IBM Applied Data Science Capstone Project
Car Accident Severity

Importance for predicting road accident Severity

- Approx. 1.35 million people die in road crashes each year
- 20-50 million suffer non-fatal injuries, often resulting in long-term disabilities
- Big problem in the way of improving public health
- Take a toll on the countries’ finance
- Multidisciplinary approach is required to understand the causes
- Great help to city planners and emergency service providers

Objectives

1. To find main features causing road accidents and factors to predict severity of road accidents

2. Train and test machine learning models for Seattle road accident dataset and find best model to predict severity of road accidents

Data Source

- Data which is recorded by Seattle Department of Transport (SDOT) is used for training and testing machine learning models for predicting severity of accident.
- Data is downloaded from link provided by IBM in applied data science capstone course
- Dataset contains 194673 rows and 38 columns . The metadata of datasetcan be downloaded from

https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf

Data Cleaning / Data Wrangling

It is process of converting data from the initial format to format that may be better for analysis. It includes following steps

1. Identifying missing data of key features and dropping the affected rows

2. Drop unnecessary columns which are unrelated to causes

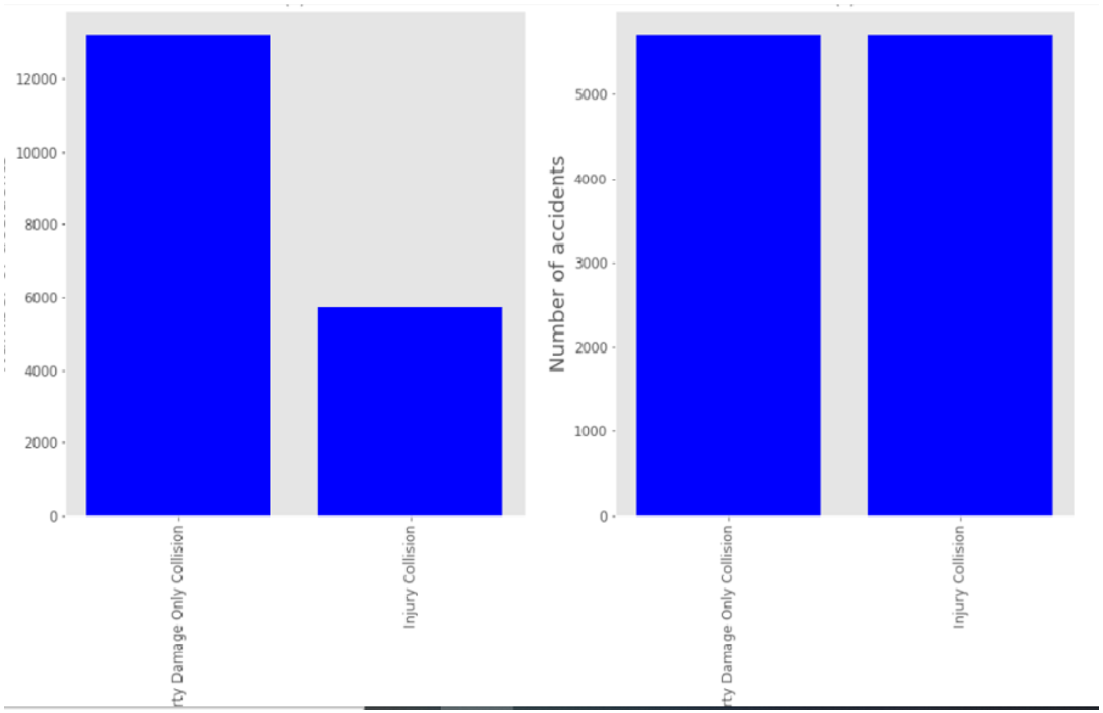
Or severity of accidents

3. convert categorical data to numerical data via one hot encoding as machine learning models do not handle categorical variables

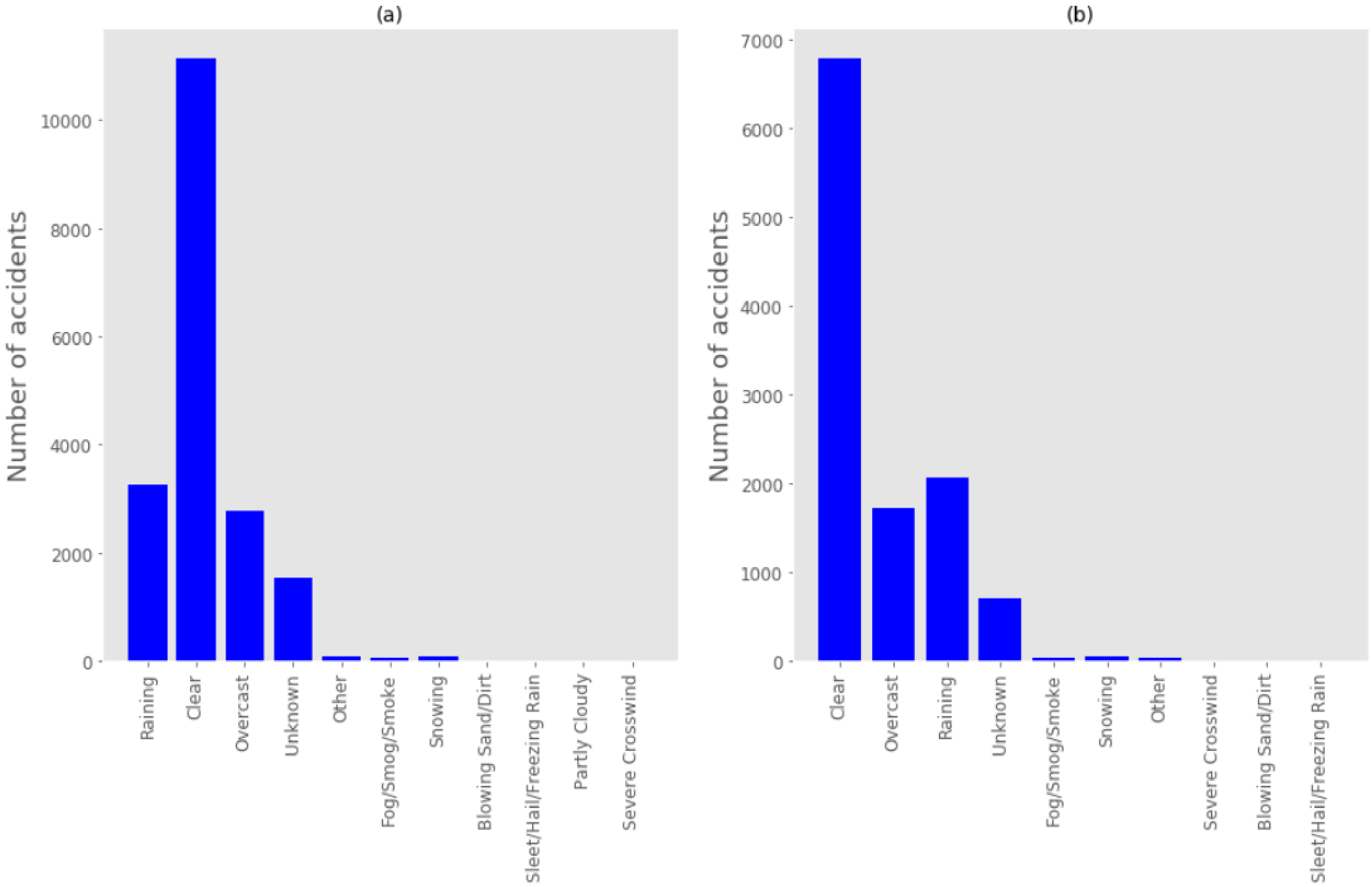
Cont...

Data Cleaning / Data Wrangling

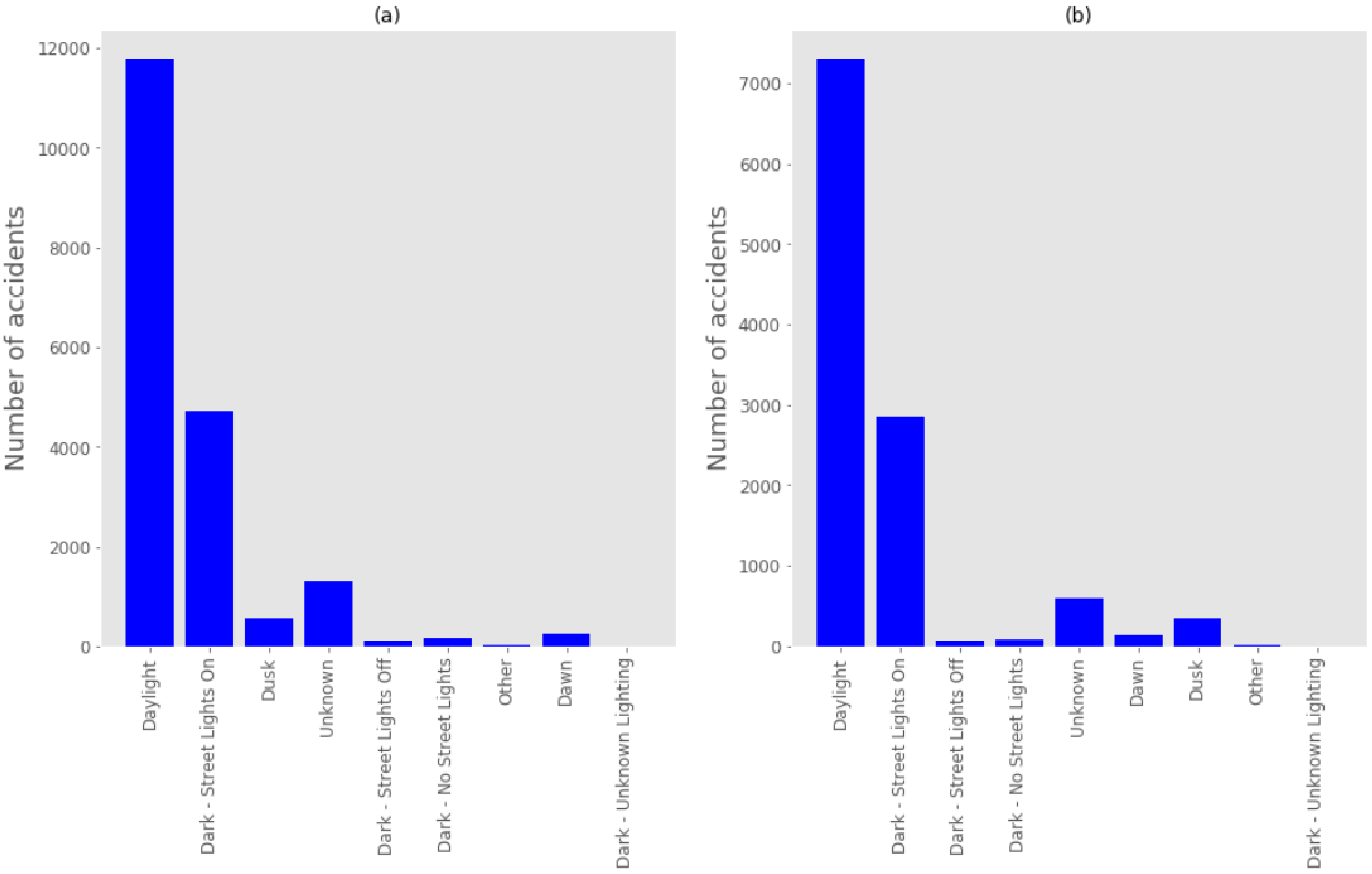
4. Dealing with unbalanced dataset – The dataset is highly imbalanced having 13221 accidents with SEVERITYCODE=1 and 5705 accidents of SEVERITYCODE=2. Data is balanced to avoid biased model.



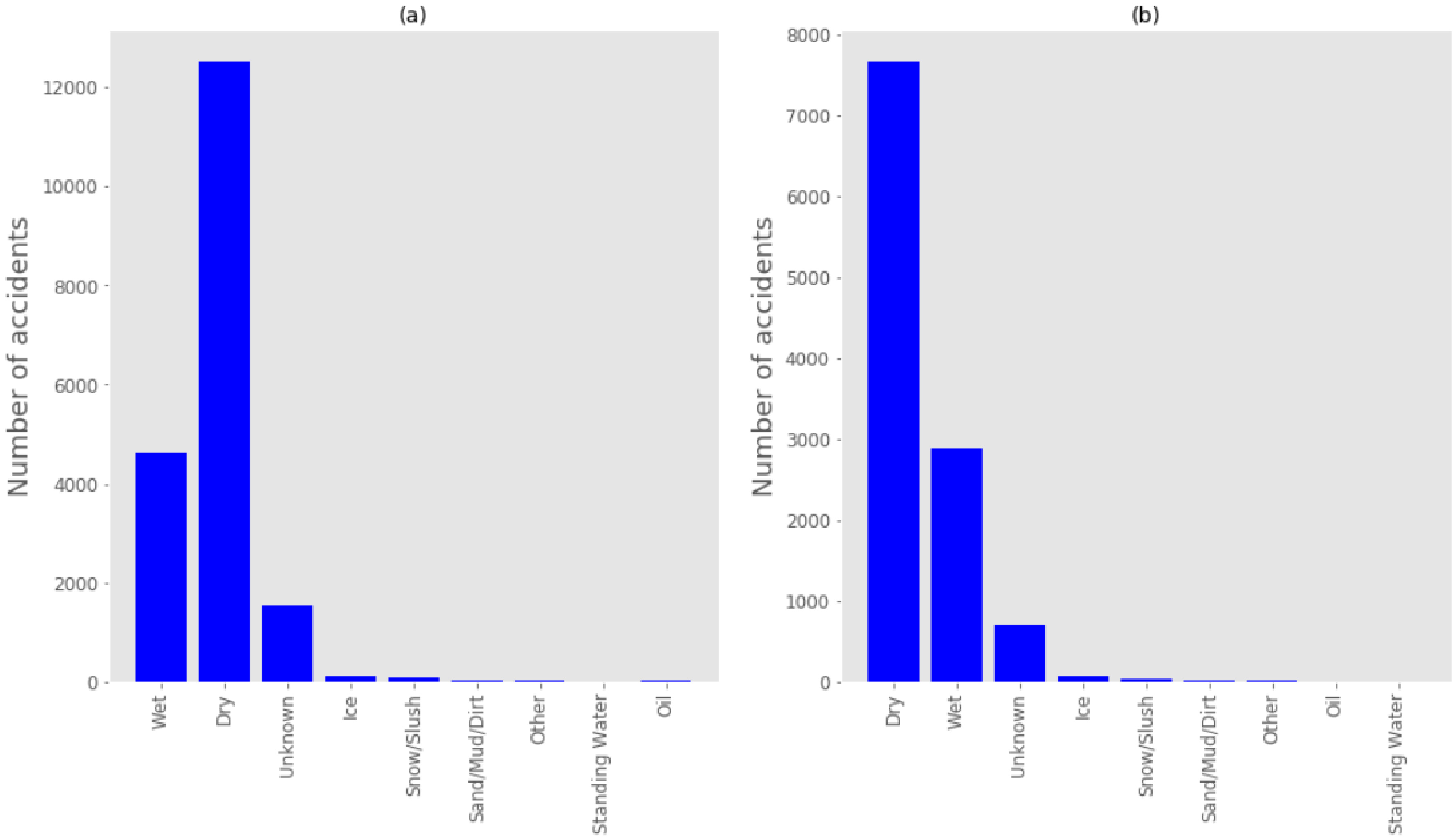
Distribution of weather conditions recorded by SDOT before and after resampling



Distribution of light conditions recorded by SDOT before and after resampling



Distribution of road conditions recorded by SDOT before and after resampling



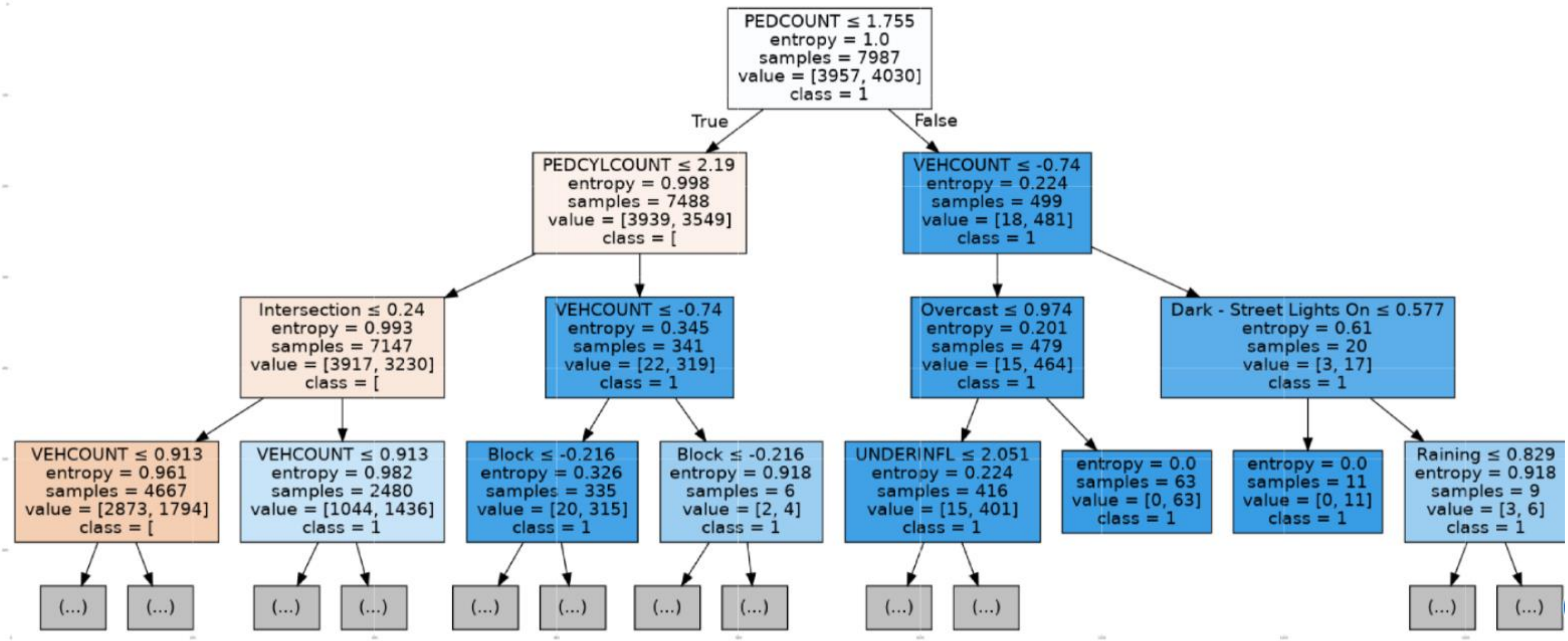
Data preprocessing before modeling

● Feature Selection – After cleaning and balancing dataset, final feature set is selected to predict the severity code. It consists of 11410 rows and 33 columns.

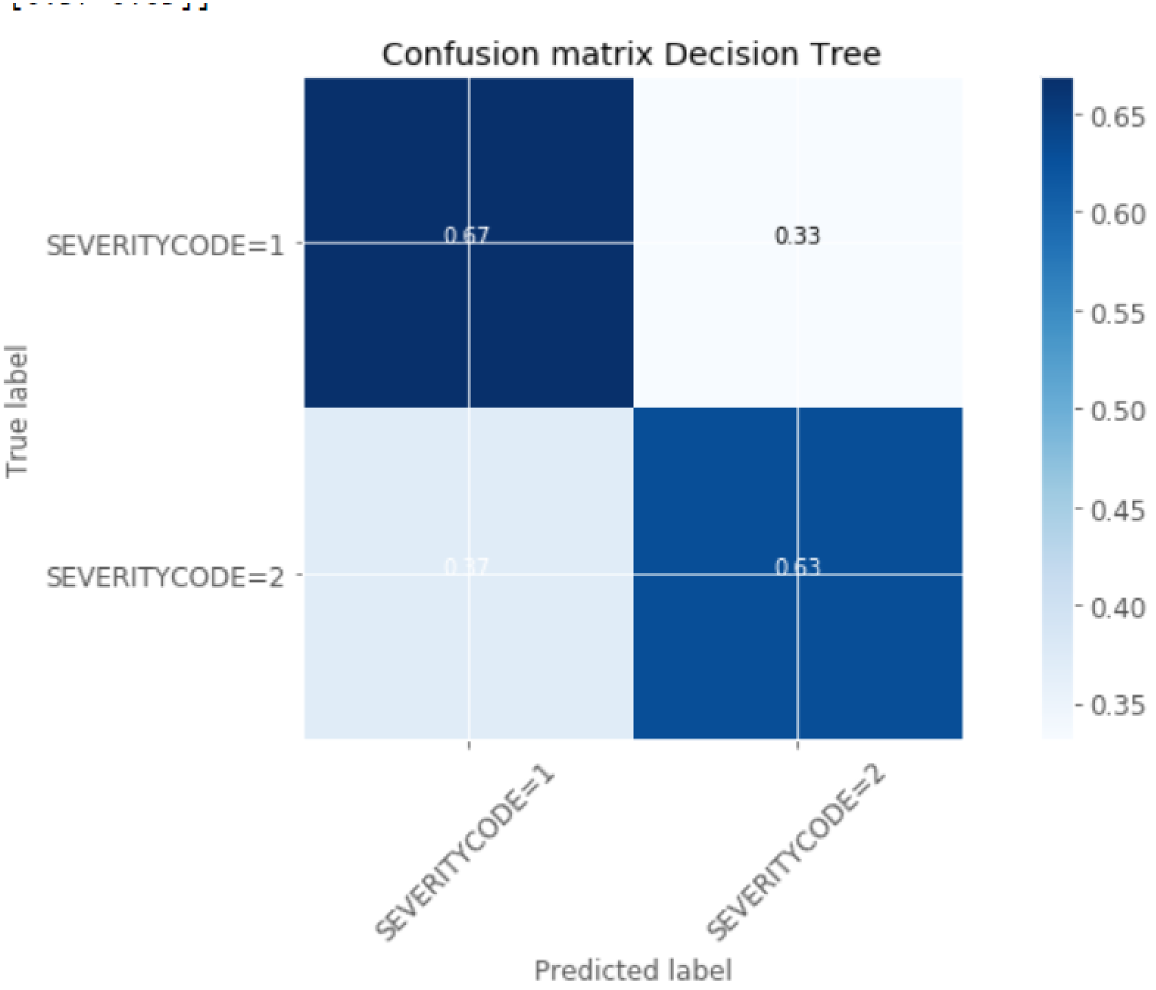
●Standardization – feature dataset is standardized having a distribution with zero mean and unit variance. This is important step in order to ensure that the model is not biased towards or away from certain feature types.

●Train – Test split – Finally data were split into training and testing datasets. 70% of balanced data were used for training the model and 30 % reserved for testing.

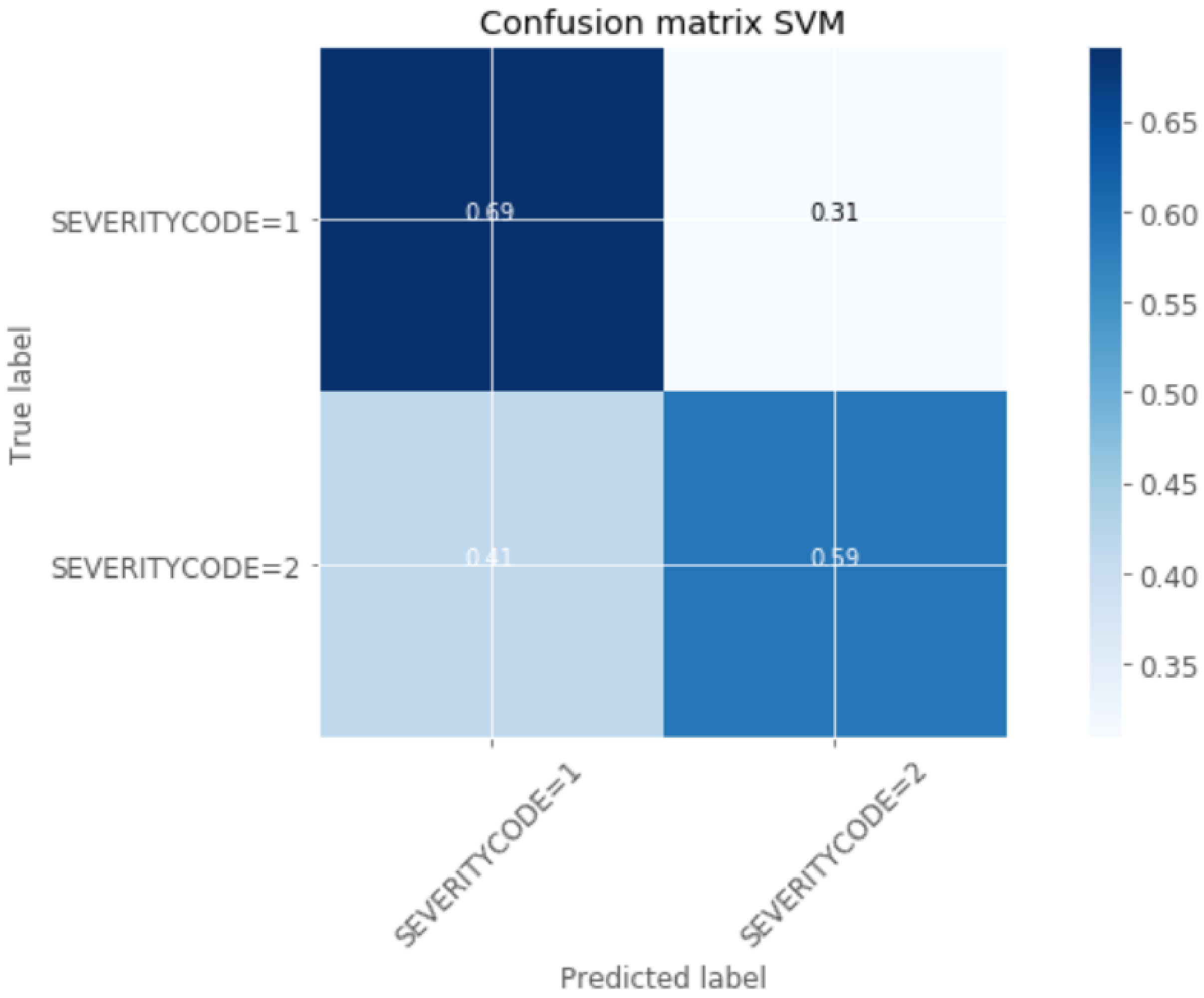
1. Decision Tree Model



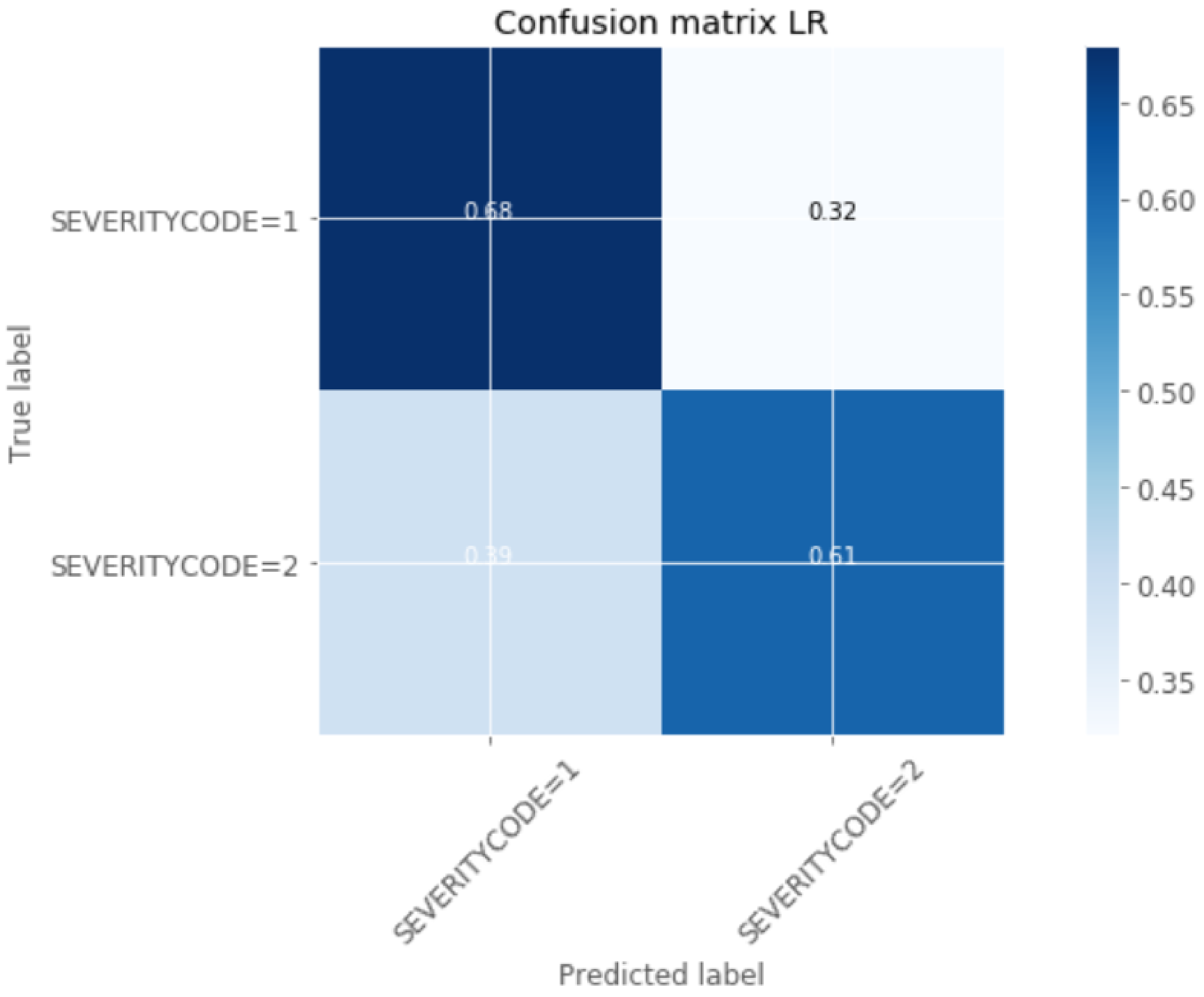
1. Decision Tree Model



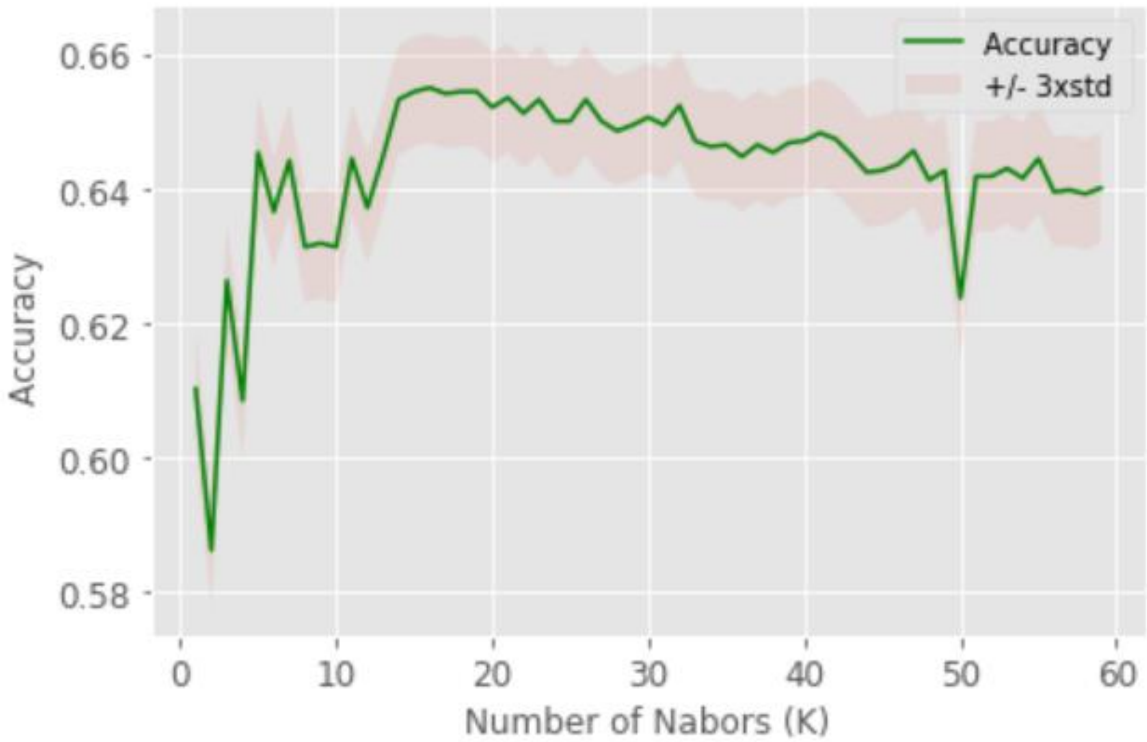
2. Support Vector Machine Model

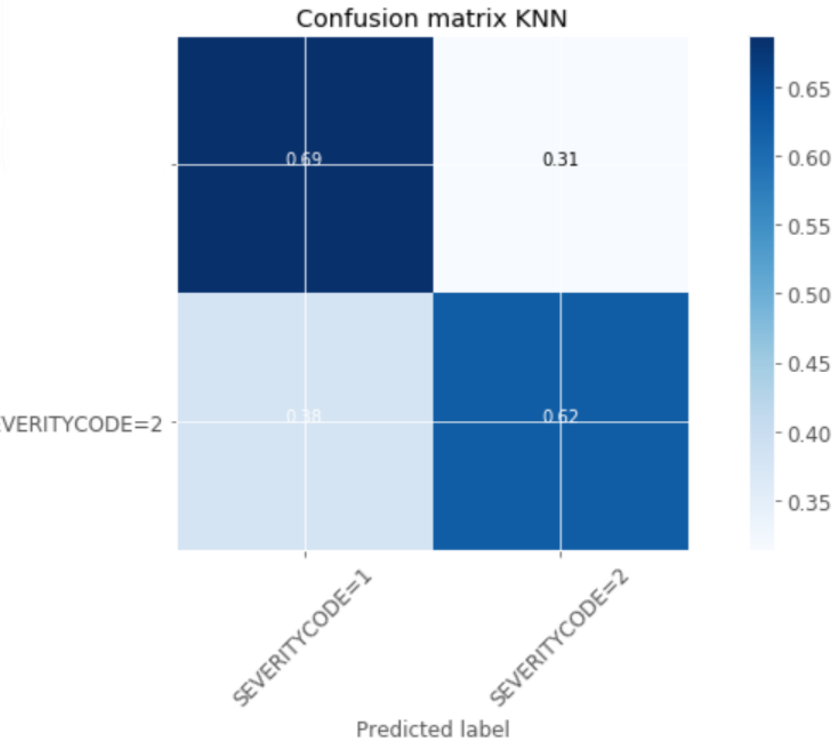


3. Logistic Regression Model

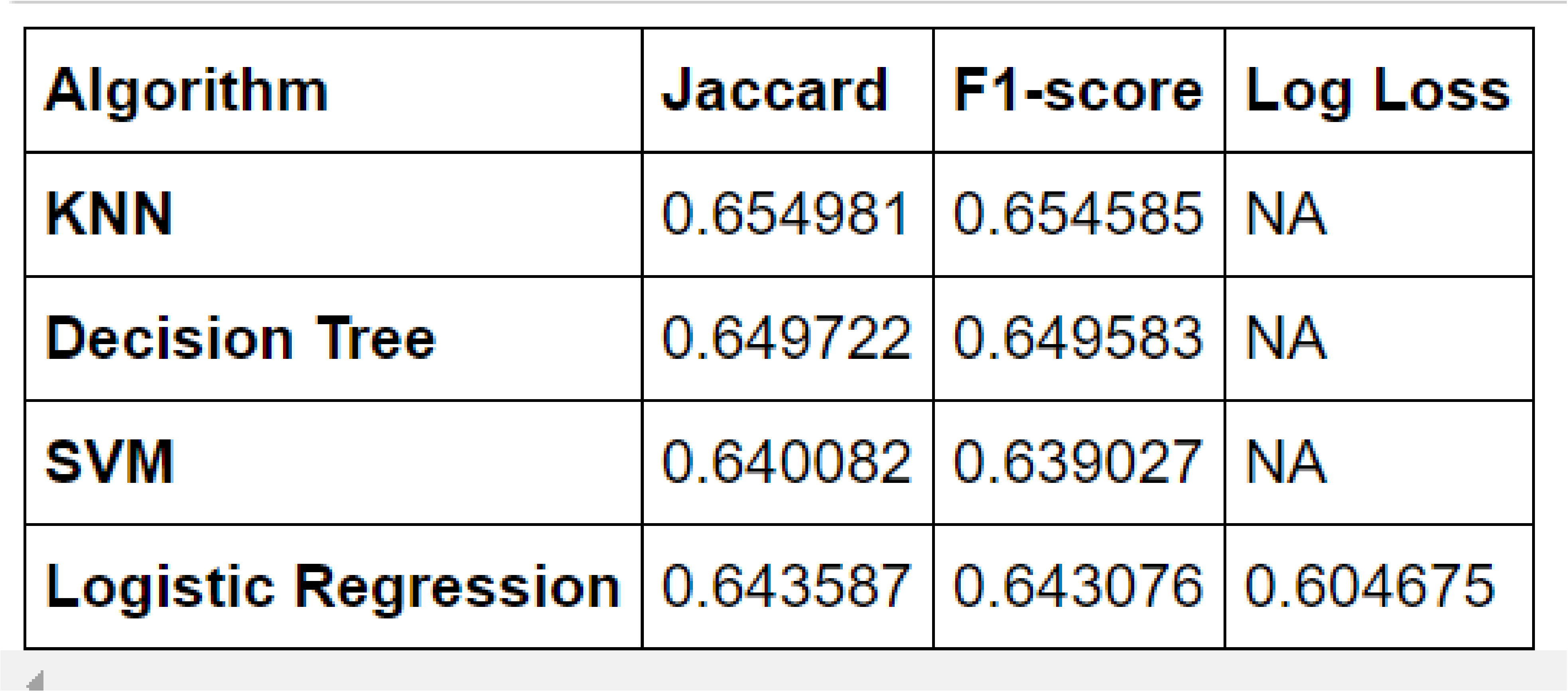


4. K- Nearest Neighbor Model





Comparitive performances of each of the four machine learning models



Algorithm	Jaccard	F1-score	Log Loss
KNN	0.654981	0.654585	NA
Decision Tree	0.649722	0.649583	NA
SVM	0.640082	0.639027	NA
Logistic Regression	0.643587	0.643076	0.604675

Conclusions and Future Work

- In this study data from Seattle Department of Transport is used to train and test the models for predicting the severity of accidents.
- The main purpose of building this model is to allow the emergency services in Seattle to allocate their resources in such way as to bet deal with accidents.
- The KNN,Decision Tree and LR models performed well with F1 Score between 0.64-0.65. Given the nearly similar performances of KNN and Decision tree models, Decision tree model is preferred.
- Cont….

Conclusions and Future Work

- Model can be adopted to any road traffic network in any part of the world
- In future model could be improved to predict severity on continuum running from 1-4 rather than predicting a binary severity of 1 or 2.

References

1. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

2. https://www.coursera.org/learn/applied-data-sciencecapstone/supplement/Nh5uS/downloading-example-dataset

3.https://github.com/snehalkolekar/Coursera_Capstone/blob/master/week%202%20assignment.ipynb