**Assignment - 3**

**Packages and tools used:**

For the classification, I used the following models:

* LogisticRegression (1)
* DecisionTreeClassifier (2)
* KNeighborsClassifier (3)
* LinearDiscriminantAnalysis (4)
* GaussianNB (5)
* SVC (6)

These are the classification models from the sklearn package. For plotting the visualization graphs- matplotlib was used.

**Analytical goals:**

The goal of classification was to build a model so that the NaN values in “Inside/Outside’ column of the dataset could be replaced with the original (predicted) value. This was achieved by first accurately predicting the Inside/Outside values with the data already present, testing that prediction with already separated test data and finally replacing the NaN values with the predicted values.

The features given to the model (x1, x2, x3, x4) are 'Description', 'Weapon', 'District', 'Premise'. Initially only the first 3 features were used which gave accuracies of:

62.32, 75.38, 71.81, 62.48, 54.56 and 53.2 for the 6 models (in order).

By the addition of the 'Premise' feature this accuracy rose to 97.92, 98.11, 96.23 for 1, 2 and 3.

The predicted value (y) was then used to construct the new dataset by replacing the NaN values in ‘Inside/Outside’.

I chose the DecisionTreeClassifier as it had the max accuracy among the models with 98.11%.

**Parameters & Options:**

The first choice I faced was in encoding the categorical data into Ordinal encoding/ Onehot Encoding. I chose the former which had 3 features, but on addition of the fourth (‘Premise’) which had a lot of categories, I switched to OneHot encoding. Apart from feature selection and encoding selection, I didn’t have much to pick from as I had already obtained a high accuracy using the prebuilt models from sklearn itself.

The following multi bar chart shows the number of crimes per category. We have crime categories on the x axis and crime count on the y axis. The blue bar is the inside (crime happening indoors), the green bar is outside (crime taking place outdoors). This is the crime data prior to adding the dropped data points from the classifier model.

Chart, histogram

Description automatically generated

The below visualization shows almost the same multi-bar chart but with no data points dropped. We can see that the pattern of bars remains the same even though their individual count increases. That is because the classifier model correctly classifies the crime happening indoors/outdoors. The only deviation can be seen in the Larceny category (7th bars) where the indoor/outdoor ratio changes. This can be attributed to the ‘Weapon’, ‘District’ and the ‘Premise’ features.

Chart, bar chart, histogram

Description automatically generated

**Difficulties faced:**

* The dataset is typically made up of categorical data, so all types of analysis was difficult to do as there was few to no numerical fields.
* The crime categories are several which made it difficult to get 100% accuracy, and unevenly distributed.
* Hard to plot- as there were very large categories and tiny ones- so the scale was hard to choose (which also affected the font size for the rows.)

**Problems Remaining:**

* The Nightingale Rose chart is finished for the crimes per day analysis with the crime categories displaying. Need to extend it to days/week and months/year timescales.