Model Choice – Random Forest Classifier random oversampling applied

* Supervised learning (knowledge of our values/classifications)
* Confusion Matrix
* Test, Training split – 75/25
* 15 features from 20 – split datetime due to having string information
* Oversampling because of having an imbalanced dataset – also tested undersampling
* Most important features [year(time measurement), humidity, visibility]
* Humidity %, visibility (km),

For the machine learning portion of our project we tested supervised learning methods and our model choice of all the models we tested was the Random forest classifier with random oversampling applied due to its performance in comparison of the other models. We used the random oversampling class because our dataset was imbalanced with 82% of the data belonging to one class. In our model we used 15 of the 20 features in our dataset with the most important features being one of our measurements of time, humidity, and visibility. Our accuracy score in the end came out to be 94% and we had an F1 score of 93.