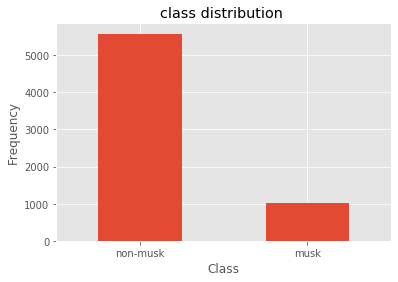
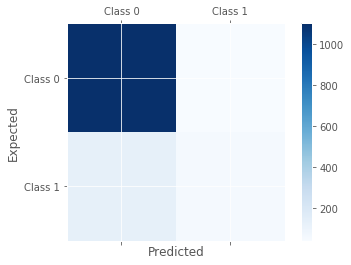
**Pre-processing the data:**

1. Finding any missing values in the dataset
2. Converting numerical data to non-numerical using a function handle\_non\_numerical\_data()
3. Finding the most important features using Extra trees Classifier. Even the most important features had correation values less than 1. So, all the features were used in training the model.
4. Checking the dataset imbalance:



The dataset was imbalanced. We have more features for class non-musk over class musk. In a dataset with highly unbalanced classes, if the classifier always "predicts" the most common class without performing any analysis of the features, it will still have a high accuracy rate.

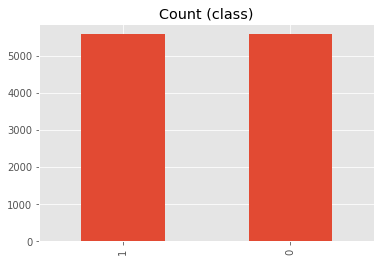
The confusion matrix was created which showed the values of predicted class’s vs. the expected values.



Class 1 has some false negatives and false positives showing false predictions.

1. Re-sampling the dataset

To deal with data imbalance re-sampling was performed. Over-sampling means that the copies of the class musk will be created in order to make the classes have same number of records. This can cause over-fitting. But under-sampling resulted in high accuracy but lower F1- scores due to a result in loss of information. So, over-sampling was chosen and the count of both the classes was made equal.



1. Feature scaling the features was performed to handle highly varying values.

**Model Description :**

A deep neural network was created with one input layer, 2 hidden layers and one final output layer. The activation function used for hidden layer was relu while after the final output layer, sigmoid activation was used.

In addition, 2 dropout layers were used with the peobability of 0.5, after each hidden layer.

