IML Hackathon 2019

Participants:

Or Nokrean (206223166)

David Nir (203487293)

Noam Delbari (315005066)

Idan Gabay (312415706)

Task: 2

We started by taking a look at our data (which was big).

We immediately noticed that there are some useless features there, such as the number of the row in the data file.

We also suspected that the fields “Updated on”, “ID”, wards/ward, and checked how they affected the results of our models.

We split the “Date” field into three new features: “Day”, “Month” and “Hour” (which combines the minutes and the AM/PM part of the “Date”).

We dropped the columns we deemed not helpful.

We applied a map to the entire data matrix to change Boolean values to integers.

We then encoded the “Block” column into dummy parameters.

We also replaced all naN’s with the median of the column.

We split some columns into categorial columns.

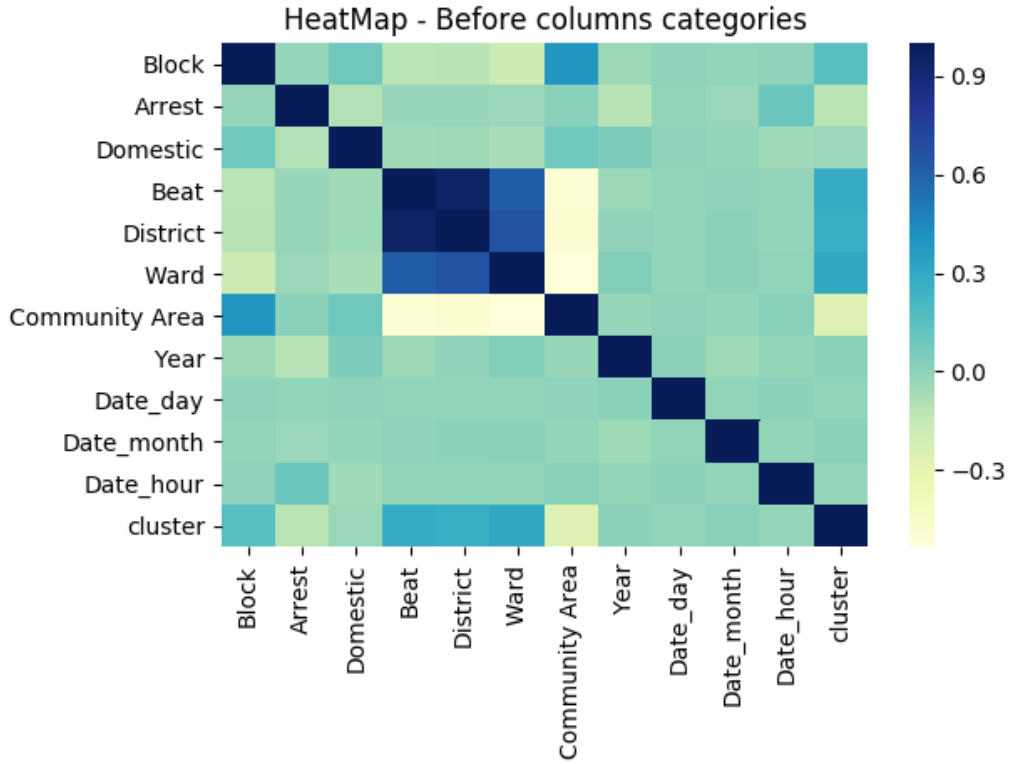
We tried running the data with the following algorithms:

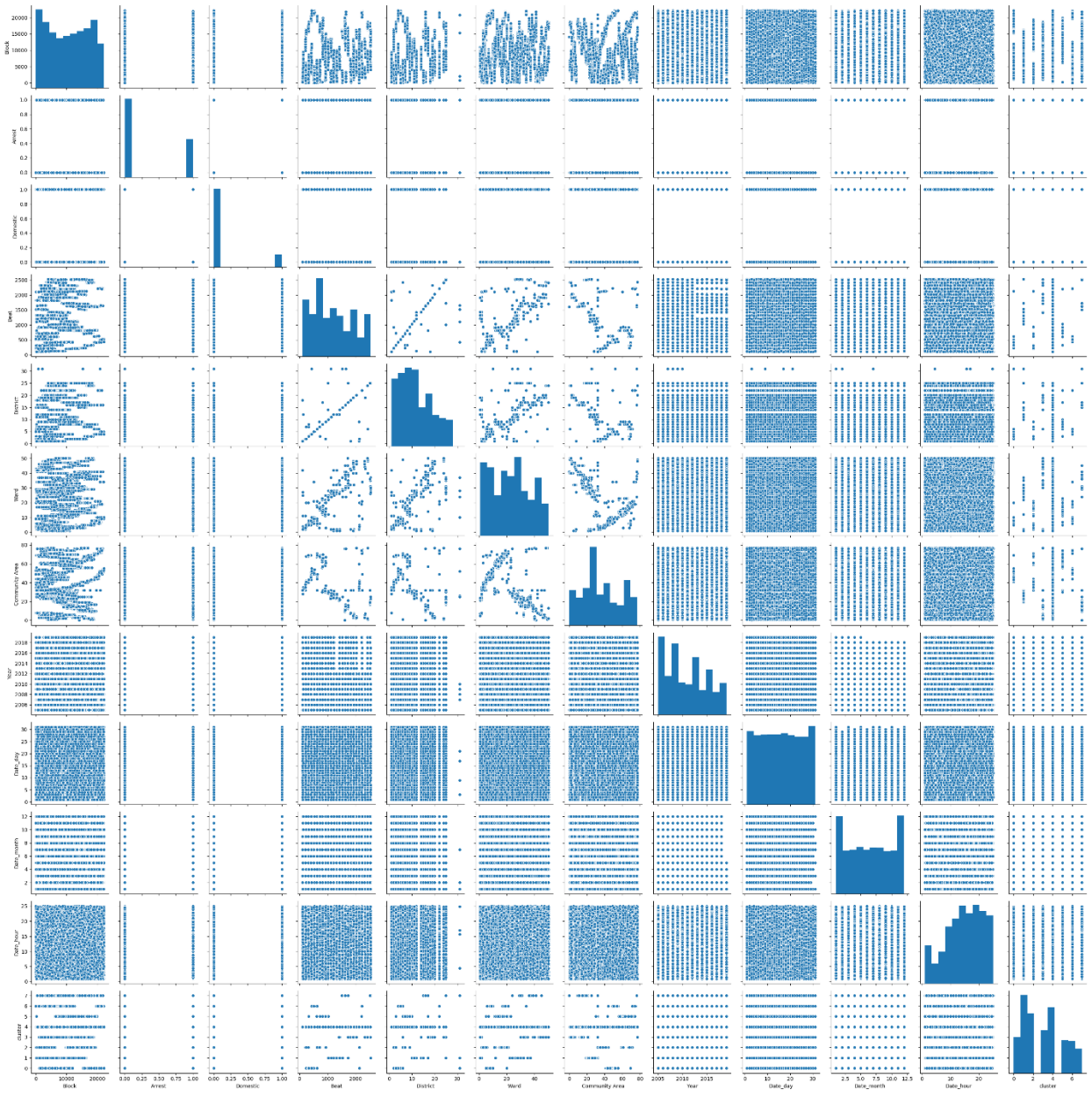
- GradientBoostingClassifier
- DecisionTreeClassifier
- Perceptron
- K-nearest-neighbours
- LogisticRegression
- RidgeRegression
- RandomForest
- BaggingRandomForest
- AdaBoost on DecisionTreeClassifier
- AdaBoost on RandomForest
- AdaBoost on Perceptron

We eventually came to a conclusion that the BaggingRandomForest was the best, it gave the most solid answers and was hard to overfit.

We noticed that all of the models can’t get better than 65% success rate and we have tried (but did not succeed) to improve that, using other models or techniques, or by manipulation of the data.

We noticed that we have columns which are irrelevant by their own but are very important together, these are “Latitude” and “Longitude”. We didn’t manage to create a new feature to take advantage of them and thus we couldn’t rely on them.




Block
Arrest
Domestic
Beat
District
Ward
Community Area
Year
Date_day
Date_month
Date_hour
cluster