**Applied Data Science Capstone Project**

**Prediction of Car Accidents Severity**

**Introduction/Business Problem**

The primary audience for this data science project is the unit responsible for safety analysis within Seattle Department of Transportation. They will be interested in achieving the highest levels of safety performance for the general public through monitoring and controlling the rate of accidents in Seattle. For this to be achieved, data analysis has to applied for the data set in hand, a predictive model for accidents severity to be developed. Through this model, factors leading to the most severe accidents, as a priority, can be identified and controlled through targeted mitigation actions at the root causes.

**Data**

The data that will be used for this project will be provided from Seattle Department of Transport, traffic management division. The data is a historical data set from 2004 to present on collisions, with the size of 194673 observations (rows) and 38 attributes (columns).

Among the 38 attributes within the data set is information about accidents severity (SEVERITYCODE attribute) which our model will predict with the highest possible level of accuracy as a function of the selected features from the data set.

A link providing description to the attributes of data set can be found [here](https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf).

**Methodology**

Data shape, types, unique values and columns were explored, it was discovered that the attributes of interest are categorical variables where a new data frame was created dropping unnecessary attributes and retaining the following 8 attributes;

[['SEVERITYCODE', 'JUNCTIONTYPE', 'INATTENTIONIND', 'UNDERINFL', 'WEATHER','ROADCOND', 'LIGHTCOND', 'PEDROWNOTGRNT', 'SPEEDING']]

Missing data was explored and it appeared that some columns were missing data;

JUNCTIONTYPE 6329

INATTENTIONIND 164868

UNDERINFL 4884

WEATHER 5081

Each column was handled individually before applying one hot encoding to convert 7 categorical variables into binary variables so they can be used in a Feature data frame for model building to predict SEVERITYCODE. No missing data was dropped.

In Model Building, we used the feature data set to build an accurate model based on classification algorithms since data was labelled, split the data to train and test the model, then reported the accuracy of each model to choose the best one. We used the following 4 machine learning classification algorithms:

* K Nearest Neighbor (KNN)
* Decision Tree (DT)
* Support Vector Machine (SVM)
* Logistic Regression (LR)

**Results**

Each of the 7 attributes were explored Four classification algorithms were developed using 7 variables to predict the target variable; “SEVERITYCODE”. Model accuracy was tested through Jaccard index and F1-score. KNN provided the lowest accuracy relative to the other three models, which in their turn provided almost equal accuracy levels.

The following table summarizes the 4 models accuracy

|  | **Jaccard index** | **F1-score** |
| --- | --- | --- |
| **KNN** | 0.64 | 0.64 |
| **DT** | 0.71 | 0.61 |
| **SVM** | 0.71 | 0.61 |
| **LR** | 0.71 | 0.62 |

**Discussion and Recommendations**

Accidents severity data was explored and several models were developed using 7 attributes to predict severity levels. Since the different models accuracy did not exceed 71% in terms of jaccard index,it is recommended that models can be further developed using different attributes, especially those attributes that are related to location, in order to enhance the machine learning model predictability.

**Conclusion**