DS 3000 Group 34 Phase 3: Model selection, Training and Evaluation

Teammates:

Isabella Hernandez, Han-Mac Kim, Julia Ouritskaya, Henry Renninger

Title:

Exploring Factors Contributing to Vehicle Collisions in New York City

1. Algorithms to Explore:

Decision Tree:
Decision Trees are intuitive and capable of capturing non-linear relationships, making them suitable for analyzing the complex factors involved in vehicle collisions. With decision trees, we will be able to use classification and regression techniques. While they excel in handling both categorical and numerical data, their tendency to overfit makes the algorithm sensitive to variations in the data and requires careful handling.

Random Forest:
An extension of decision trees, Random Forests also are used for classification and regression tasks and aggregate the predictions of numerous trees, enhancing accuracy against overfitting. It is a much slower algorithm, but since this project isn't running in real-time, this is superfluous. This algorithm is proficient in handling large datasets with high-dimensional feature space, making it ideal for deriving insights from varied factors contributing to vehicle collisions.

k-Nearest Neighbors (KNN):
KNN is a straight-forward algorithm with no training phase that classifies data based on its similarity to neighboring points. It is particularly effective for small to medium-sized datasets and requires careful selection of the number of neighbors (k) and distance metric. However, its performance may decline with an increasing number of features and larger datasets.

2. Training and Evaluation

The dataset will be partitioned into training and testing subsets (e.g., 70% training, 30% testing). We will use cross-validation (like k-fold) to ensure the model's performance is consistent across different subsets of the data. The training process will also include preprocessing steps like encoding categorical data and normalizing/scaling numerical data. For evaluation, we will use the accuracy score and metrics like Accuracy, Precision, Recall, and the F1 score since the project deals with collision frequencies and factors.

3. Hyper-parameter tuning

- Decision Tree: Key parameters include `*max_depth*` (to control tree depth), `*min_samples_split*` , and `*min_samples_leaf*` (to avoid overfitting).

- Random Forest: Similar to Decision Trees, but with additional parameters like `*n_estimators*` (to determine the number of trees in the forest) and `*max_features*` (to specify the number of features to consider for the best split).
- K-Nearest Neighbors: Key parameters include `*n-neighbors*` (number of neighbors) and `*distance metric*` (like Euclidian).

4. Expected Measure of Accuracy

For the classification models, we will primarily focus on the Accuracy metric and the Classification Report, which includes Precision, Recall, and F1 Score, trying to classify so we can see how accurate the prediction is. For regression, we would use Mean Squared Error (MSE), trying to see how far our predicted value is from regression.