# CAPSTONE PROJECT REPORT(Vehicle Collision)

## 1. Introduction

### 1.1 Background

These days vehicle accidents are getting very serious issue globally leading to deaths and serious injuries.Here we are concerned abour car collisions and will try to understand the factors on which collisions happens or occurs.

### 1.2 Problem

Problem occurs with business understanding of the above mentioned issue. We need some solution to reduce or eventually stop this collisions. In an effort to reduce the frequency of car collisions in a community, an algorithim 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.

### 1.3 Intrest

Road Transport Corporation, Drivers & Highway Authorities.

## 2. Data

### 2.1 Data Sources:

Data was given by Coursera only. Used that only.

### 2.2 Data Understanding:

Our predictor or target variable will be 'SEVERITYCODE' because it is used to measure the severity of an accident from 0 to 5 within the dataset. Attributes used to weigh the severity of an accident are 'WEATHER', 'ROADCOND' and 'LIGHTCOND'.

Severity codes are as follows:

0 : Little to no Probability (Clear Conditions)

1 : Very Low Probablility - Chance or Property Damage

In it's original form, this data is not fit for analysis. For one, there are many columns that we will not use for this model. Also, most of the features are of type object, where they should be numerical type.

We must use label encoding to covert the features to our desired data type. With the new columns, we can now use this data in our analysis and ML models.

## 3. Methodology

We have done some EDA by looking into the data. Did Standard scaling for normalizing the data set.

Label Encoding for converting object data into integer data. Removed unnecessary columns.

Our data is now ready to be fed into machine learning models.

We will use the following models:

### 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 concequences of a decision. It 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.

## 4. Result & Evaluation

We have evaluate our models by following metrics:

1. For Decsion Tree(Accuracy Score,F1 Score, Jaccardian Score)
2. For KNN (Accuracy Score,F1 Score, Jaccardian Score)
3. For Logistic Regression(Accuracy Score,F1 Score, Jaccardian Score & Log Loss)

Sinnce our target variable is binary, so Logistic Regression found to be the bpest alogorithm for this problem.

## 5. Discussion:

In the beginning of this notebook, we had categorical data that was of type 'object'. This is not a data type that we could have fed through an algoritim, so label encoding was used to created new classes that were of type int8; a numerical data type.

Once we analyzed and cleaned the data, it was then fed through three ML models; K-Nearest Neighbor, Decision Tree and Logistic Regression. Although the first two are ideal for this project, logistic regression made most sense because of its binary nature.

Evaluation metrics used to test the accuracy of our models were jaccard index, f-1 score and logloss for logistic regression. Choosing different k, max depth and hyparameter C values helped to improve our accuracy to be the best possible.

## 6. Conclusion:

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