2. Data

The data set used in this project can be found at https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv. This data set contains the driving conditions, the number of people and vehicles involved in the crash, and the severity of the crash.

There were several problems regarding this data set. Some entries were missing crucial data required for this algorithm. For example, some columns were filled with an "Unknown" in the number of vehicles or persons injured. To remedy this, I had decided to drop the row as a whole as I believed filling in the data with the mean number of cars/people would not be an accurate representation of the car crash.

3. Methodology

After cleaning up the data, I had simply plotted three graphs. One was a simple histogram plotting the severity code against its frequency. This plot informs us that a severity code of 1 was most common among the data set.

Another plotted graph that I had generated to help view the data was the collision code and the frequency. Similarly to the one above, it plots the amount of a specific collision code. However, in contrast to the one above, the collision code is more specific, stating that the collision of code 10 or 11, which is either "entering at an angle" or "both going straight, both moving, sideswipe" was the highest amongst the rest.

The last two graphs plotted are scatter plots of damage only to property and a collision involving human injuries taking into account the number of people and vehicles involved. Looking below we see that property damage accidents mostly involve two or more vehicles and multiple people.

The last graph below (red) is the same as the one above (blue) except it graphs the collisions that involve injuries.

To predict the accident severity, I had implemented both a decision tree and K-nearest neighbor (KNN). However, the plotted decision tree had looked cluttered, and not much information could be taken from it.

Instead, the KNN implementation was much easier and helpful. I had first tested the accuracy for the value of K between 1 and 9 inclusive to see which one would result in a higher accuracy value.

4. Results

I had found that a K value of 9 resulted in the highest accuracy value at around 0.699. I then predicted the value of y_hat and it had produced 12 correct values out of 20. Using the KNN model, I had gotten around 68.5% accuracy for predicting the car accident severity.

5. Discussion

Based on the results, I believe if I had incorporated more variables to predict the target variable, the severity, the accuracy would be higher. Additionally, this project led to thinking, what if instead of predicting car accident severity after accidents had occurred, what if we had used previous car crashed and the road, weather, lighting conditions, and speeding and other data to predict what is the likelihood one would get into a car crash given those conditions.

6. Conclusion

In this study, I analyzed the relationship between the number of people injured, the number of vehicles damaged, what kind of collision, and the severity of the collision. I built classification models, specifically the KNN model, to predict the severity of a car accident.