# Predicting Car Accident Severity

# Car Accident Severity

- Car accidents can post economical and social harm to the society.
- Accidents can be caused by different factors:
    i. Weather
    ii. Road conditions
    iii. Light conditions
    iv. Not paying enough attention during driving
    v. Speeding
    vi. Under influence of drugs or alcohol.

- Predict the severity of an accident given some factors can help to analyze the car accidents, also it can help to give some suggestions to the target audiences about how to reduce the number of accidents.
- Target audiences:
    i. Local Seattle government
    ii. Police
    iii. Car drivers

# Data Acquisition and Cleaning

- Data was collected by the Seattle Police Department and Accident Traffic Records Department from 2004 to 2020.
- Recorded the car accidents which have taken place in the city of Seattle, Washington from the year 2004 to 2020.
- In total, 194673 rows and 37 independent variables in the raw dataset.
- Target variables is “SEVERITYCODE”, and relevant independent variables are “INATTENTIONIND”, “UNDERINFL”, “ROADCOND”, “LIGHTCOND”, “WEATHRE”, and “SPEEDING”.
- Other irrelevant columns were dropped.

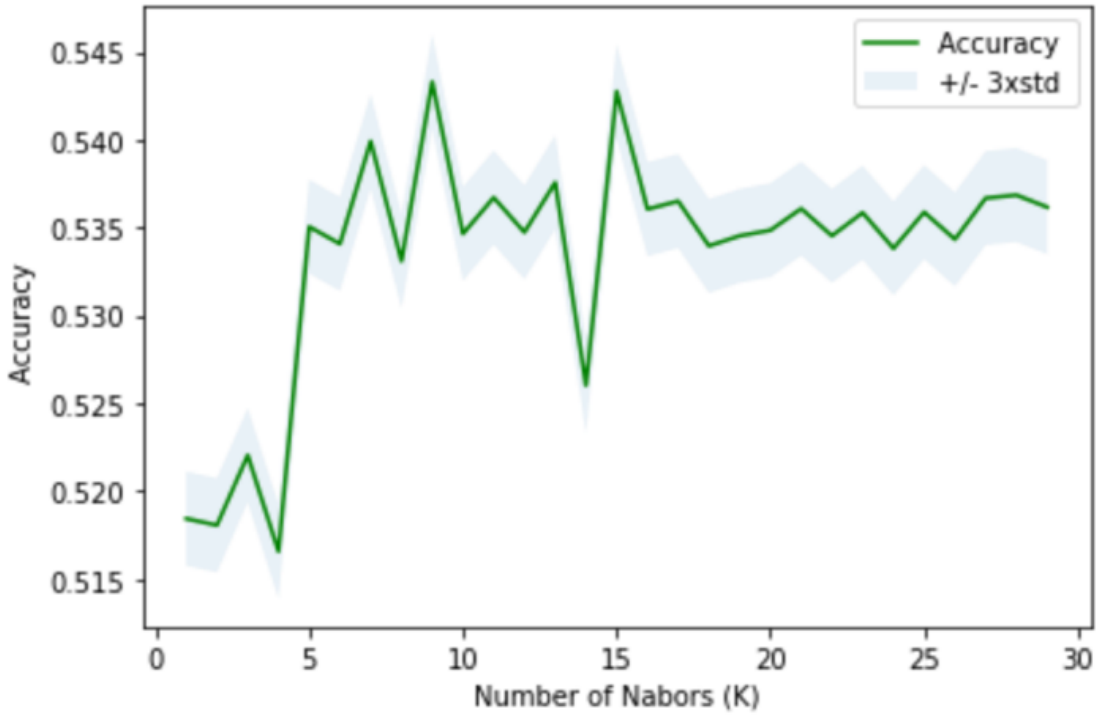
- Target variable was imbalance and it is balanced by using the technique of downsampling.
- Features are of object data types and they were converted into numerical data types.
- Missing values were replaced by frequency.

# Methodology

- Data was split into train set and test set:
    i. 30% was used for testing
    ii. 70% was used for training
- Three machine learning models were used:
    i. Logistic regression
    ii. Decision Tree (max depth = 4, criterion = “entropy” )
    iii. K-Nearest Neighbor
- Jaccard score, F1-score and Log loss were calculated for evaluating the models.

# K-Nearest Neighbor

- To decide the value for k to be used for building kNN model, an iterative process was done to calculate the accuracy of prediction using different k values.




Best accuracy = 0.5433 with k = 9

# Result

| | Algorithm | Jacard | F1-score | LogLoss |
|---|---|---|---|---|
| **0** | KNN | 0.415206 | 0.538047 | NA |
| **1** | Decision Tree | 0.123197 | 0.454320 | NA |
| **2** | Logistic Regression | 0.288168 | 0.539320 | 0.669559 |

# Conclusion

- Models were built for predicting the severity of car accident given some factors.
- Accuracy of the models still can be improved:
    i. if there were fewer missing values in the dataset
    ii. if it is a balance dataset for target variable
    iii. changing the maximum depth of the tree
- Based on the dataset and the machine learning models analysis, factors like weather conditions, road conditions, light conditions, not paying enough attention during driving, driving at high speed, and under influence of drugs or alcohol have some impact on the severity of the accidents.