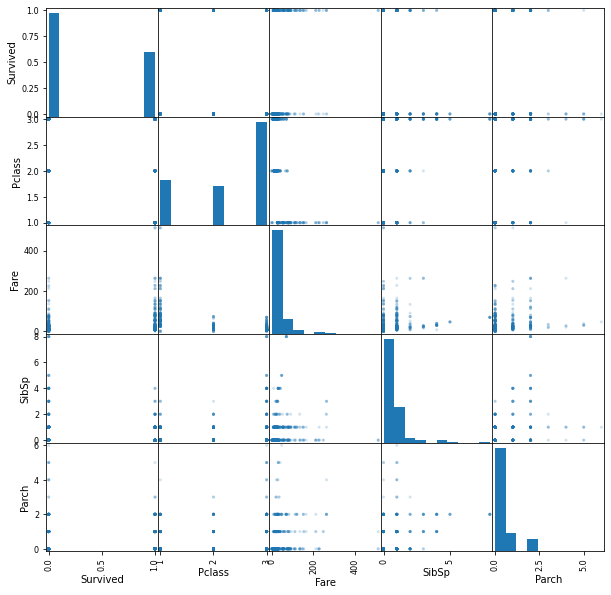
**PART C. DATA VISUALIZATION**



In the first plot (Survived by Class) the number of survivors is somewhat similar across all classes and around 100. However, the most deaths occurred in the 3rd class in towering numbers compared to classes 1 and 2.

The second plot (Survived by Gender) shows that the number of women that survived was roughly 2.5 times the male survivor count. On the other hand, in terms of deaths, there were nearly 5 times more deaths for men than for women.

**SCATTER MATRIX**

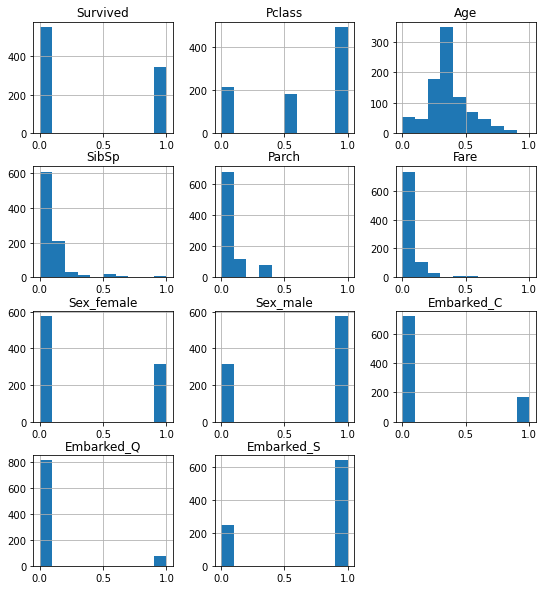


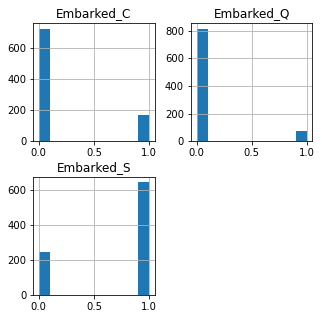
The main diagonal of the matrix that compares each variable vs itself, shows the distribution of the data for each variable. The plots cells comparing variables with discrete values show the data grouping either in columns or rows depending on the comparison being done above or under the diagonal as expected with the pivoted axis. On this particular view, the variables fare, siblings/spouses and parent/children show a similar shape. However, there is not a particularly visible “correlation” of variables in the plot

The plot is helpful in determining not only the approximate shape of the variables but also at detecting ranges for the values and spotting possible outlier values. This plot also helps to raise questions regarding preconceptions regarding the data, for instance, the appearance of very low fare passengers on first class, in fact, all three classes appear to have comparable values when you would expect more clear “buckets” of fares.

**PART C. DATA TRANSFORMATION**

**HISTOGRAMS FOR ALL COLUMNS**



**HISTOGRAMS FOR PORT OF EMBARKMENT VARIABLES**

With the created histograms after the data normalization process, we can clearly that most of the passengers embarked in Southampton. Two thirds approximately corresponding more than 600 passengers over the total of 891

**PART E. MODEL BUILDING**

DISPLAY CROSS-VALIDATION RESULTS

Test Size Min Mean Max Range

Test Size: 0.1000 || Metrics: 0.7250 0.7965 0.8875 0.1625

Test Size: 0.1500 || Metrics: 0.7467 0.8005 0.8421 0.0954

Test Size: 0.2000 || Metrics: 0.7361 0.8006 0.8592 0.1230

Test Size: 0.2500 || Metrics: 0.7121 0.8038 0.8507 0.1386

Test Size: 0.3000 || Metrics: 0.7097 0.7978 0.8871 0.1774

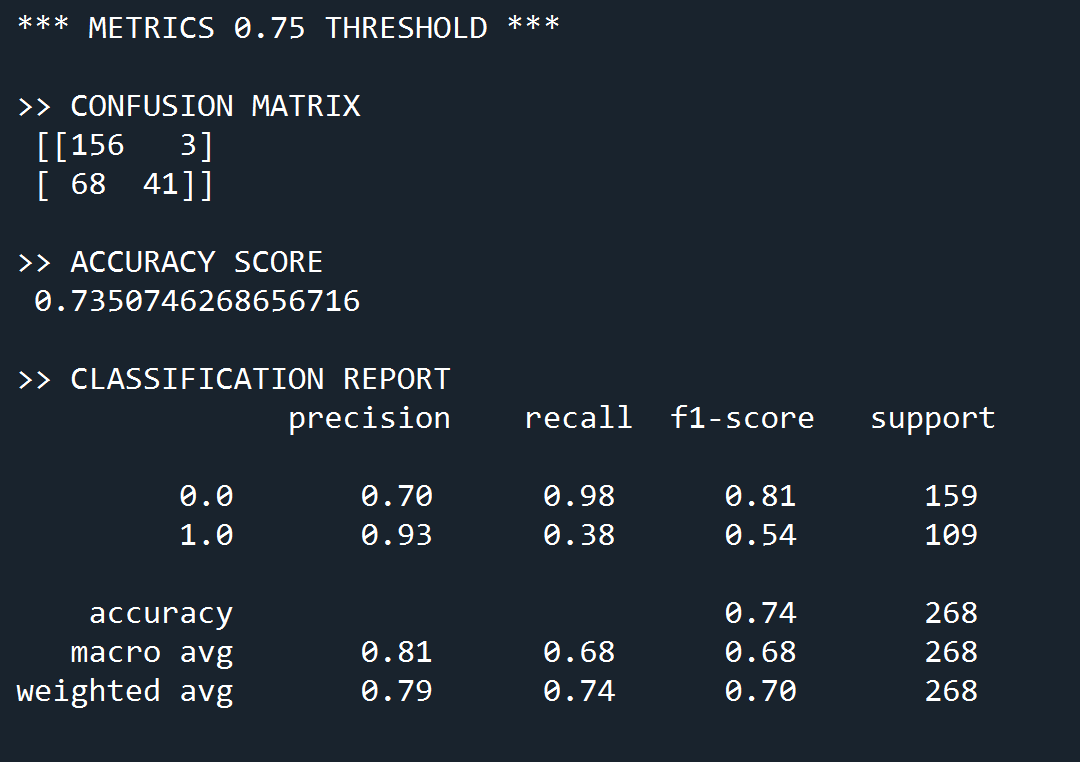
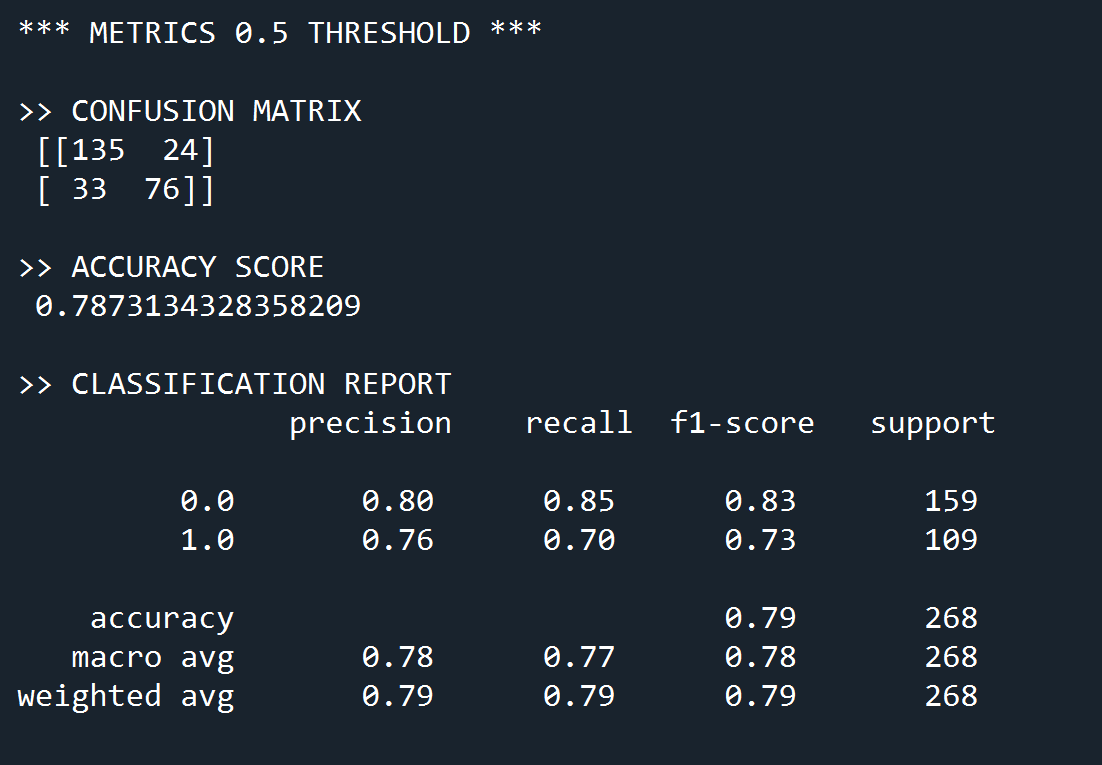
Test Size: 0.3500 || Metrics: 0.7241 0.7928 0.8448 0.1207

Test Size: 0.4000 || Metrics: 0.7170 0.7901 0.8889 0.1719

Test Size: 0.4500 || Metrics: 0.6939 0.7959 0.8980 0.2041

Test Size: 0.5000 || Metrics: 0.7045 0.7929 0.8667 0.1621

With these results, we can observe that the metrics for the test size = 25% provides the higher values for accuracy in the mean with a narrow range (min,max) compared with test sizes that got higher max values but have higher spreads such as test size = 40% or 45%. These results show a high accuracy zone between 15% and 25% test size

**PART F. MODEL TESTING**

**ACCURACY TEST DATA VS TRAINING DATA**

During the cross-validation stage, the results for all splits analyzed were higher than 0.79 on average. On this final test, only the 0.5% threshold produced results close to what was expected in terms of accuracy.

**METRICS COMPARISON THRESHOLDS 0.5 AND 0.7**

By taking into account the axis used by scikit learn to produce the matrix we can analyze the results properly. For instance threshold(0.5) has 76 True Positives TP and threshold(0.7) has 41 TP. This is a little different from what we saw in class but only in terms of the organization of the data in the matrix but not the metrics.

For the final tests performed while varying the threshold value, the results showed that the threshold(0.5) produces higher accuracy and recall values than the threshold(0.7). At the same time threshold(0.7) produces much higher precision for the positive class (0.93) but also almost extreme values for the recall (0.38 - 0.98) in the positive and negative class. These extreme values indicate that the number of True Positives produced was high (small win) in contrast with the few times overall that the results were actually correct (big loss).

With these results, the threshold(0.5) appears to be preferable with more balanced predictions. The threshold(0.7) would seem better at predicting the negative class, but in this case, this is a warning sign considering that the basic distribution of the data showed that 2/3 of the passengers died so this model might not beat a random guess.