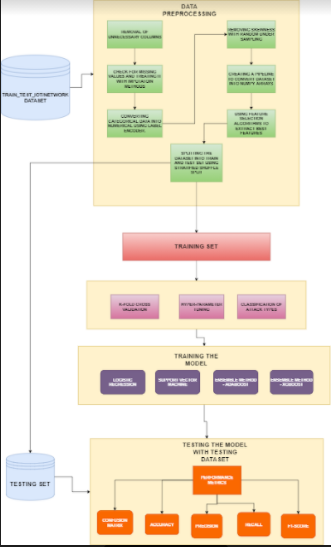
Review-2 Document-ISAA



**Link for datasets:**

<https://www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15-Datasets/bot_iot.php>

**Description:**

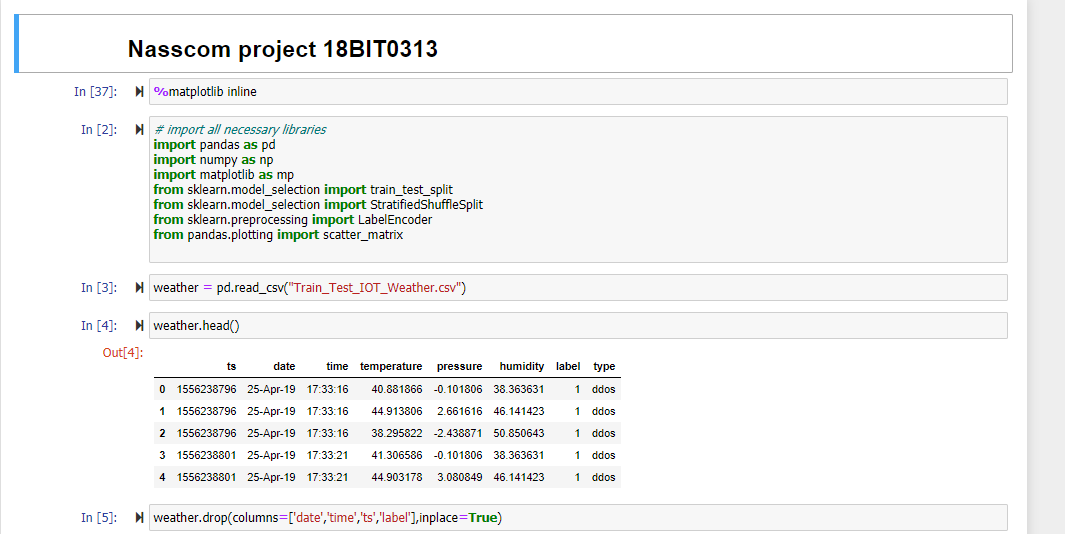
In this project, we have built machine learning models to classify and detect various types of intrusions like, DDoS, ransom ware, Backdoor, Injection, Password, XSS, MITM, scanning, DoS, etc. We have tried to visualize and understand the datasets. The dataset was preprocessed and the best features were extracted using feature selection algorithms. After this the data was split into train and test sets using stratified shuffled split.

We trained our models and then tried to improve them using hyper-parameter tuning and K-fold cross validation.

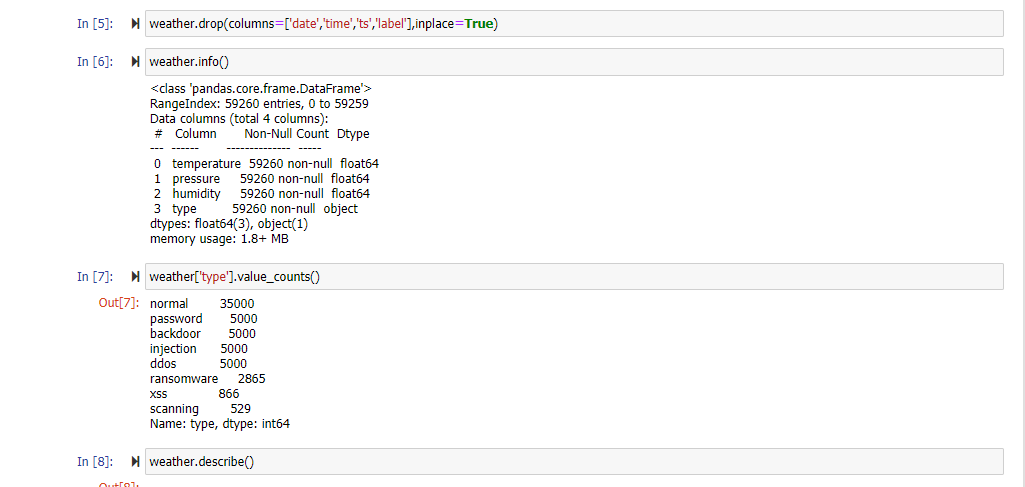
We tested our models and determined various performance metrics like accuracy, confusion matrix, precision, recall, F1-score. At last after training and testing of the various models with the dataset, we will compare the models performance and find the most efficient of them all.

*IOT Weather*

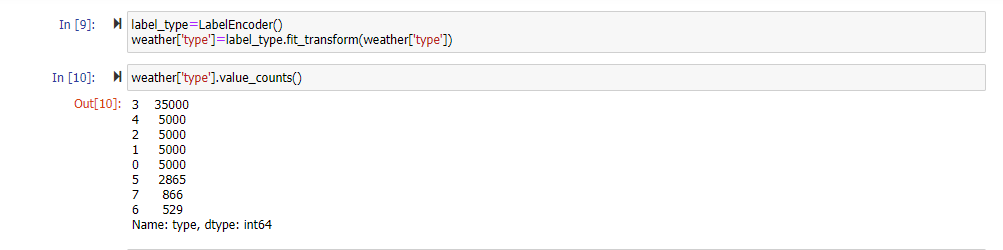
1) Include required packages and import the dataset into a pandas dataframe



2) Begin preprocessing by dropping redundant columns



3) Use label encoding to encode type to integer



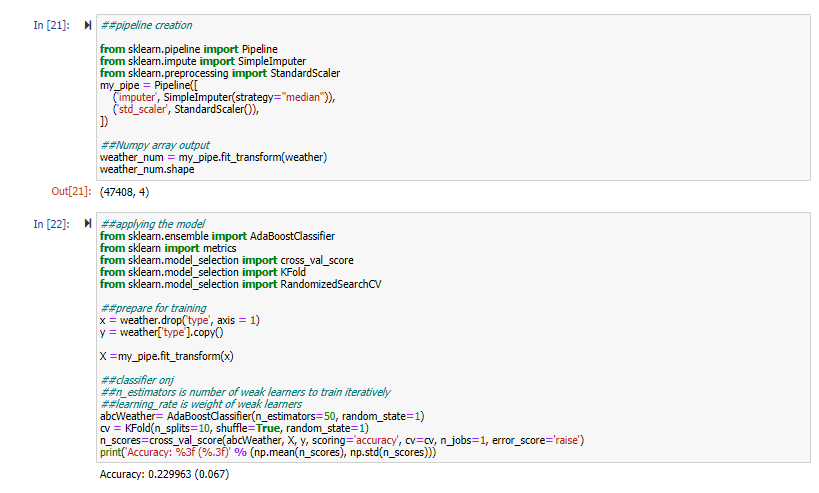
4) Using train\_test\_split and stratified\_split to split the df



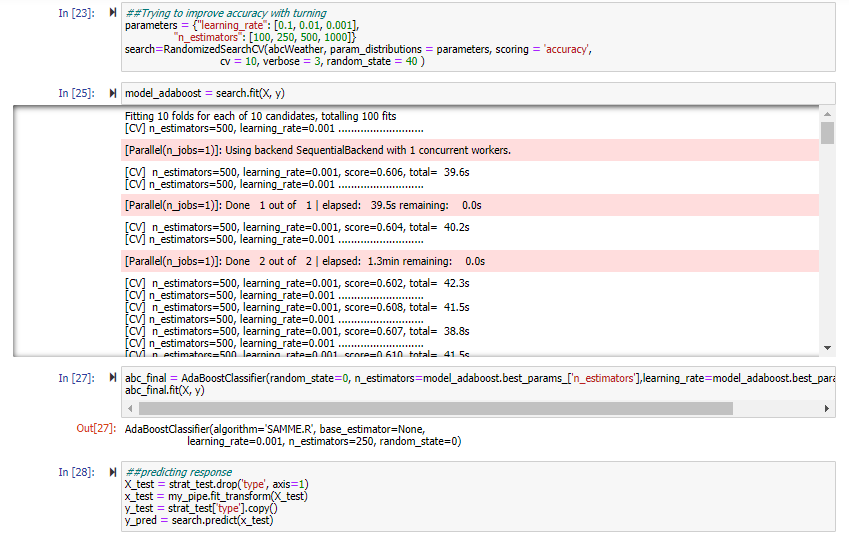
5) Study correlation and scatter matrix

****

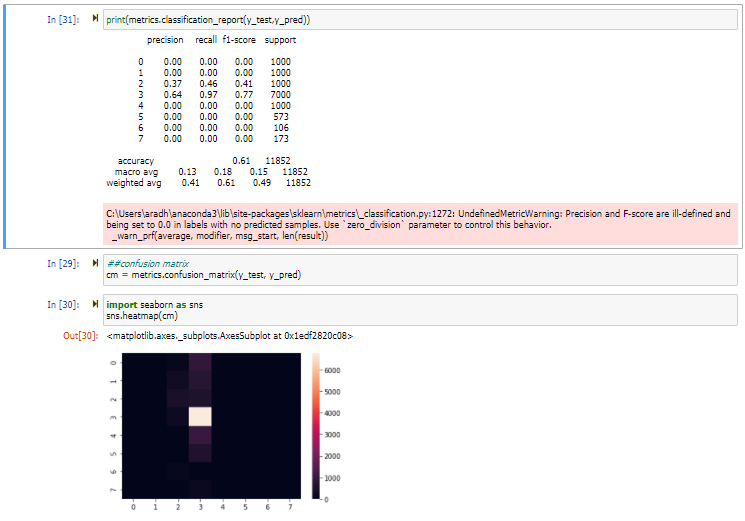
6) Create pipeline for numpy array and create basic adaboost model. Check for accuracy



7) Try to improve accuracy with hyperparameter tuning



8) Check results with metrics and confusion matrix

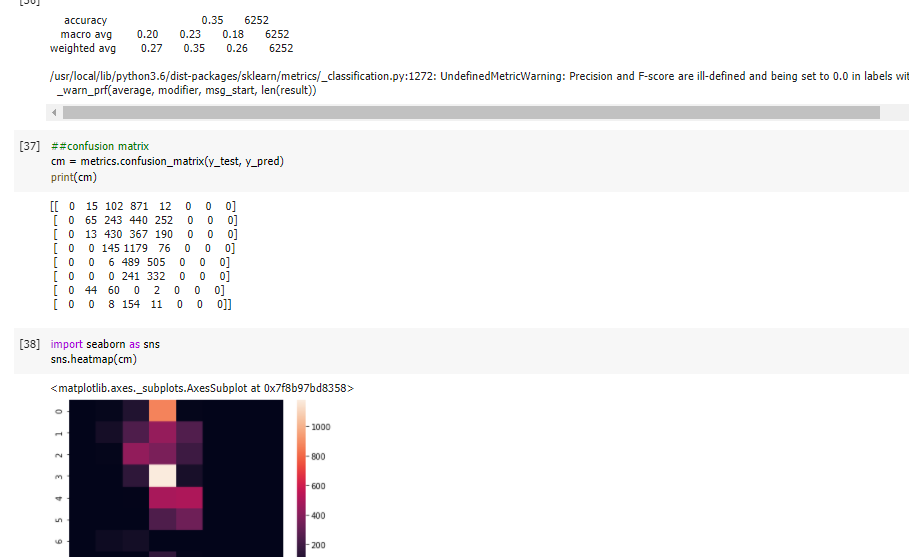


Accuracy is 0.61 but confusion matrix is not ideal

A) After reducing skewness due to ‘normal’ types

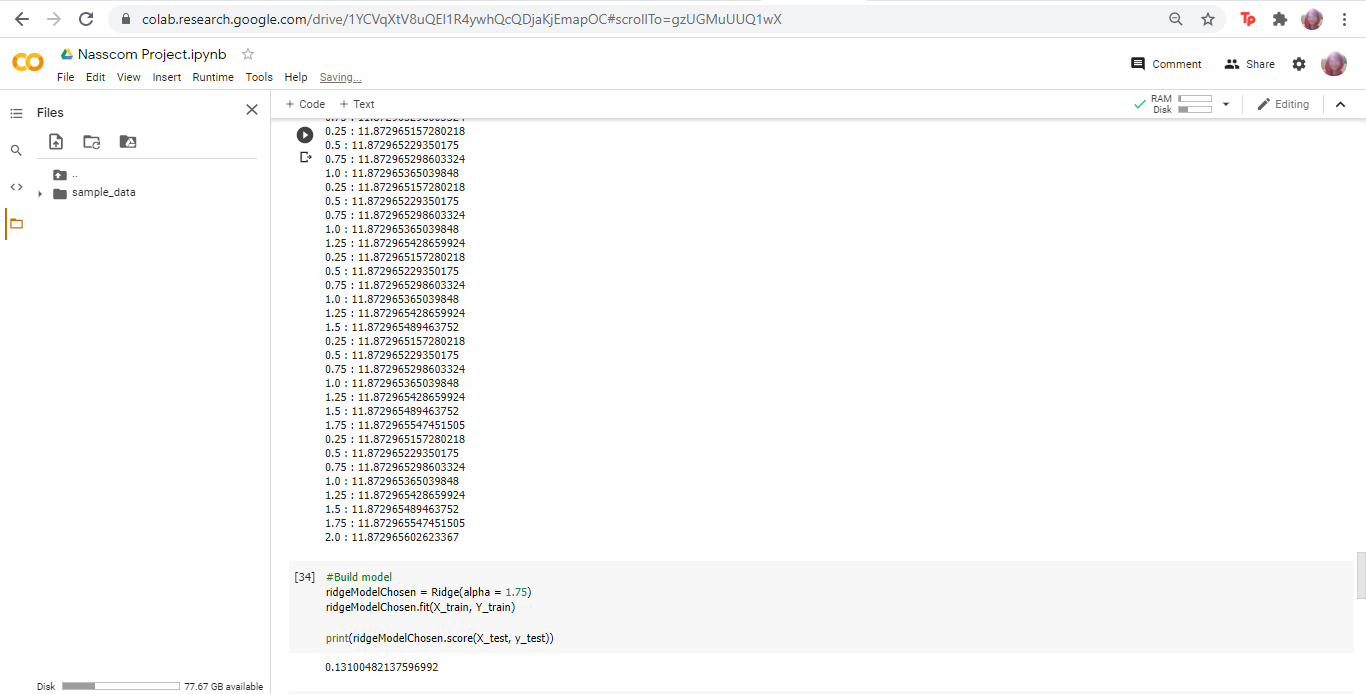


With hyperparameter tuning

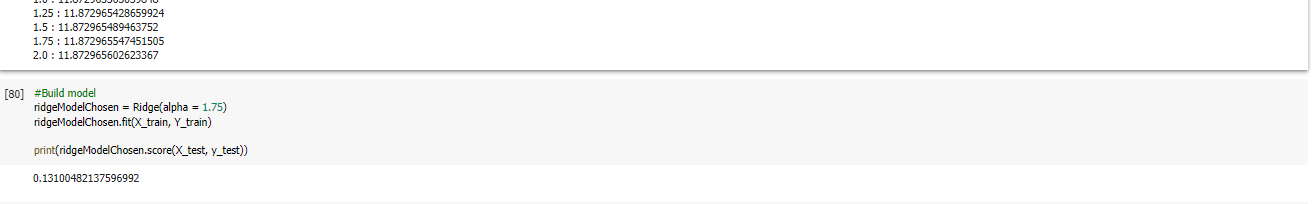


Accuracy decreased but confusion matrix is improved

B) Trying Ridge regularization

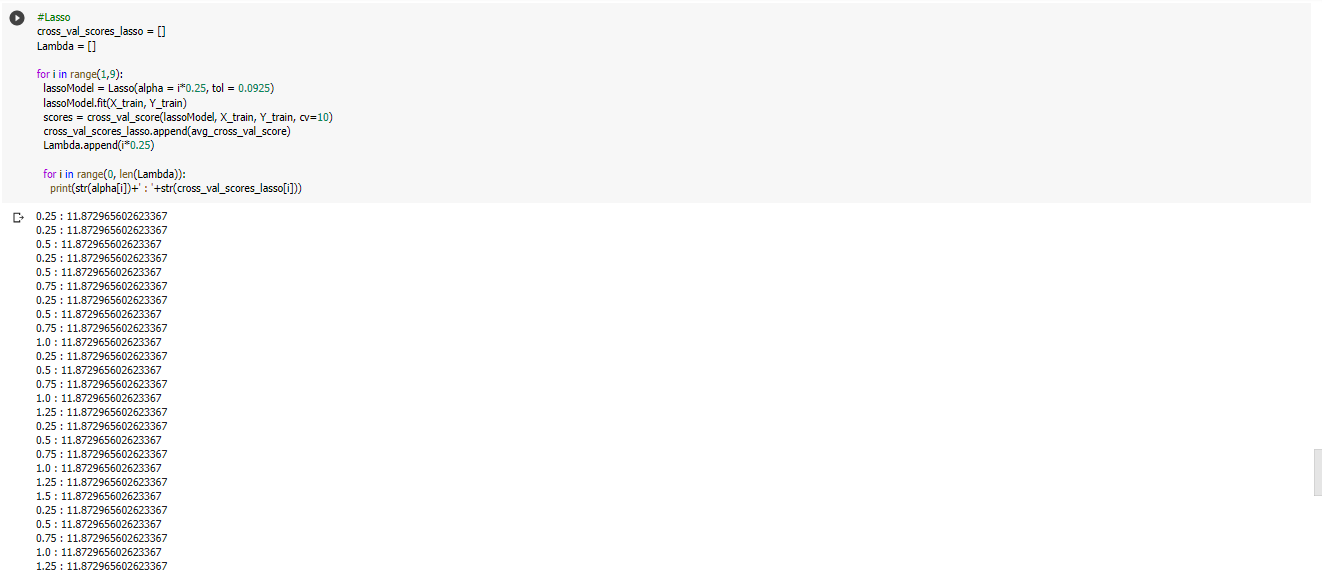


Result

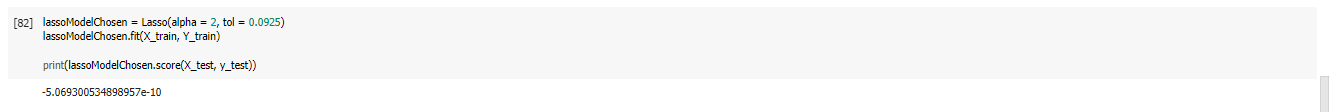


Accuracy has not improved

C) Trying Lasso regularization



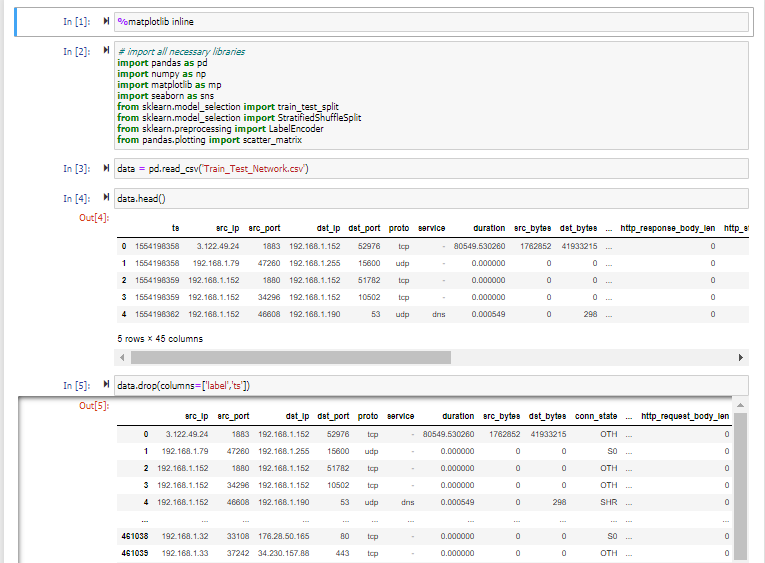
Results



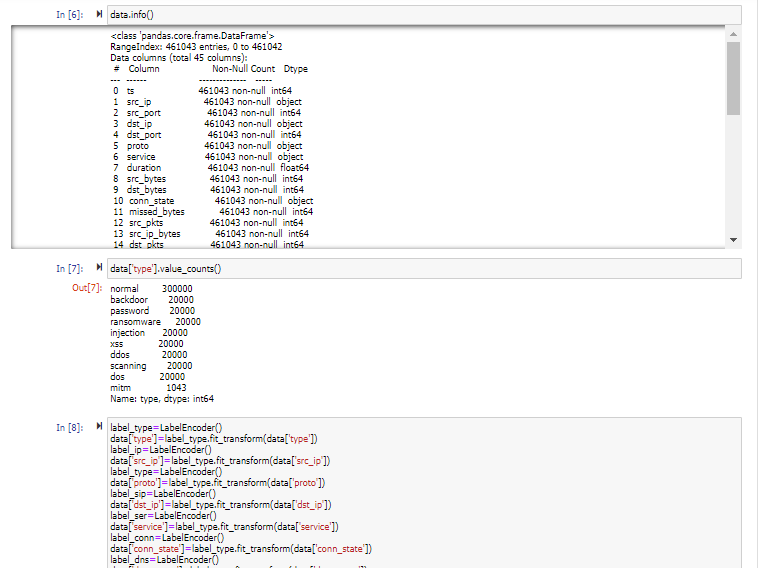
Accuracy has really decreased, dataset is not ideal for regularization

*IOT Network*

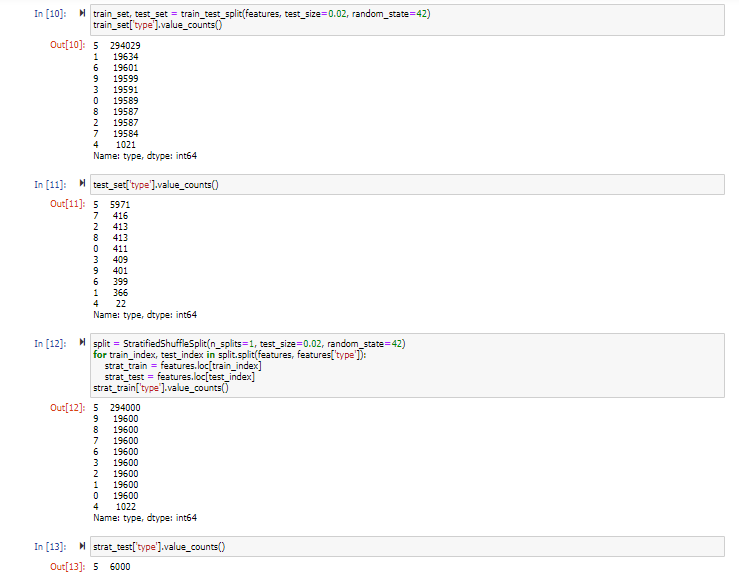
1) Importing dataset and necessary packaging, dropping redundant columns



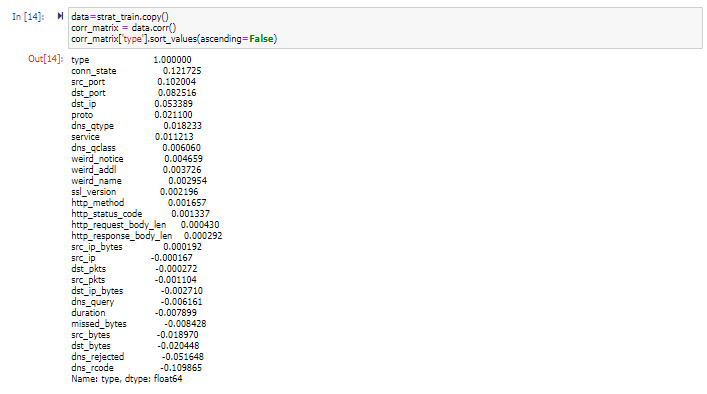
2) Studying dataset and using labelEncoding for all non-numerical columns, selecting only relevant sounding columns for the dataset



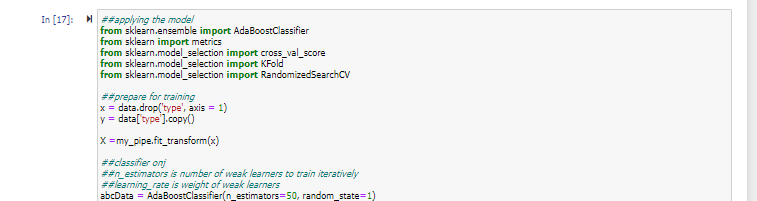
3) Splitting the dataset



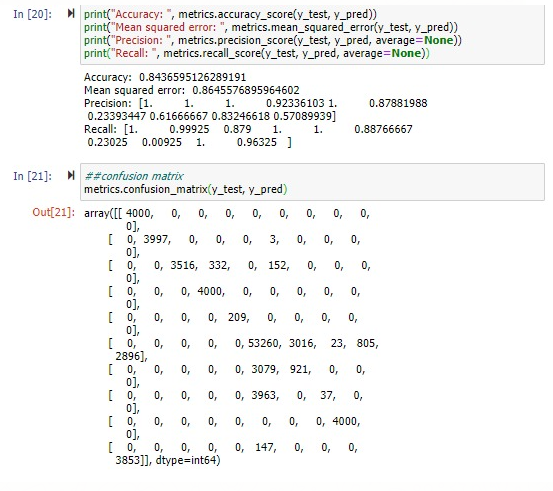
4) Checking correlation

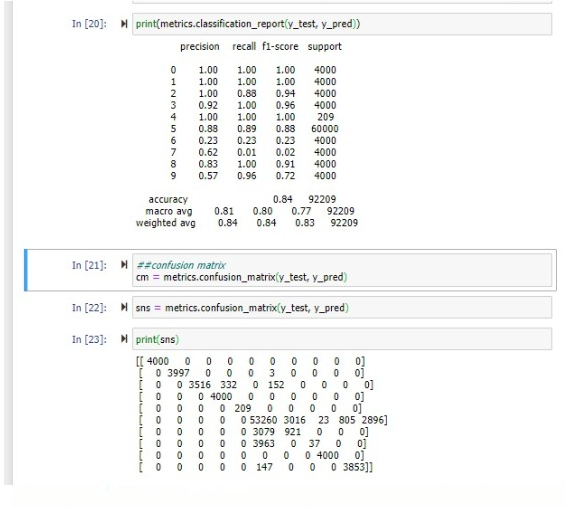


5) Pipeline creation, creating adaboost model



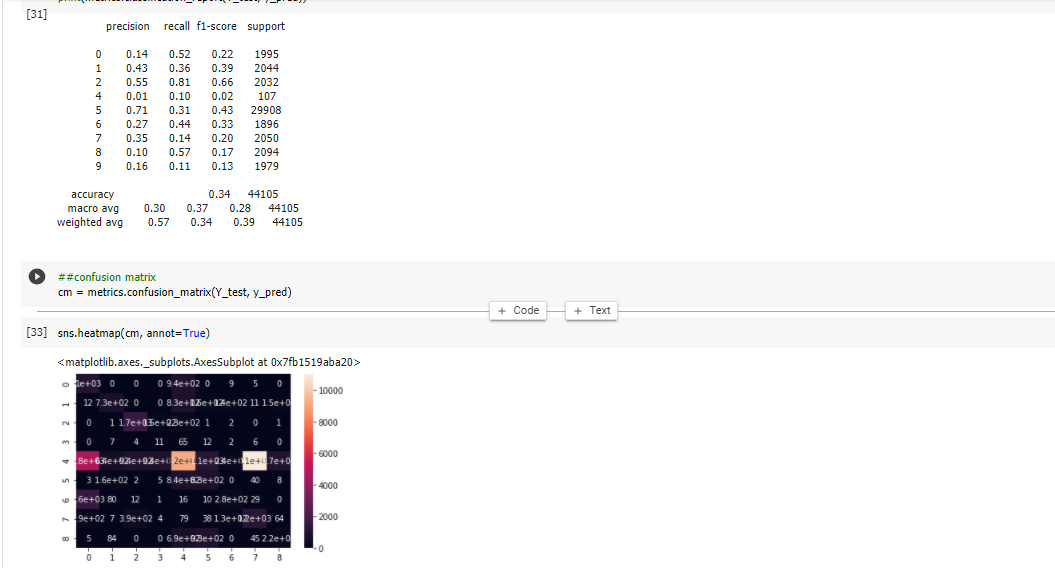
6) Results without any further preprocessing



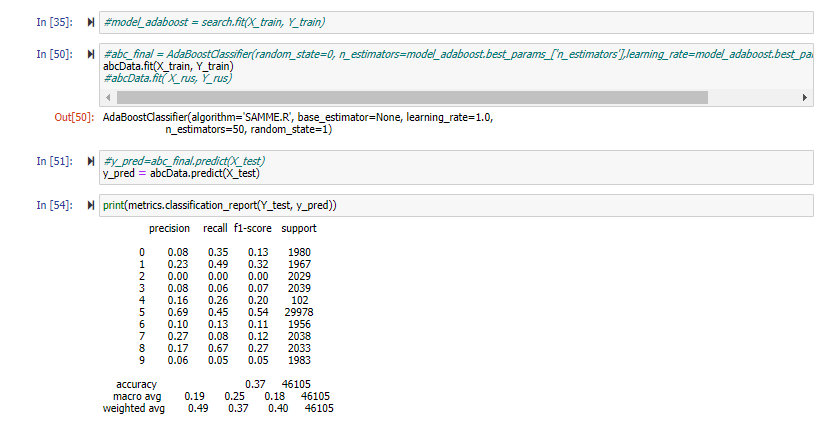


Results are very good, testing with other preprocessing to check. I am not expecting accuracy to increase without overfitting

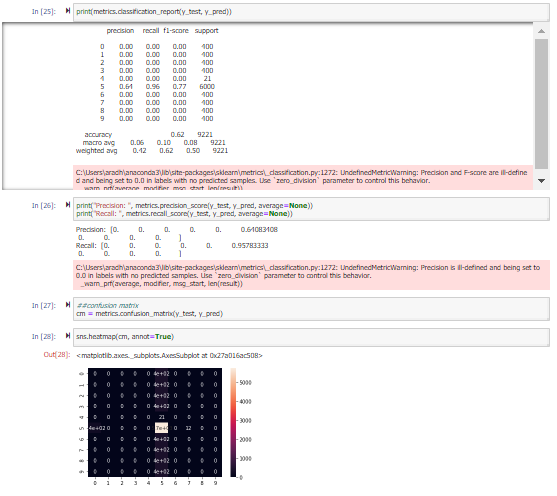
A) Without stratified shuffle split, with reduction of “normal” to 7000 from 35000, with HPT, without rus, test size = 0.1



B) With only RUS and no hyper parameter tuning



C) With hyperparameter tuning after removing skewness of dataset



C) RUS without Hyper Parameter Tuning

* Results did not greatly improve, however the correlation matrix was a little more regular

D) RUS with Hyper Parameter Tuning

* Less than .85 accuracy but good correlation matrix

***Final results***