DATA SCIENCE

IBM Capstone Project

Predict Car Accident Severity

Priyanka Dave

10/10/2020

1. Introduction

1.1 Background

Every year car accidents cause hundreds of thousands of deaths worldwide. According to a research conducted by the World Health Organization (WHO) there were 1.35 million road traffic deaths globally in 2016, with millions more sustaining serious injuries and living with long-term adverse health consequences. Globally, road traffic crashes are a leading cause of death among young people, and the main cause of death among those aged 15{29 years. Road traffic injuries are currently estimated to be the eighth leading cause of death across all age groups globally, and are predicted to become the seventh leading cause of death by 2030.

Leveraging the tools and all the information nowadays available, an extensive analysis to predict traffic accidents and its severity would make a difference to the death toll. Analyzing a significant range of factors, including weather conditions, locality, type of road and lighting among others, an accurate prediction of the severity of the accidents can be performed. Thus, trends that commonly lead to severe traffic incidents can help indentifying the highly severe accidents. This kind of information could be used by emergency services, to send the exact required staff and equipment to the place of the accident, leaving more resources available for accidents occurring simultaneously. Moreover, this severe accident situation can be warned to nearby hospitals which can have all the equipment ready for a severe intervention in advance.

Consequently, road safety should be a prior interest for governments, local authorities and private companies investing in technologies that can help reduce accidents and improve overall driver safety.

1.2 Problem

Data that might contribute to determining the likeliness of a potential accident occurring might include information on previous accidents such as road conditions, weather conditions, place of the accident, type of vehicles involved in the accident, information on the people involved in the accident and the severity of the accident. This projects aims to forecast the severity of accidents with previous information that could be given by a witness informing the emergency services.

1.3 Interest

Governments should be highly interested in accurate predictions of the severity of an accident, in order to reduce the time of arrival and to make a more efficient use of the resources, and thus save a significant amount of people each year. Others interested could be private companies investing in technologies aiming to improve road safeness.

2. Data

2.1 Data Source

Download dataset from Click Here.

2.2 Feature Selection

The data set contains information on the place, type of collision, number of vehicles involved in collision, number of people involved in collision, weather and lighting conditions and type of intersection where it occurred, X and Y coordinates along with accident identification numbers.

Initial analysis was performed for the selection of the most relevant features for this specific problem, reducing the size of the dataset and avoiding redundancy, click here. With this process the number of features was reduced from 38 to 12.

2.3 Description

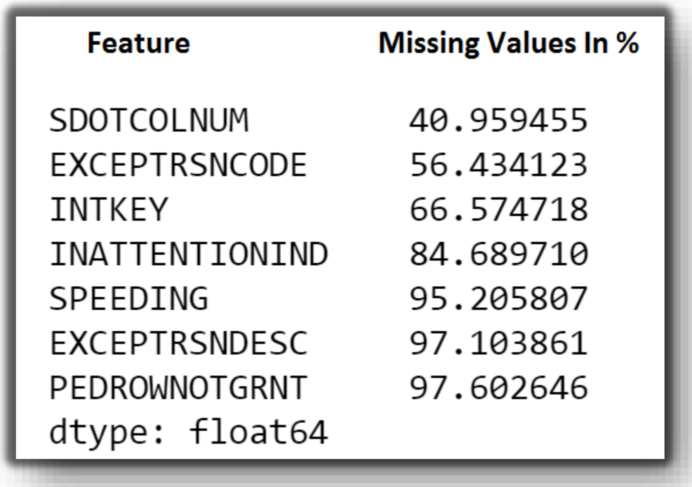
The data set resulted from the feature selection stage having 194673 samples and 12 features. Below is the list of features and its description:

Features	Description
SEVERITYCODE	Target variable.
X	X co-ordinate
Y	Y co-ordinate
ADDRTYPE	Collision address type.
COLLISIONTYPE	Collision type.
PERSONCOUNT	Total number of people involved in the collision.
VEHCOUNT	Total number of vehicles involved in the collision.
JUNCTIONTYPE	Category of junction at which collision took place
UNDERINFL	Whether or not a driver involved was under the influence of drugs or alcohol.
WEATHER	Description of the weather conditions during the time of the collision.
ROADCOND	Condition of the road during the collision.
LIGHTCOND	Light conditions during the collision.

2.4 Data Cleaning

The data cleaning is the process of giving a proper format to the data for its further analysis.

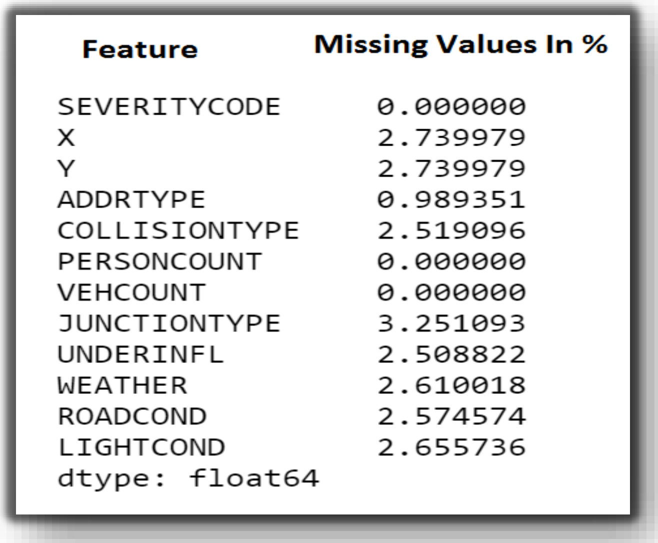
- Initially features having identification number, duplicate features and irrelevant features were removed.
- Features having more than 40% missing values were removed.



```
  Feature               Missing Values In %

SDOTCOLNUM             40.959455
EXCEPTRSNCODE          56.434123
INTKEY                 66.574718
INATTENTIONIND         84.689710
SPEEDING               95.205807
EXCEPTRSNDESC          97.103861
PEDROWNOTGRNT          97.602646
dtype: float64
```

- Initial data set was imbalanced. 30% Vs. 70% ratio was there in target distribution. Used down sampling method to balance data set.
- Removed remaining rows having missing values by checking its proportion with data set. Missing values less than or equal to 3% removed from the data set.



```
  Feature          Missing Values In %

SEVERITYCODE       0.000000
X                  2.739979
Y                  2.739979
ADDRTYPE           0.989351
COLLISIONTYPE      2.519096
PERSONCOUNT        0.000000
VEHCOUNT           0.000000
JUNCTIONTYPE       3.251093
UNDERINFL          2.508822
WEATHER            2.610018
ROADCOND           2.574574
LIGHTCOND          2.655736
dtype: float64
```

- Merged duplicate and ambiguous categories to remove ambiguity.
  - Feature **UNDERINFL** had ambiguous categories. Merged **Y** with **1** and **N** with **0**.
  - Feature **WEATHER** had ambiguous categories. Merged **Sleet/Hail/Freezing Rain** with **Snowing**, **Other** with **Unknown**, **Blowing Sand/Dirt** with **Severe Crosswind**.
  - Feature **ROADCOND** had ambiguous categories. Merged **Other** with **Unknown**.
  - Feature **LIGHTCOND** had ambiguous categories. Merged **Other** with **Unknown**, **Dark - Street Lights Off** with **Dark - No Street Lights**.
- Converted categorical values to numeric .
- Then normalized data using data scaling method.

3. Exploratory Data Analysis

First, distribution of the target values was visualized. The plot confirmed that it is a balanced dataset as we have down sampled majority class from original dataset. Below figure shows category distribution in dataset before and after down sampling data.

- **1: Property Damage**
- **2: Injury**

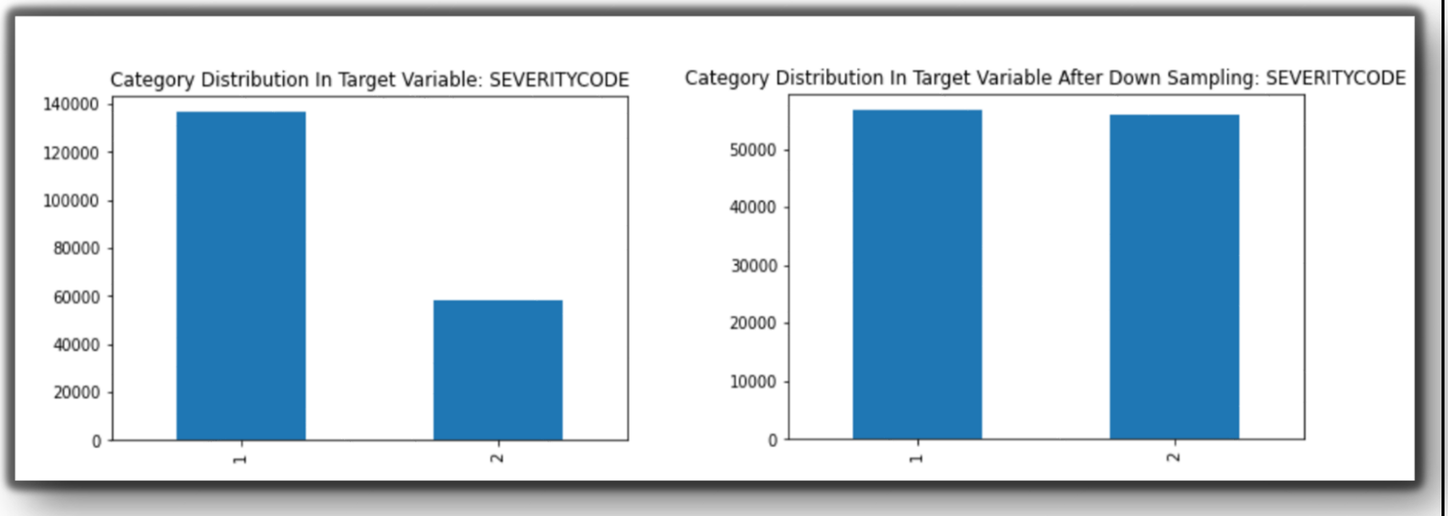


Figure 1: Category distribution in target variable before and after down sampling

Below graph shows frequency of Injury in collision is lower than frequency of Property Damage.

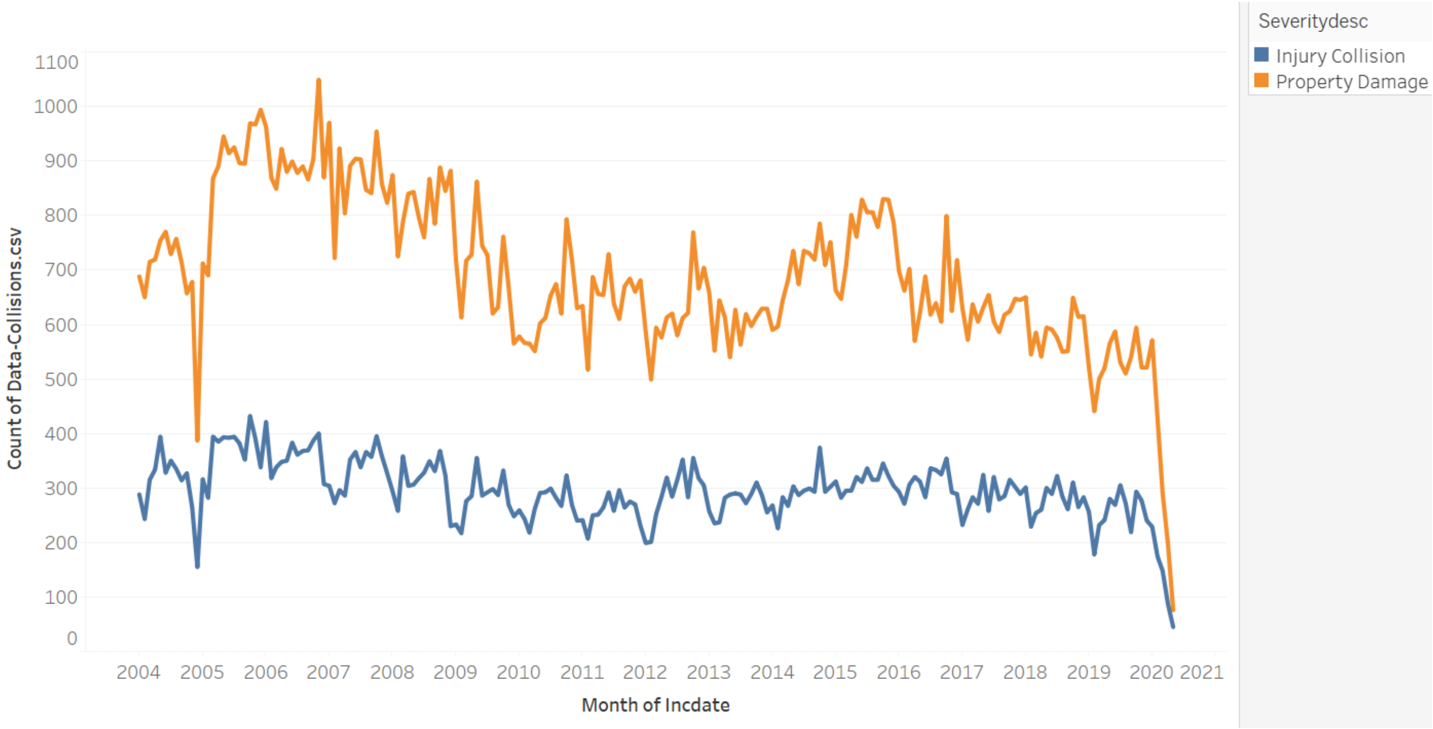


Figure 2: Month wise collision data from 2004 to 2021

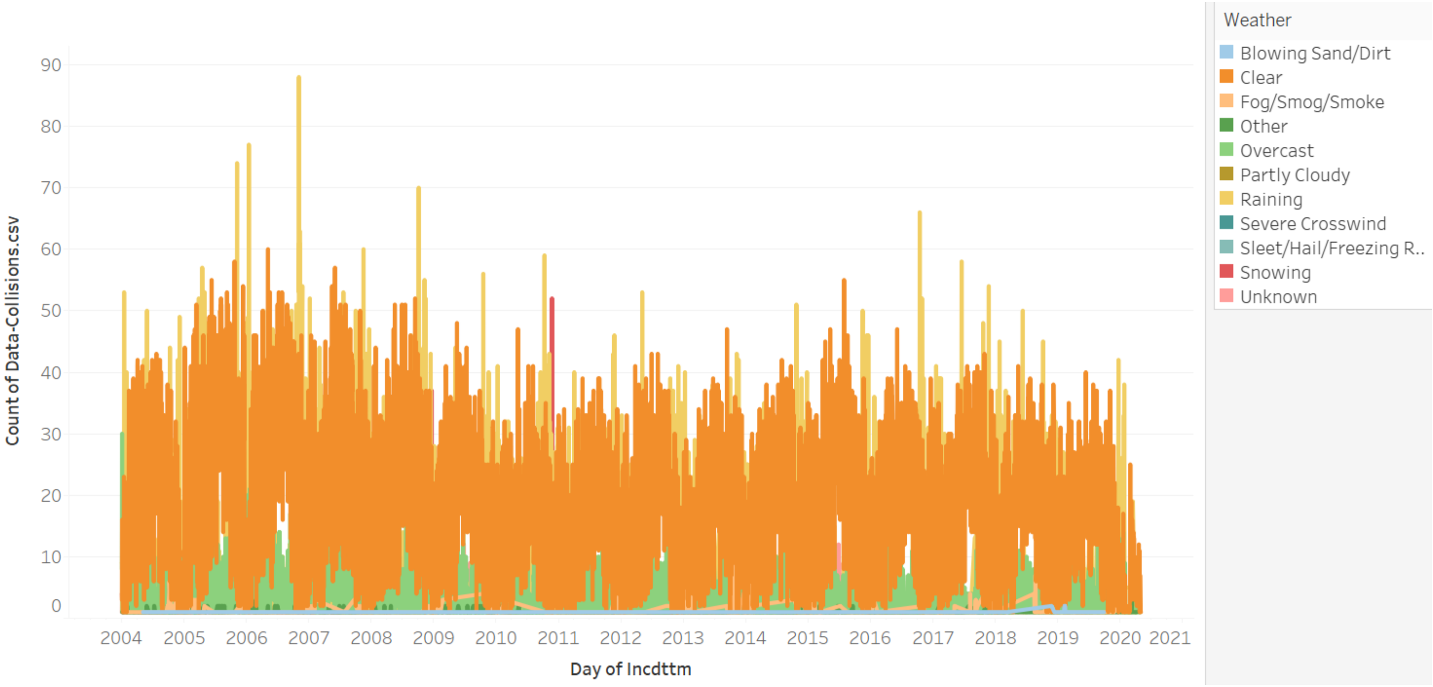


Figure 3: Daily collision data categorized by Weather

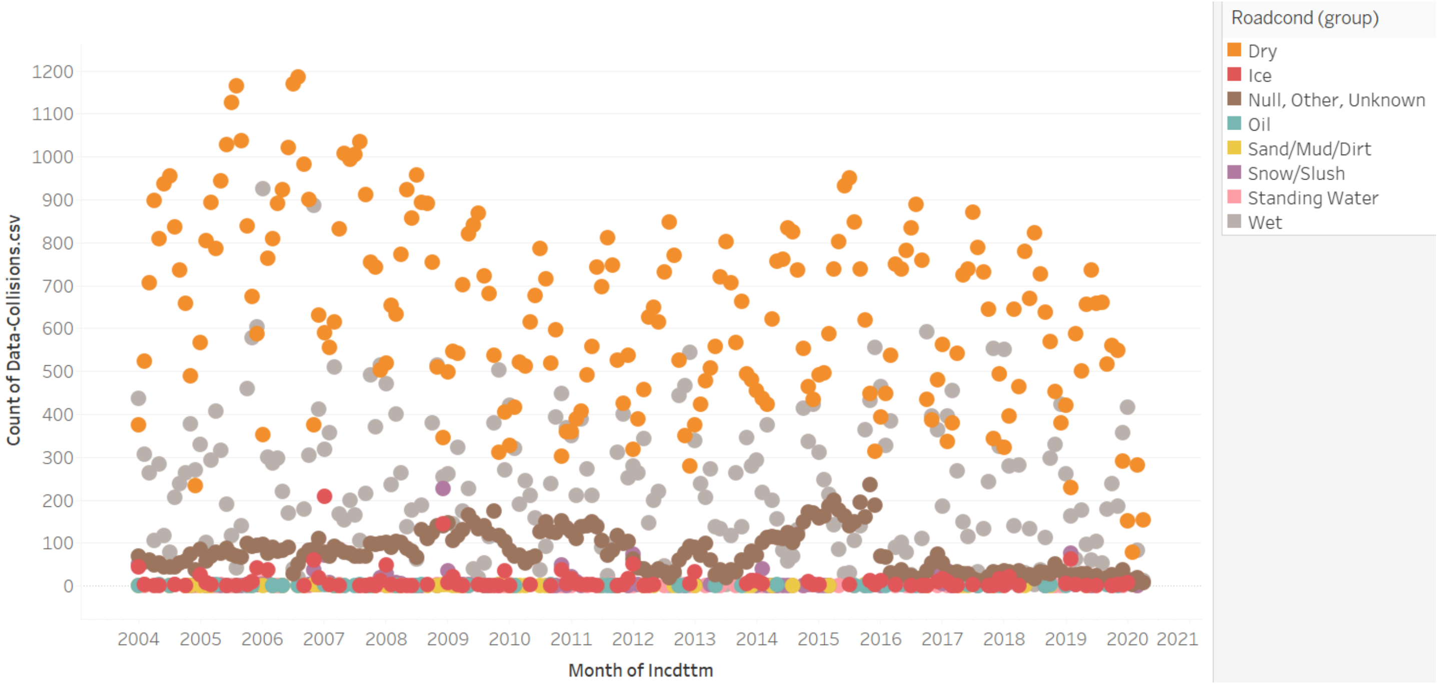


Figure 4: Monthly collision data categorized by Road condition

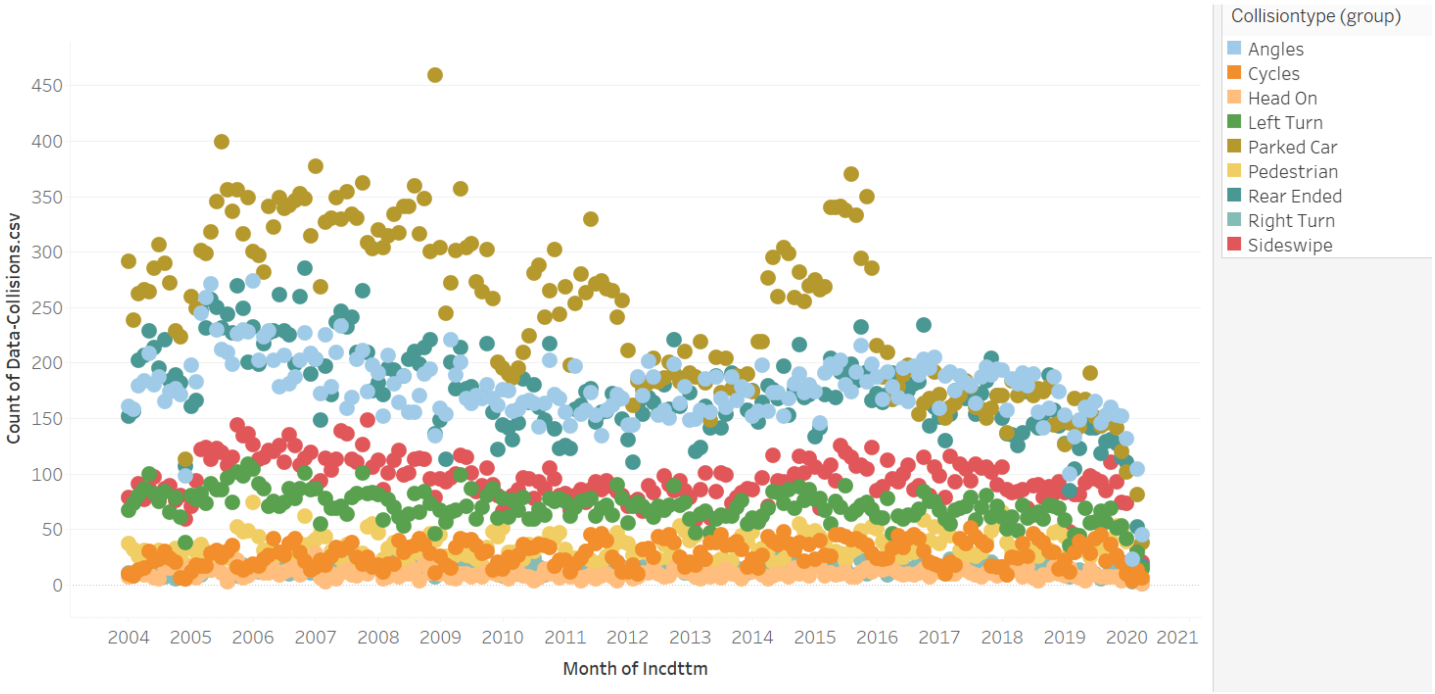


Figure 5: Monthly collision data categorized by Collision type

Then checked category distribution of other features with respect to target variable along with data distribution in continuous features and visualized results. It helped to merge duplicate and ambiguous categories within data set.

It also helped to check pattern in data before performing down sampling and after performing down sampling. If data distribution gets changed after applying any up sampling or down sampling technique, that means either we may have generated bad data or have lost some important information from original data set. Here pattern remains intact after applying down sampling technique that means we have not lost any important information.

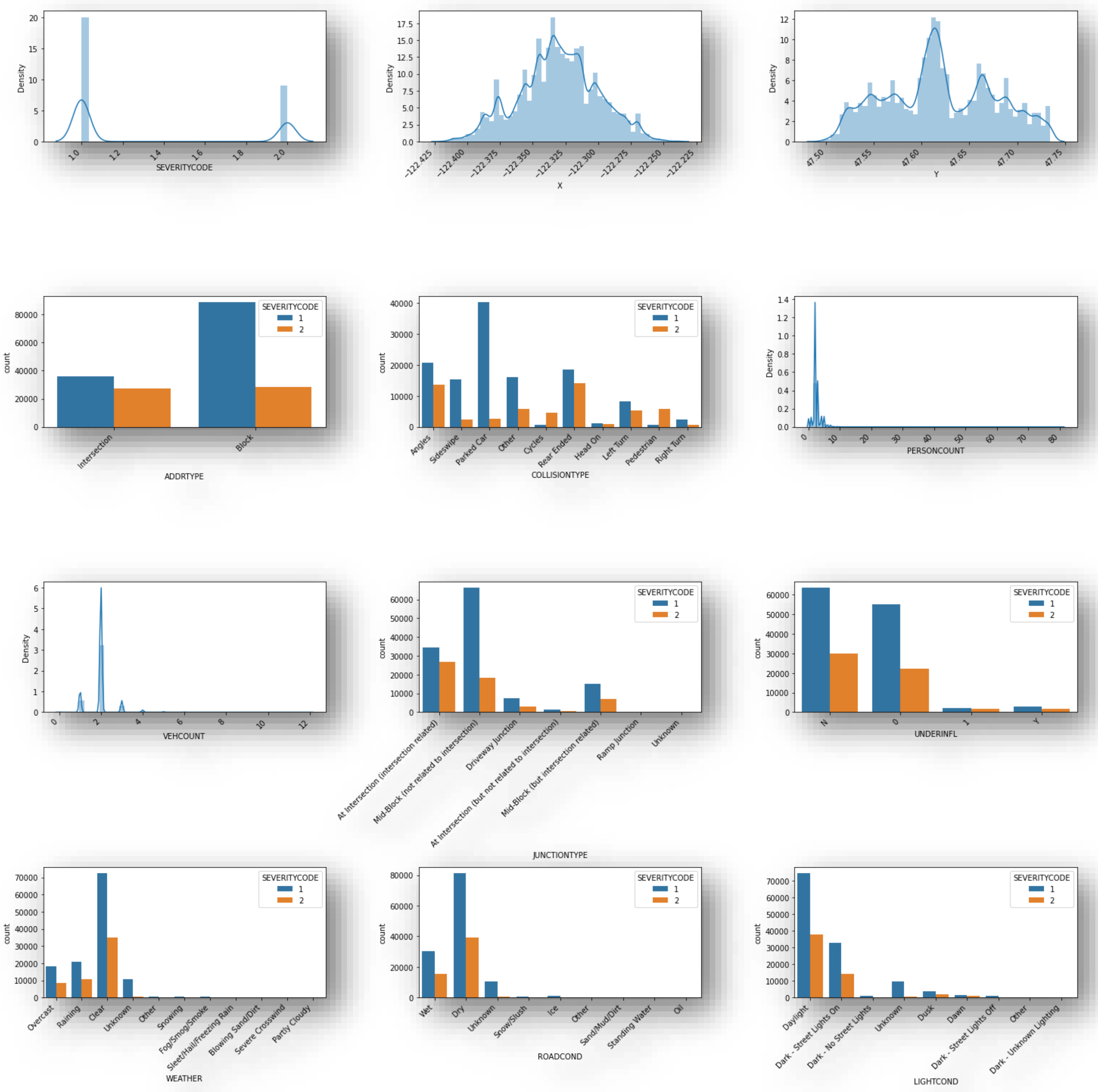


Figure 6: Data distribution with respect to the target variable: Before down sampling

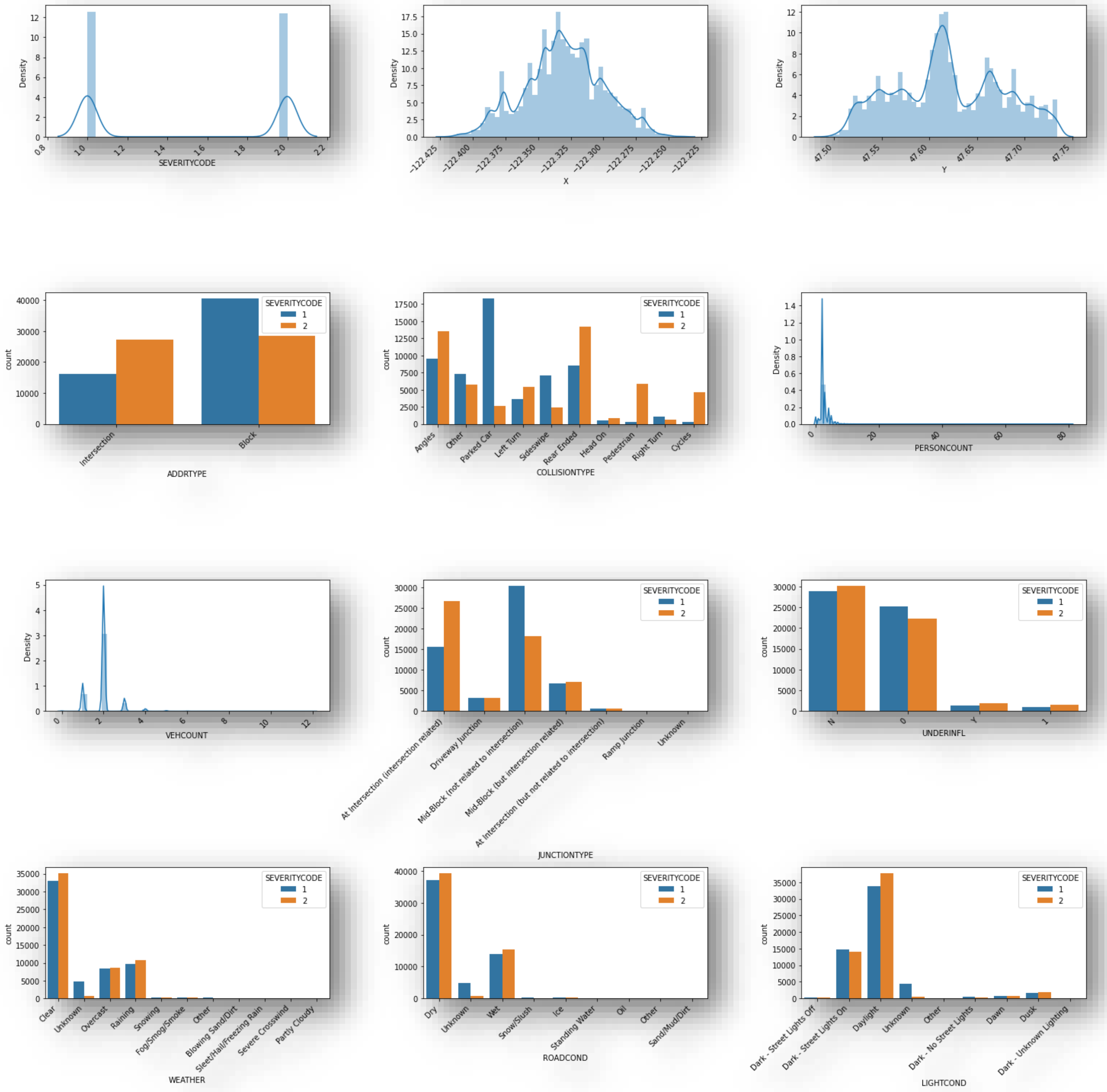


Figure 3: Data distribution with respect to the target variable: After down sampling

Below graph shows that outliers exists in data set. Removed outliers.

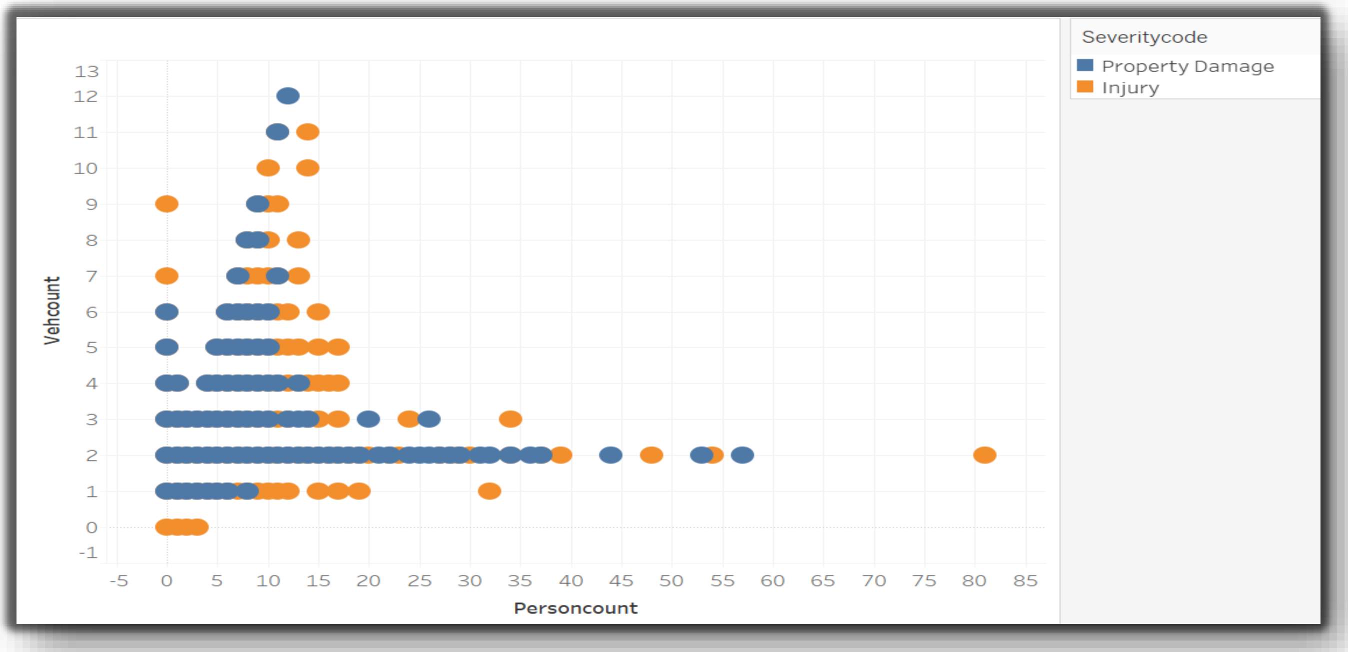


Figure 4: Scatter plot of PERSONCOUNT Vs. VEHCOUNT

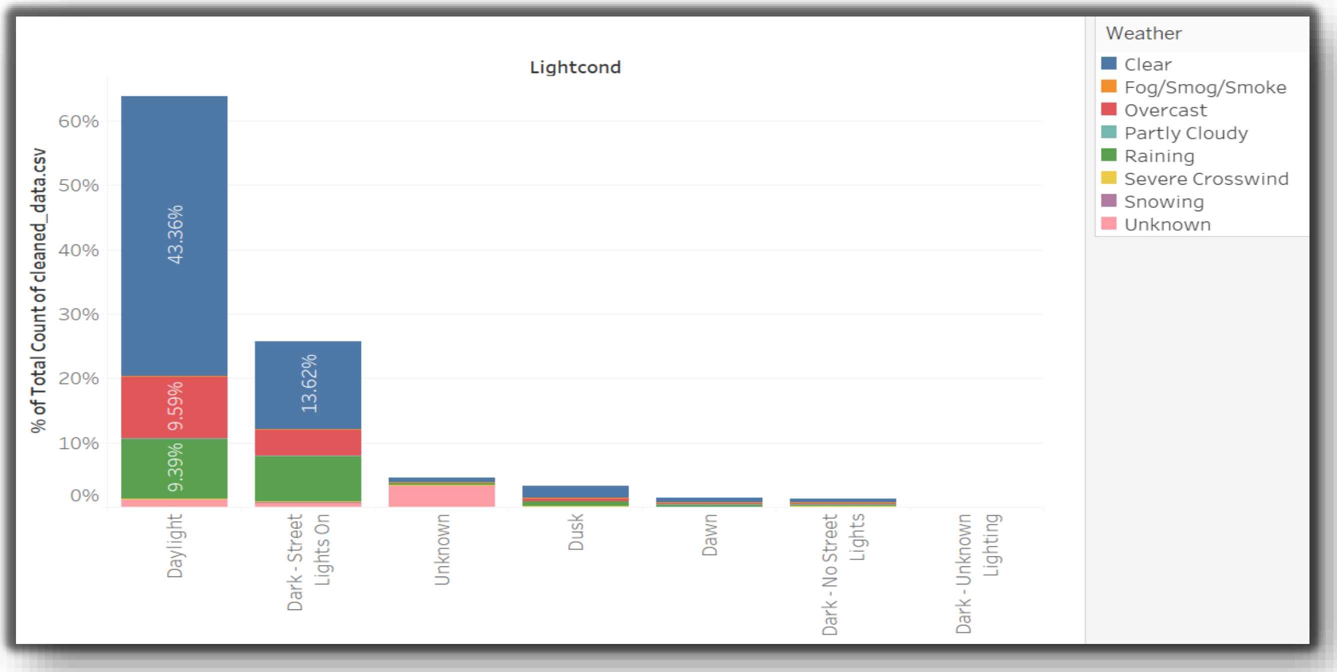


Figure 5: Category Distribution In LIGHTCOND Vs. WEATHER

Above graph shows that maximum collisions happened in Daylight when weather condition was during clear, overcast and raining.

4. Predictive Modeling

Different classification algorithms have been tuned and built to predict accident severity. These algorithms provides supervised learning. Accuracy and computational time have been compared in order to determine the best suited algorithm for his specific problem.

First, prepared training and testing data set from **102937** samples and 48 features by splitting original data set into 70/30 ratio. Took 70% data for model training and remaining 30% for model testing.

Four classification algorithms were used to compare result:

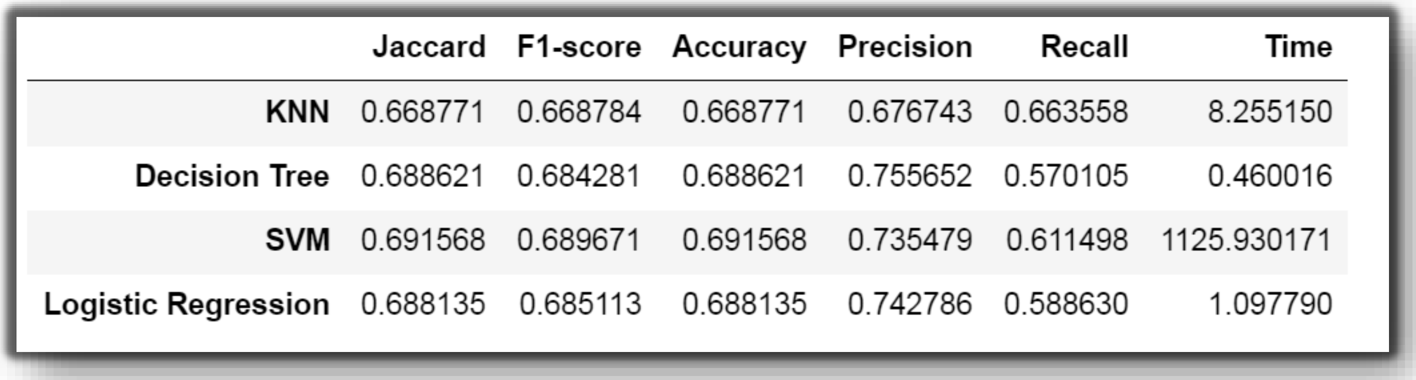
1. K-Nearest Neighbor
2. Decision Tree
3. Support Vector Machines
4. Logistic Regression

Also each model was evaluated using different evaluation methods:

1. Accuracy
2. F1-Score
3. Jaccard similarity score
4. Precision
5. Recall
6. Computational Time

5. Results

Below table shows the results of evaluation of each model.



	Jaccard	F1-score	Accuracy	Precision	Recall	Time
KNN	0.668771	0.668784	0.668771	0.676743	0.663558	8.255150
Decision Tree	0.688621	0.684281	0.688621	0.755652	0.570105	0.460016
SVM	0.691568	0.689671	0.691568	0.735479	0.611498	1125.930171
Logistic Regression	0.688135	0.685113	0.688135	0.742786	0.588630	1.097790

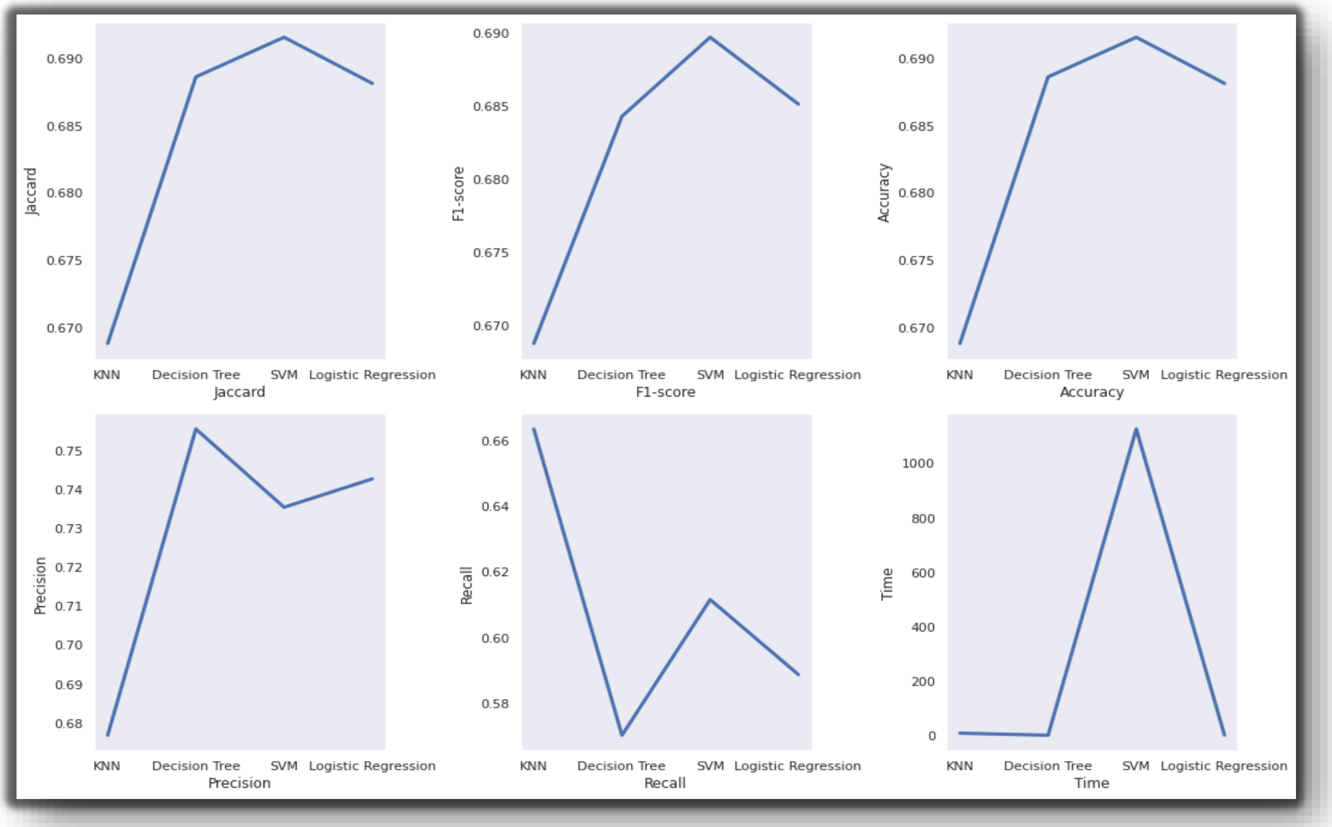


Figure 6: Comparison between different models

Here precision means the % of predicted collisions involved injuries were truly involved injuries. The recall instead, is the % of collisions truly involved injuries that were properly predicted. For this specific problem, the recall is more important than the precision as a high recall will favor that all required resources will be equipped up to predict collisions involve injuries.

F1 score and Jaccard similarity score for SVM is best. However SVM is taking maximum to train. KNN, Decision Tree and Logistic regression having almost similar accuracy.

Value of Recall is highest for KNN.

6. Conclusion

In this study, I analyzed the relationship between severity of an accident and some characteristics which describe the situation that involved the accident. Initially I thought that features such as atmospheric conditions, the lighting vehicle count would be the most relevant ones, yet I identified the road category and type of collision among the most important features that affect to the gravity of the accident. I built and compared 4 different classification models to predict whether an accident would have a high or low severity. These models can have multiple application in real life. For instance, imagine that emergency services have a application with some default features such as date, time and department/municipality and then with the information given by the witness calling to inform on the accident they could predict the severity of the accident before getting there and so alert nearby hospitals and prepare with the necessary equipment and staff. Also by identifying the features that favor the most the gravity of an accident, these could be tackled by improving road conditions or increasing the awareness of the population.