IML HACKATHON

We chose the Chicago Police assignment, as we found it very interesting.

The Dataset is composed of many features, some of them were useful such as the location and the blocks, but we found that many features didn't give us much insight, such as the year, the date in which the case was updated, therefore, at the preprocess stage, we performed feature selection, keeping the features we found to the most correlated with a specific type of crime committed.

In addition, at the preprocess stage, we handled cases where certain feature cells were empty, for example, if the x coordinate or the y coordinate was empty, we calculated the mean of other coordinates in rows where the fields that are related to locations such as Ward, Block, Beat and so on that had identical values, or almost identical values.

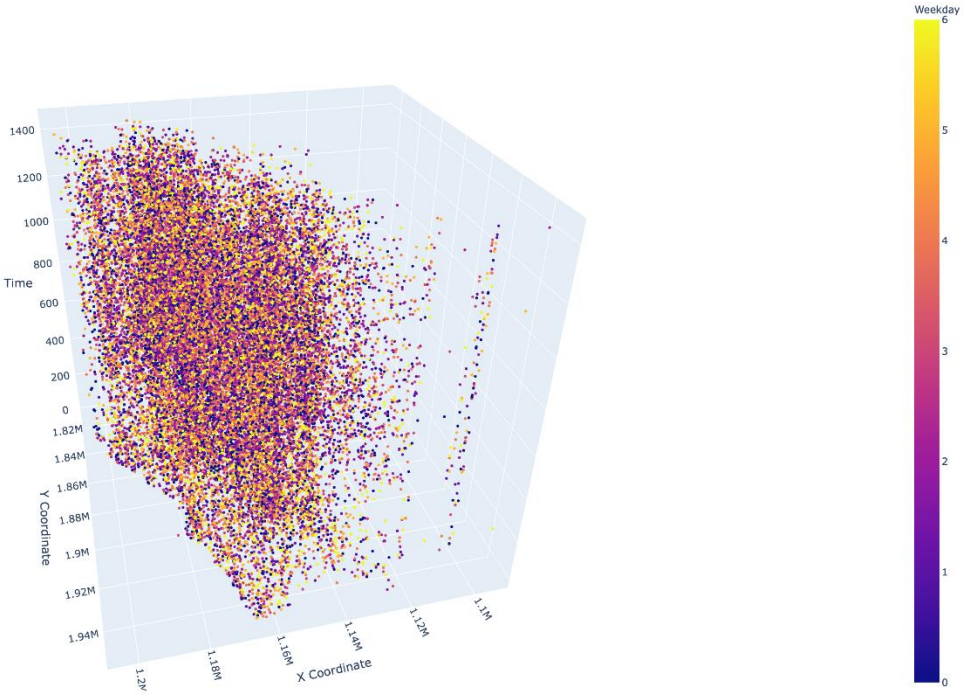
Since our problem is classification problem we used our knowledge from class to create an ensemble – a random forest of trees which decreased our miss-classification error.

As part of our assignment we tried a variety of approaches on how to handle the pre- process, at first we tried a basic pre- process, then, after experimenting in adding features we found that we could add some dummy variables that turned out to be very helpful!
In addition we tried different parameters for different classifiers to explore and know which one will get the best bias and variance – generalizing but not over fitting .

For the secondary challenge, we decided that clustering would the best method to find the crime hotspots, so we divided the crimes by weekdays, then we divided all the crimes for every weekday into 100 clusters, then we calculated the centroids for each cluster, and chose 30 centroids randomly since we want the centroids to be distributed over more options- this way we increase the chance of sending the police cars to the right areas and have as much crime prevention as possible.

When we first attempted to find the clusters, we found that they were too central in their position relative to the time axis, that was because the time scale in minutes was very small in comparison to the distance scale in the x,y plane, the largest distance margin was around 120 thousand feet while the largest time margin was 1440 minutes, so we normalized both the time and the distance margins and then calculated the centroids.

We present the distribution of crimes by x,y coordinates and the time in which they happened by the minute, with 0 being midnight, and 1439 being 23:59, the labels represent the days.



Here we present the clusters of crime, by their time and location, we plot the centroids as well, the labels represent the clusters.