ML 1- Yilmaz

2024-2025

Name of the group members (max 2 people): Aaryan Sumesh, Sami Saleh

If you are working alone, please state it clearly.

If you are working in groups, only one group member should submit the proposal.

Must haves:

* Link to dataset
* <https://catalog.data.gov/dataset/motor-vehicle-collisions-crashes>
* Information about your data
  + Meaning of attributes
  + Dimension
  + Number of instances
  + How many missing values
  + Is it uniform? Skewed?
  + What are the class distributions?
* What are you classifying/predicting?
* How will this be useful?
* Plans for preprocessing

Don’t just insert your answers in the above, make it look like you're writing the portions of the final report. See the final report sample I posted at Schoology.

As soon as you are ready with your proposal, please see me, so that I can approve it.

Also, please add your project to the spreadsheet I will provide under Q1 Project Folder.

Link to dataset - <https://catalog.data.gov/dataset/motor-vehicle-collisions-crashes>

Meaning of attributes

* CRASH DATE: Date of the car accident in format day/month/year
* CRASH TIME: Time of car accident in format hour:min on 24 hour clock
* BOROUGH: If the car accident occurred in a New York borough, that location is given from 5 choices: Brooklyn, Bronx, Manhattan, Queens, Staten Island
* ZIP CODE: 5 digit area code where car accident occurred
* LATITUDE: Latitude coordinate of car accident location
* LONGITUDE: Longitude coordinate of car accident location
* LOCATION: Location of incident in form (Latitude, Longitude)
* ON STREET NAME: Name of street where car accident occurred
* CROSS STREET NAME: Name of street that crosses the on street at the accident
* OFF STREET NAME: Building address outside where accident occurred
* NUMBER OF PERSONS INJURED: Total persons injured in car accident
* NUMBER OF PERSONS KILLED: Total persons killed in car accident
* NUMBER OF PEDESTRIANS INJURED: Total pedestrians (people walking on street/sidewalk) injured in car accident
* NUMBER OF PEDESTRIANS KILLED: Total pedestrians (people walking on street/sidewalk) killed in car accident
* NUMBER OF CYCLIST INJURED: Total amount of people injured who were riding bikes at the time of the car accident
* NUMBER OF CYCLIST KILLED: Total amount of people killed who were riding bikes at the time of the car accident
* NUMBER OF MOTORIST INJURED: Total amount of people injured who were in or driving cars at the time of the car accident
* NUMBER OF MOTORIST KILLED: Total amount of people killed who were in or driving cars at the time of the car accident
* CONTRIBUTING FACTOR VEHICLE 1: A label of a possible reason that the first driver was involved in the car accident
* CONTRIBUTING FACTOR VEHICLE 2: A label of a possible reason that the second driver was involved in the car accident (if there were 2 cars involved in the accident)
* CONTRIBUTING FACTOR VEHICLE 3: A label of a possible reason that the third driver was involved in the car accident (if there were 3 cars involved in the accident)
* CONTRIBUTING FACTOR VEHICLE 4: A label of a possible reason that the fourth driver was involved in the car accident (if there were 4 cars involved in the car accident)
* CONTRIBUTING FACTOR VEHICLE 5: A label of a possible reason that the fifth driver was involved in the car accident (if there were 5 cars involved in the car accident)
* COLLISION\_ID: Unique ID label for each instance of an accident
* VEHICLE TYPE CODE 1: A label for the form of transportation of the first vehicle
* VEHICLE TYPE CODE 2: A label for the form of transportation of the second vehicle
* VEHICLE TYPE CODE 3: A label for the form of transportation of the third vehicle
* VEHICLE TYPE CODE 4: A label for the form of transportation of the fourth vehicle
* VEHICLE TYPE CODE 5: A label for the form of transportation of the fifth vehicle

Dimension

* The dimension of the data set is 28 (28 attributes)

Number of instances

* 1048575

How many missing values

* BOROUGH: 376652
* ZIP CODE: 376836
* LATITUDE: 75283
* LONGITUDE: 75283
* LOCATION: 75283
* ON STREET NAME: 257041
* CROSS STREET NAME: 543032
* OFF STREET NAME: 792862
* NUMBER OF PERSONS INJURED: 17
* NUMBER OF PERSONS KILLED: 30
* CONTRIBUTING FACTOR VEHICLE 1: 3757
* CONTRIBUTING FACTOR VEHICLE 2: 178615
* CONTRIBUTING FACTOR VEHICLE 3: 970983
* CONTRIBUTING FACTOR VEHICLE 4: 1030528
* CONTRIBUTING FACTOR VEHICLE 5: 1043535
* COLLISION\_ID: 1048575
* VEHICLE TYPE CODE 1: 8659
* VEHICLE TYPE CODE 2: 249127
* VEHICLE TYPE CODE 3: 975403
* VEHICLE TYPE CODE 4: 1031430
* VEHICLE TYPE CODE 5: 1043737

Is it uniform? Skewed?

* NUMBER OF PERSONS INJURED and NUMBER OF PERSONS KILLED are both heavily skewed to the right

What are you classifying/predicting?

* We will be predicting the severity of car crashes. We will group car crashes into three categories based on how lethal the crashes are. The three class labels will be NON-LETHAL, SOMEWHAT-LETHAL, and VERY-LETHAL.

How will this be useful?

* People can use our model and using the information that the model uses, they can predict whether or not the location is prone to car accidents. People can use this in urban planning, where they can see where to invest money in improving both pedestrian and driver safety.

Plans for preprocessing

1. Fix the missing values in all of the above attributes. Also, we have to use binning to create the 3 class labels that we are going to predict. We need to decide what method to use (equal width, depth …).