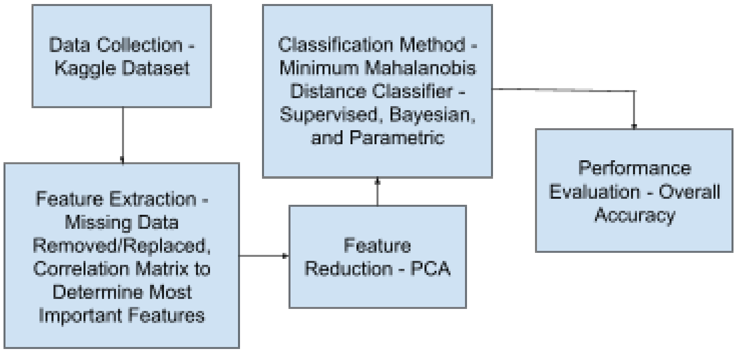
U.S. Car Accidents (2016-2023) - Group 9 - Milestone 3

By: Clay Shubert, Nagendra Upadhyay, Robert Grady Williams, Zach Williams

Part 1: Pipeline


Data Collection - Kaggle Dataset
Feature Extraction - Missing Data Removed/Replaced, Correlation Matrix to Determine Most Important Features
Feature Reduction - PCA
Classification Method - Minimum Mahalanobis Distance Classifier - Supervised, Bayesian, and Parametric
Performance Evaluation - Overall Accuracy

Part 2: Experimental Design

We started with the raw data from Kaggle and determined the data types present in the dataset using a pandas dataframe. We noticed that there were two datetime objects in 'Start_Time' and 'End_Time', so we extracted the fields 'Year', 'Month', 'Day', 'Hour', and 'Weekday' from 'Start_Time' and gave them their own column entries. In addition, we expanded the data to include a new column 'Time_Duration' by subtracting the 'Start_Time' from the 'End_Time'. 'Start_Time' and 'End_Time' were then removed because they are covered by the added columns and 'Time_Duration'. After this we inspected the newly created 'Time_Duration' field for any negative values or outlying values and replaced them with the median 'Time_Duration' value. We then label encode any string object fields to be integers for future steps.

Following that, we used a correlation matrix based on the 'Severity' value to determine the most important features and dropped unimportant features from the data. We removed multiple columns such as 'Civil_Twilight', 'End_Lat', 'End_Lng', 'Country', and 'Airport_Code' because they had very little positive or negative correlation to the accident severity, were overlapping data already contained in other columns, or had only one value among all samples.

Finally, we created training, testing and validation sets by splitting the data to have 20% testing entries and of that 20%, half was made the validation set. So in total 80% training data, 10% testing data, and 10% validation data. After this we applied PCA dimensionality reduction and reduced the dataset to 20 features, down from 36, but intend to test with other values.

Part 3: Results

For this milestone we reported the accuracy of the minimum Mahalanobis distance classifier on the PCA reduced data. The overall accuracy was 38.34% with 10,000 training samples. Because of the size of the full dataset, we decided to go with a much lower number of samples to train the model for this milestone. As a result, the accuracy rate is not as good as we expect compared to the state of the art example. Despite this, it is still currently better than a random guess because there are four potential classes. We have decided to keep working on this issue by trying different feature reduction amounts and different classifier algorithms. We hope to be able to achieve a similar performance to the state of the art notebook, so our goal is around 90%. Our current task allocation is that all group members contribute to every portion of the project such as pre-processing, processing, and post-processing.