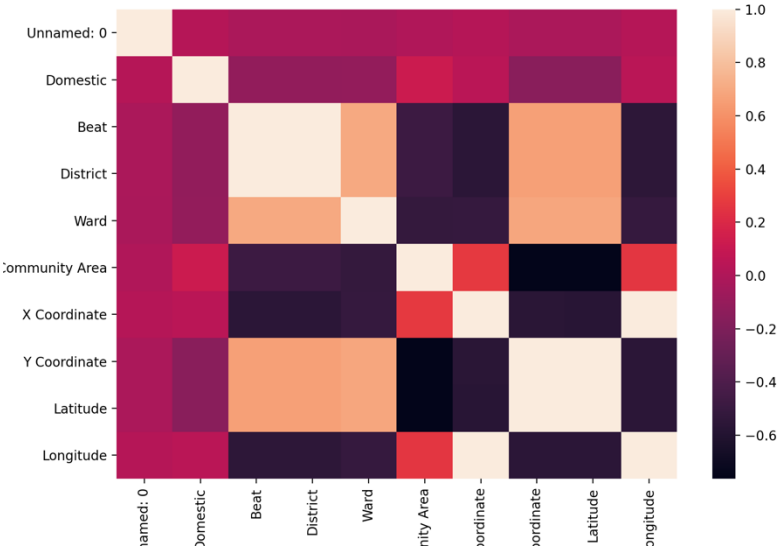
Challenges 2: Help the Chicago Police Prevent Crime!

Group name: the bootsrappers
Group members: Einat Sarig, Niv Ben-Salmon, Keren Ben-arie, Dana Adam

First assignment

STEP 1- Cleaning and Preprocessing of the data

We decided to divide the data by 70% train, 20% validation and 10% test.
First, we made some basic validation checks, making sure we are getting the proper types in each feature and that the given values are legal.
After that, we tried to determine which of the features is relevant to our analyzing process.
We realized there are many features describing the location and started off with a correlation map between them:



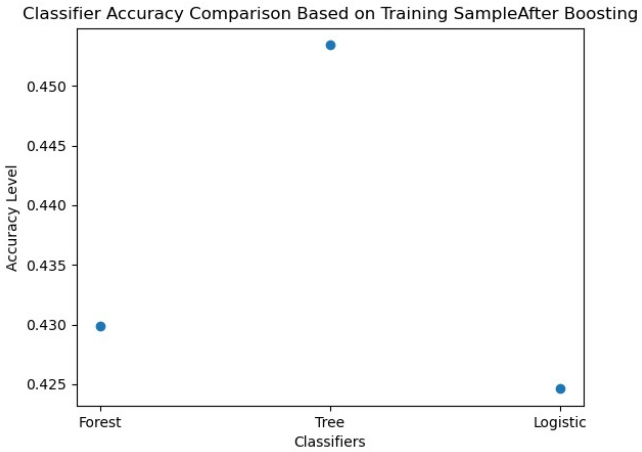
As we can see in the attached plot, there is a very high correlation (almost identical) between x coordinate and longitude, the y coordinate and latitude, beat and district. In that case, there is no need in holding all these features, that is why we decided to delete the longitude, latitude, and district features.
In addition to that, we removed the ID, Case number and Year features because those had no influence on the decision process (all crimes given had the same year).
After trying to handle both the "Block" and "Location Description", we saw that it resulted in that there was overfitting, or with a less significant handling it did not lead to any significant improvement of the model. That is the reason we decided to remove these features also.

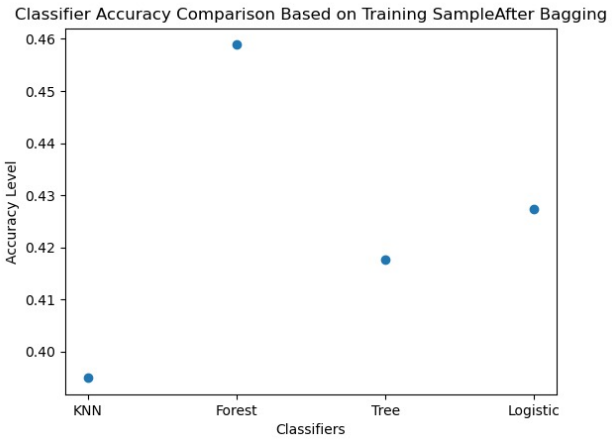
The date feature was a bit more complicated since it has date and time- 2 different features that need to be considered separately since they might have different influence on our predictions, so we split it to 2 different columns.

For the final step of pre-processing, we analyzed the free text column - Location description. We chose 20 different significant keywords and made a binary column for each one of them to identify a certain behavior based on specific sites such as schools, banks, streets etc.

STEP 2- Building models and training

We build 4 different models – KNN, random forest, Decision tree and logistic regression . For each of these models we decided which of the Meta-Algorithms Adaboost or Bagging improve the model performance.





After making that decision, we trained our model on all the data we were given, including the extra surprise data and the validation and test data.
The model we submitted is the most trained model we could have get, after using pickle .

Second assignment

In this step, we did not see a good reason to use the given dates (the argument) since we do not have any important data on the month or year and there are only 7 days a week – not a good enough indicator to determine the location. For that reason, we decided to use clustering and build a K-means model. In addition, to get the most relevant coordinates and times we calculated coordinates' distances from each cluster center and finally we chose the clusters centers which we found was most frequent. This model suits the most to our needs and gives us a pretty good evaluation of the location in case of the given dates.