

San Francisco Crime Classification

By: Nourhan Waleed



Email: [nouraaboelsoaoud@yahoo.com](mailto:nouraaboelsoaoud@yahoo.com)

Notebook Link: [Link](https://www.kaggle.com/nourhanaboelsoaoud/nourhanwaleedfinalproject)

**Main Charts/Plots:**

1.Correlation Heat Map(Training-Data):

Between all features to show which columns depend on which, and to what extent.

Graphical user interface

Description automatically generated

2.Correlation Heat Map(Testing Data):

Between all features to show which columns depend on which, and to what extent, for training data

Graphical user interface

Description automatically generated

3.Bar-Plot

To see the frequency of each crime

Graphical user interface, application

Description automatically generated

4.Bar Plot:

To see which PdDiscrit (Training-Data)has the most crimes

Graphical user interface, application

Description automatically generated

4.Bar Plot:

To see which PdDiscrit (Testing-Data)has the most crimes

Graphical user interface

Description automatically generated with medium confidence

5.Number of Crimes per District:

Graphical user interface, text

Description automatically generated

6.Different Crimes Occurrence Per Hour:

Graphical user interface, application

Description automatically generated

7. San Francisco Crimes Trend by Month &Year(Training Data):

Graphical user interface, text, application

Description automatically generated

8. San Francisco Crimes Trend by Month &Year(Testing Data):

Graphical user interface, application

Description automatically generated

9.Crime Rate by Day Of the Week(Training Data):

Graphical user interface, application

Description automatically generated10.Crime Rate by Day Of the Week(Testing Data):

Graphical user interface, application, table, Excel

Description automatically generated

**Data cleansing and features engineering steps:**

-Checked for null values

-Looked for unique values in each column

-Box-plots for numerical-value columns to check for outliers ,found some in the X,Y columns and fixed them

Before:

Graphical user interface, application

Description automatically generated

After:

Graphical user interface

Description automatically generated

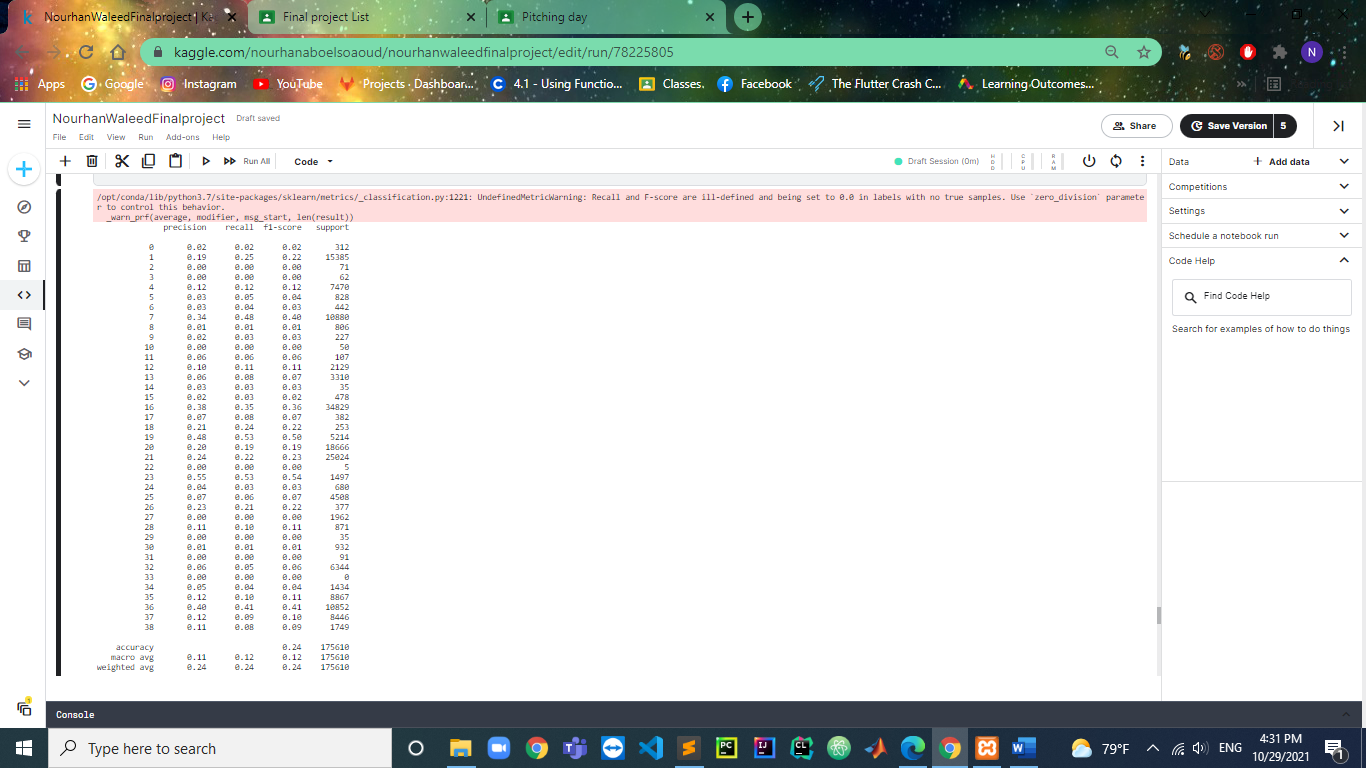
A screenshot of a computer

Description automatically generated

-Then I dropped the columns which would have no effect on classification(Address, Resolution,…….)

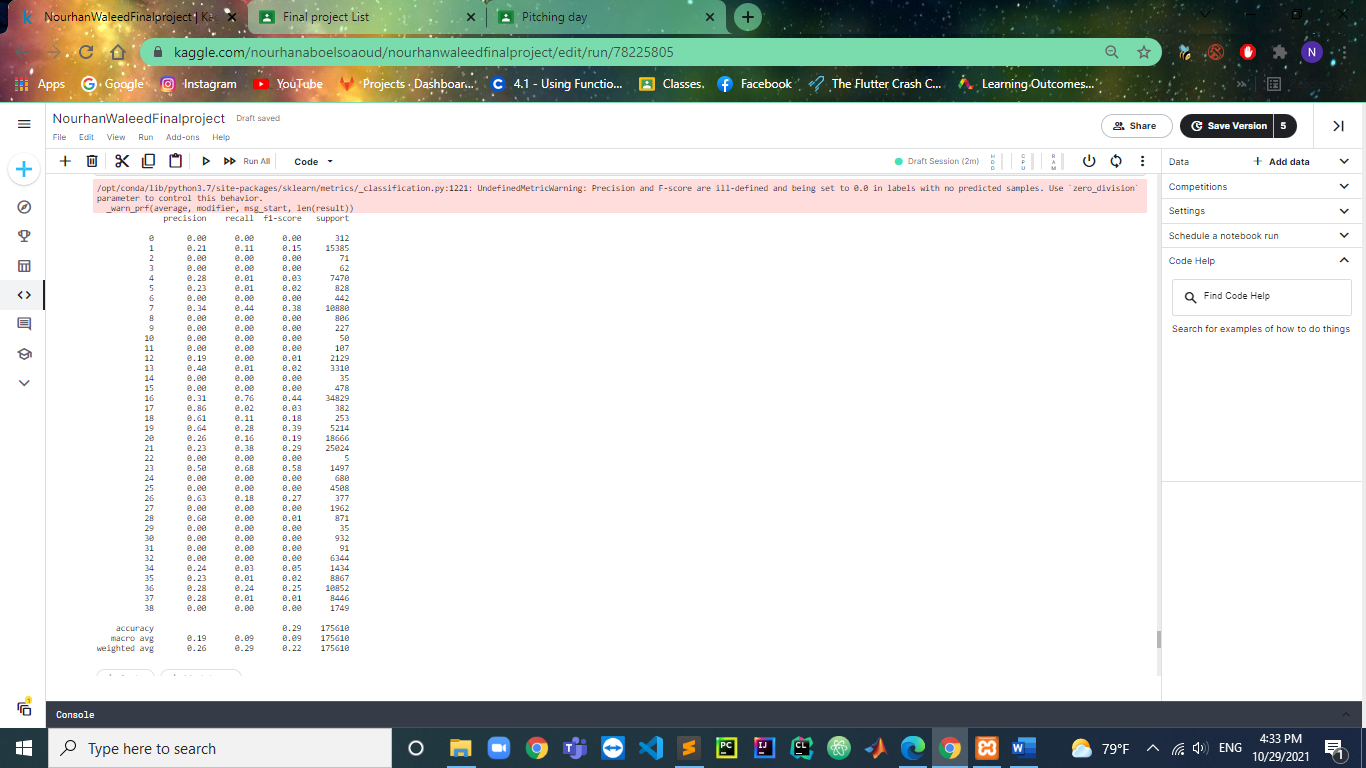
**Models, Evaluation and Tuning:**

Classification Report for Model 1 (Decision Tree):



Model Accuracy: 0.24

Classification Report for Model 2 (Random Forest Classifier):



Average Model Accuracy: 0.29

Train Accuracy: 0.3329741087838232

Test Accuracy: 0.28891862650190764

After Tuning:

Train Accuracy: 0.34567983839166105

Test Accuracy: 0.29070098513752063

Classification Report for Model 3(SVM):

Graphical user interface, text, application

Description automatically generated

Model 3 Accuracy(SVM): 0.19924263994077787

After Tuning: 0.20007402767496157

Best Model Out of the 3:

Random Forest Classifier as it produces the highest accuracy in a short time

**Future work that may be done to enhance the models:**

Give them more data to produce higher accuracy