2. Accident Dataset

The goal of this section is to indicate the sources where the data has been collected from as well as describe the meaning of each feature in the acquired dataset. It is of great importance that any data collected for the development of models intended for public use comes from reliable sources, especially when its use concerns public safety. The dataset used in this study comes from seattle.gov, the official website of the city of Seattle and provides collision data recorded from 2004 – 2020.

2.1 Data Description

There are 38 features present in the dataset and a total of 194673 records of car accidents. The first few columns provide several identification keys unique for each incident, this information is not usable for the purpose of this paper therefore it will be ignored in the analysis. The location of the accident is described by three columns, two of which refer to the latitude and longitude values while the third one contains the address. The date and time when the incident occurred is given by two columns, one containing only the date while the other containing the date and time of the collision. Next there is information on whether the accident occurred at an intersection, a block or an alley, the type of the junction and the unique keys of each intersection, crosswalk road segment where the accident occurred. There are several features showing variables such as the weather, the road condition, the lighting condition, whether the driver was inattentive, speeding was under the influence of alcohol or the pedestrian right of way was not granted. The dataset also provides information regarding the number of pedestrians, cyclists and vehicles involved in the accident and whether there was a collision with parked cars. There is information regarding the State Collision Code which uniquely describes each type of collision using a numeric value. Finally, the severity of each accident is given as a binary variable taking the value of 1 if the accident was less serious and the value of 2 if the accident was severe, this column will be used as the target data for the model to predict.

2.2 Feature Selection

The goal of this study is the development of a model that predicts the severity of an accident using real-time data, this puts some limits on the type of features that can be used since not all of the dataset's columns contain information that can be obtained in real-time. Locational variables such as intersections codes, junction types etc. as well as information regarding weather, road condition and speed limits can be acquired through the use of GPS, news and speedometers. Features containing information regarding the type of the collision cannot be used since there is no way to know this before the accident occurs. Furthermore, data on whether the driver is inattentive or under the influence of alcohol would be unknown to a machine such as electronic road signs or GPS devices therefore they should be excluded from the set of usable features.

There are more variables for which the methodology of obtaining real-time data is ambiguous. These include the number of vehicles, pedestrians, cyclists and whether parked cars were involved in an accident. These features can be converted into binary, this would mean that they will provide information on whether or not they were involved in an accident. By analyzing the frequency of vehicles, pedestrians and cyclists passing from certain locations as well as the number of parked cars in that location, signals of these frequencies passing certain limits can be generated.

2.3 Data Preparation

The process of data preparation includes removing or filling missing values, converting features into certain formats and making sure the data is balanced. This step is required in most algorithms in order for them to be implemented correctly and can significantly boost their accuracy. The dataset contains features which vary on their data types as well as a lot of missing values. Most of the features contained either discrete numeric or categorical values, the later were converted into discrete numeric values so that they can be used in the algorithm. Certain features contained only one unique value and a large number of missing values, for cases such as these missing values were converted into their own category. Such an example would be the column showing whether a car involved in an accident was speeding or not, the only unique value of this column was ‘Y’ which most probably ment ‘YES’, when a row wasn’t equal to ‘Y’ the value was missing, this indicated that the empty rows should take the value ‘N’ which would stand for ‘NO’ so that the feature would become binary. Some features contained a large number of unique values many of which had only a small number of appearances in the dataset, these values were merged with larger categories of similar characteristics. There were three variables containing information on the exact location of the accident, from these features only the ones containing the latitude and longitude were used while the third one was completely dropped. In order to convert the (lat, lon) coordinates into discrete numeric values the k-means clustering algorithm was used on the two features (Y, X) generating six distinct location clusters. The remaining rows containing missing values were removed.

There were two columns with information regarding the date and time of the accidents. One of them only contained data on the date of the accident while the other also contained data on the exact time it occurred. In order to analyze date and time in more detail, four new rows containing the month, the day of the month, the day of the week and the hour of the accident were generated while the initial columns were dropped from the dataset.

When looking at the frequency of the severity of the accidents it is obvious that the dataset is unbalanced in favor of the less severe ones and therefore biased. After the preprocessing step was done, about 67% of the remaining data had a severity of 1 while only 33% had a severity of 2. In order to tackle this problem a technique called down-sampling was used, a random sample without replacement of size equal to the number of rows containing a severity value of 2 was taken from the rows containing the value 1, the end result was a dataset with an equal amount of the two distinct values of the target data. The final dataset contains 20 columns one of which is the target data and 109380 accident records which is large enough for training and testing the model.