# **IBM Data Science Capstone:**

# **Car Accident Severity Report**

**Introduction**  
To reduce the frequency of car collisions in a community, an algorithm must be developed to predict the severity of an accident given the current weather, road, and visibility conditions. When conditions are bad, this model will alert drivers to remind them to be more careful.

The target audiences of this study are those people who really care about the traffic records, especially in the transportation department. Also, we want to figure out the reason for collisions and help to reduce accidents in the future.

**Business problem**  
The purpose of this project is to analyze the collision dataset for the city Seattle and find patterns, also determine key factors such as weather, light and road conditions, drug or alcohol influence, to provide the best traffic accident severity prediction.

**Data Understanding**  
We will use SEVERITYCODE as our dependent variable Y, and try different combinations of independent variables X to get the result.

**Data:**

We want to analyze the accident severity in terms of human fatality, traffic delay, property damage, or any other type of accident bad impact.The data was collected by Seattle SPOT Traffic Management Division and provided by Coursera via a link. It contains information such as severity code,address type, location, collision type, weather, road condition, speeding, among others.

There are 194,673 observations and 37 variables in this data set. Since we would like to identify the factors that cause the accident and the level of severity,we will use SEVERITYCODE as our dependent variable Y, and try different combinations of independent variables X to get the result. Since the observations are quite large, we may need to filter out the missing value and delete the unrelated columns first. Then we can select the factor which may have more impact on the accidents, such as address type, weather, road condition, and light condition.

Other important variables include:

• ADDRTYPE: Collision address type: Alley, Block, Intersection

• LOCATION: Description of the general location of the collision

• PERSONCOUNT: The total number of people involved in the collision helps identify severity involved

• PEDCOUNT: The number of pedestrians involved in the collision helps identify severity involved

• PEDCYLCOUNT: The number of bicycles involved in the collision helps identify severity involved

• VEHCOUNT: The number of vehicles involved in the collision identify severity involved

• JUNCTIONTYPE: Category of junction at which collision took place helps identify where most collisions occur

• WEATHER: A description of the weather conditions during the time of the collision

• ROADCOND: The condition of the road during the collision

• LIGHTCOND: The light conditions during the collision

• SPEEDING: Whether or not speeding was a factor in the collision (Y/N)

• SEGLANEKEY: A key for the lane segment in which the collision occurred

• CROSSWALKKEY: A key for the crosswalk at which the collision occurred

• HITPARKEDCAR: Whether or not the collision involved hitting a parked car

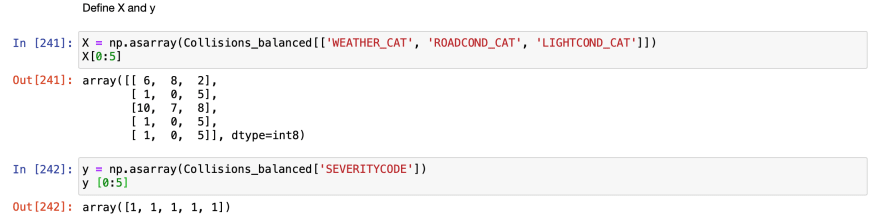
**Methodology**

The data is now ready to be modeled

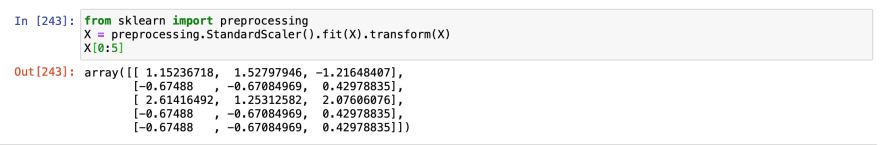
**K-Nearest Neighbor (KNN)**  
KNN will help us predict the severity code of an outcome by finding the most similar to data point within k distance.

**Decision Tree**  
A decision tree model gives us a layout of all possible outcomes so we can fully analyze the consequences of a decision. Its context, the decision tree observes all possible outcomes of different weather conditions.

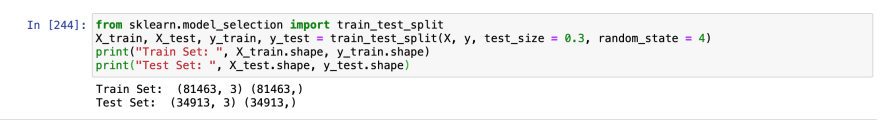
**Logistic Regression**  
Because our dataset only provides us with two severity code outcomes, our model will only predict one of those two classes. This makes our data binary, which is perfect to use with logistic regression.

Initialization:  
[](https://res.cloudinary.com/practicaldev/image/fetch/s--1zqfBjFH--/c_limit%2Cf_auto%2Cfl_progressive%2Cq_auto%2Cw_880/https:/dev-to-uploads.s3.amazonaws.com/i/igg8rmgjkd039fn3t8l2.png)

Normalizing the data:

[](https://res.cloudinary.com/practicaldev/image/fetch/s--JnYVj-dY--/c_limit%2Cf_auto%2Cfl_progressive%2Cq_auto%2Cw_880/https:/dev-to-uploads.s3.amazonaws.com/i/f70j422wk5vqsmd8e5wr.png)

Train/Test Split:

[](https://res.cloudinary.com/practicaldev/image/fetch/s--IR_8k18D--/c_limit%2Cf_auto%2Cfl_progressive%2Cq_auto%2Cw_880/https:/dev-to-uploads.s3.amazonaws.com/i/sxtmz8v8ju9xus0rohb4.png)

Then we modeled the data using 3 different machine learning algorithms: KNN, Decision tree, and Logistic regression.

**Results & Evaluation:**Then we checked the accuracy and found out that the Decision tree is the most accurate model.

**Discussion:**  
Most crashes happened in clear, dry, and bright conditions. Most days are clear, dry, and bright, so it’s no surprise that most car crashes occur under these conditions. I also found out that crashes with a distracted driver or an impaired driver are statistically more likely to result in injury, which is also not a surprise. The results of the data indicate to city officials that they should ask drivers to be more alert in ideal conditions.

**Conclusion:**  
Based on historical data from weather conditions pointing to certain classes, we can conclude that particular weather conditions have a somewhat impact on whether or not travel could result in property damage (class 1) or injury (class 2).