Did you describe the dataset, and any challenging characteristics it has?

the data set is all crimes committed in Chicago area during 4 month at the beginning of 2021,

its most valuable categories are: date, Location Description, and X,Y coordinates for task 1,
for task 2: date, and X,Y coordinates.

Most challenging characteristics was getting a direct spot of crime on map, we used grid of X and Y coordinates for that.

Did you describe (briefly) the data cleaning and preprocessing?

First we got rid of ID, Case Number and Year which were unnecessary for us, and Longitude and Location as we used X,Y coordinates.

we extract date and time from Date column, and in order to give meaning to the date, we came up with a function to make the dates be represented as continuous numbers. We did this also for the time that was given the Date column.

We used Location Description, checked if the crime was committed outside, inside-private or

inside-public, as there was good correlation between the type of crime, and where it was committed.

We split the map into grid with X and Y coordinates, from each box we counted the number of incidents,

If multiple incidents occurred, we add those coordinates as bool, and each sample that were around would be 1, while others 0.

Did you describe the considerations that guided your design of learning systems?

We used random forest as the data was scattered on map with gathering around specific areas.

Therefore, forest could box those gatherings with ease and determine which classification is best,

We trained 3 different learners distinct by the time of time committed: morning, noon and night.

From our experience bagging was better than ada-boosting, giving more stable results.

Did you describe (briefly) various methods you tried and the results you obtained?

We firstly drew all the incidents of the crimes committed from the training set (by the x and y coordinates) on a mesh board to get a "feeling" of how the data looks. By this we saw the data is very mixed, and not separable in any way, and so we agreed that we should try the Random Forest classifier, and the Ada-boost classifier, and use whatever brings us better results. After hours of testing and running simulations, it was obvious that the Random Forest was better (with about 5%-8% better success rate). As mentioned before, in order to give more meaning to the X and Y coordinates we built a grid to divide the "map" of Chicago, used special functions for the time and date, divided the data to morning, noon and night timeframes and more .

Did you describe the learning system you ended up using?

After lots of testing, the best learning system we came with was Random Forest classifier. We also checked Ada-boost classifier, but the Random Forest classifier was better

Did you provide a prediction (and explanation) of the generalization error you expect your system to have?

From testing and plotting graphs, we expect an estimated generalization error of about 48%-50%. We have performed many tests, checking different variables , such as the max depth of the tree, committee size, amount of leaves and more, and after close inspection, we came up with the variables that give us the best results. Needless to say, we used the strategy we learned at the course of splitting the data to train, validation and test.