Traffic Accident Severity Prediction

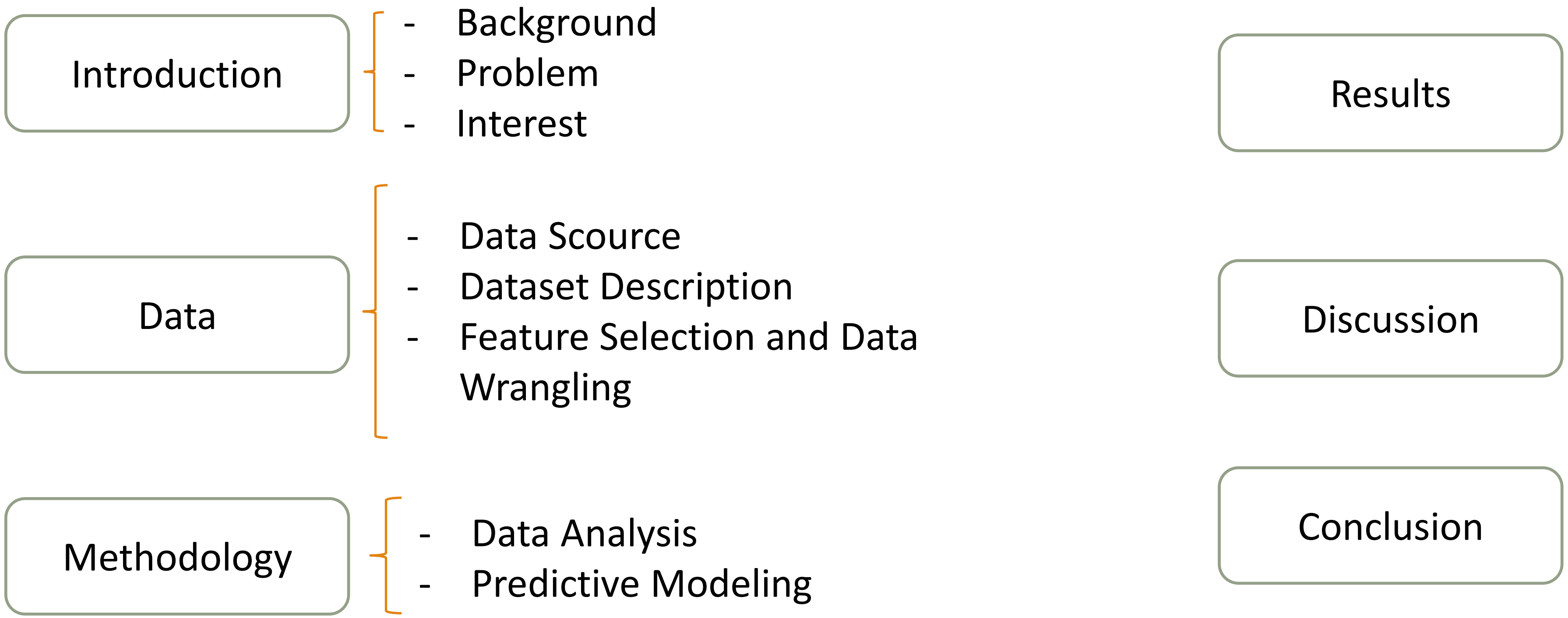
Applied Data Science Capstone

SANTIAGO ANDRES GRANDA BRAVO

sgranda996@gmail.com

OCTOBER 2020

Overview



Introduction

- Background
- Problem
- Interest

Data

- Data Scource
- Dataset Description
- Feature Selection and Data Wrangling

Methodology

- Data Analysis
- Predictive Modeling

Results

Discussion

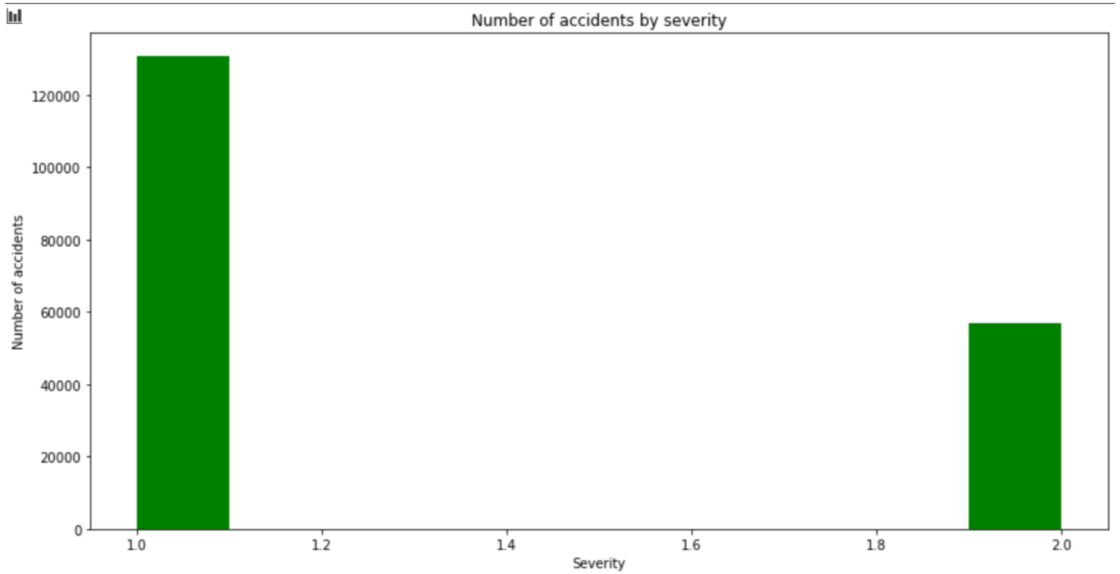
Conclusion

Introduction

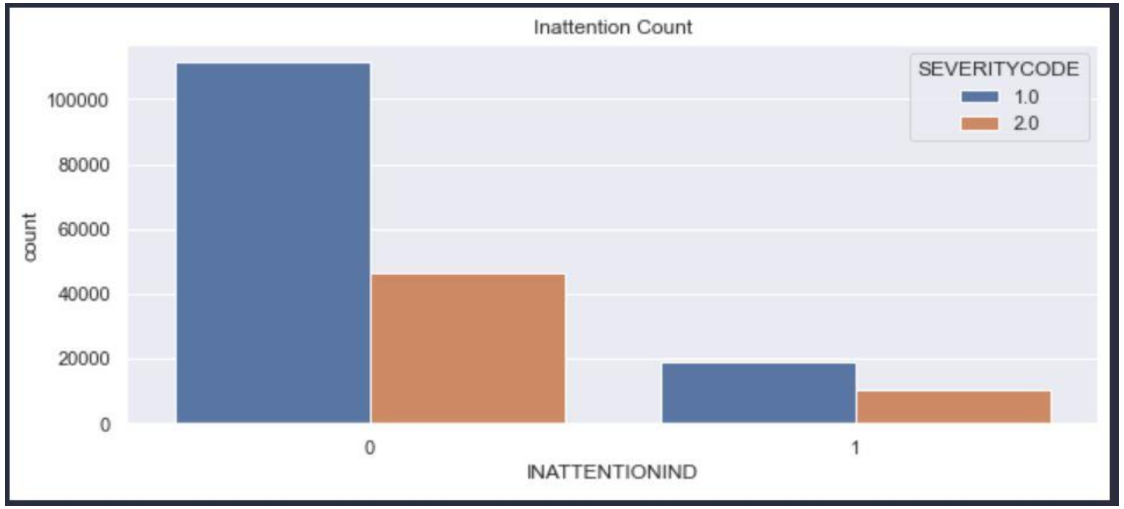
- Traffic accidents cause not only a huge amount of deaths around the world but produces economic costs to governments.
- Traffic accidents are the 8th cause of mortality in different countries and are expected to rise to the 3rd rank by 2020

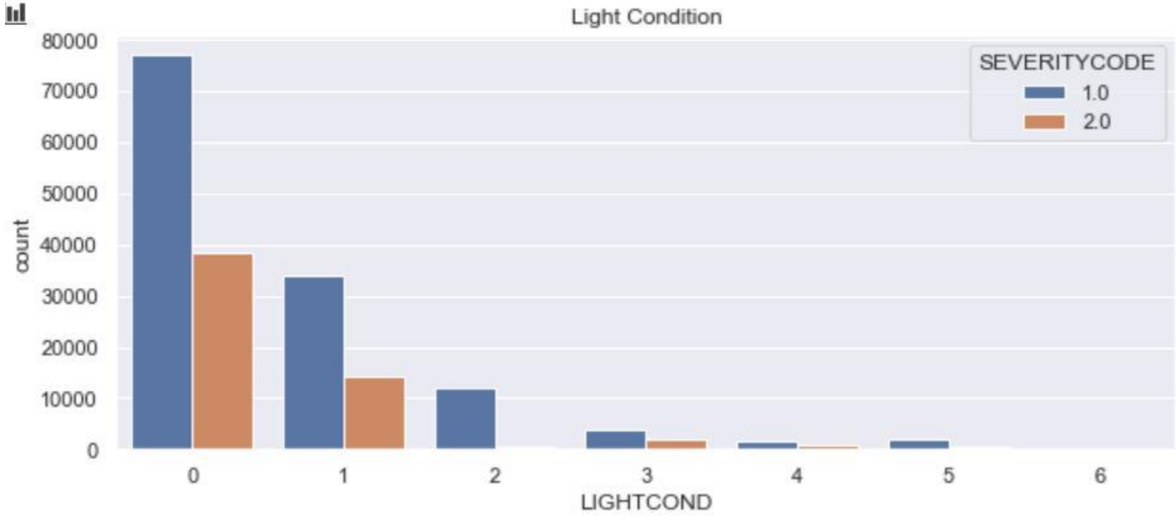
Data

- The dataset contains information of the type of collision, location, weather, the severity of the collision, and many other attributes that occurred, which are reported and described on the dataset.



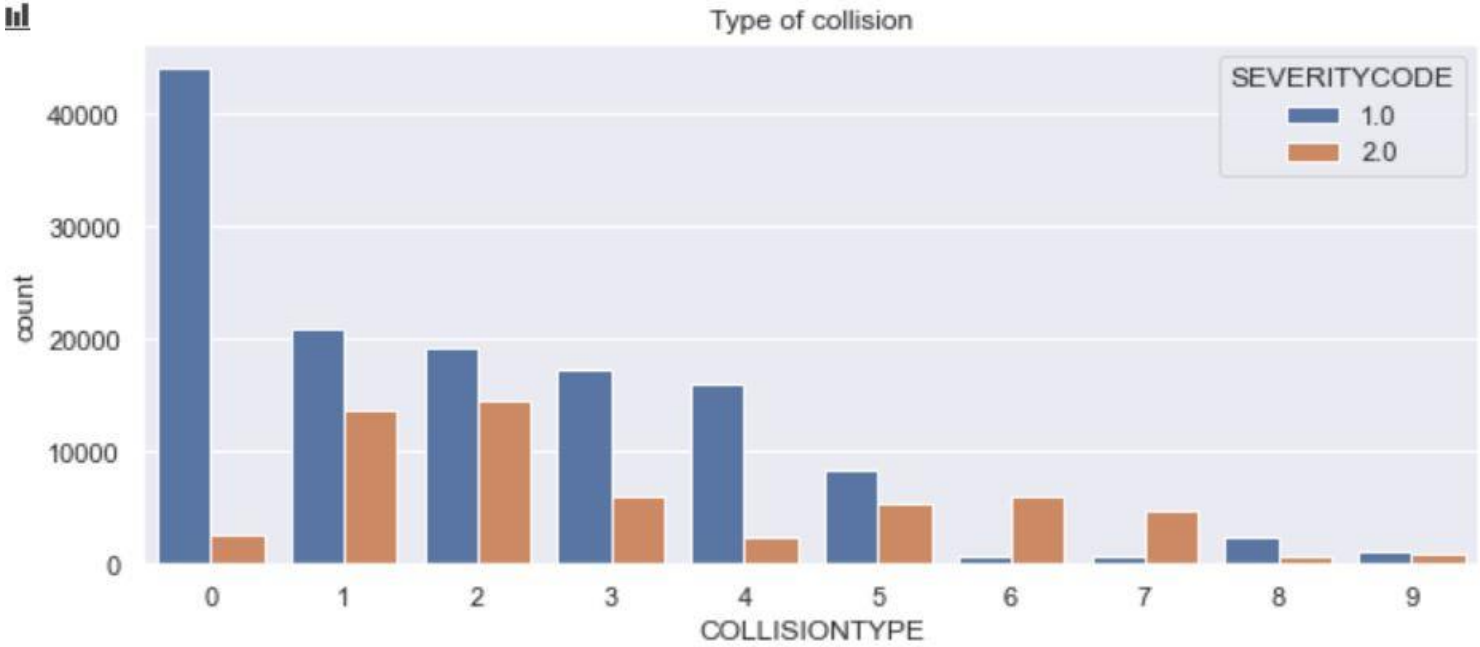
Dataset features examples devided by the severity





Accidents by inattention

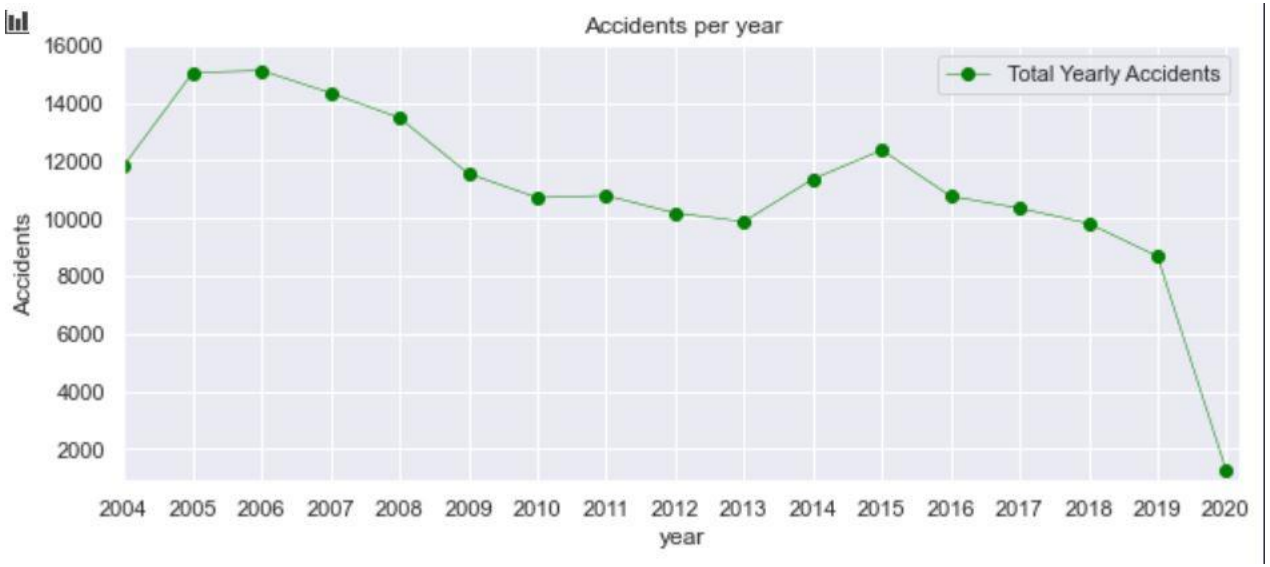
Accidents by light condition



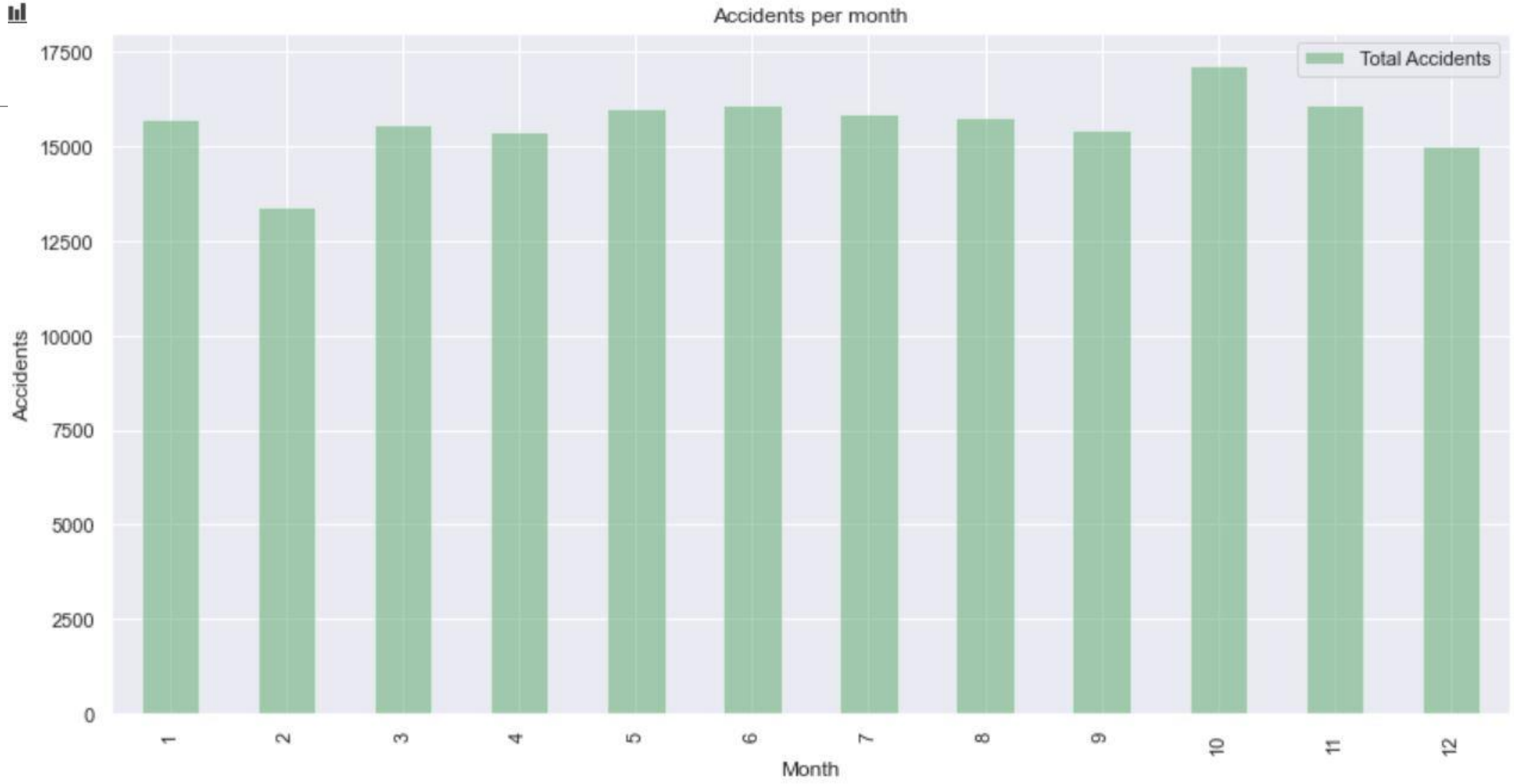
Accidents by type of collision

Methodology

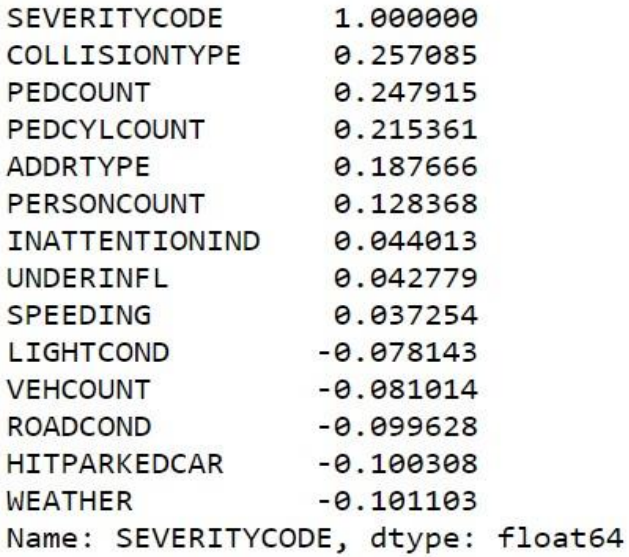
Data Analysis



Accidents per year



Accidents per month



```
SEVERITYCODE      1.000000
COLLISIONTYPE     0.257085
PEDCOUNT          0.247915
PEDCYLCOUNT       0.215361
ADDRTYPE          0.187666
PERSONCOUNT       0.128368
INATTENTIONIND    0.044013
UNDERINFL         0.042779
SPEEDING          0.037254
LIGHTCOND        -0.078143
VEHCOUNT         -0.081014
ROADCOND         -0.099628
HITPARKEDCAR     -0.100308
WEATHER          -0.101103
Name: SEVERITYCODE, dtype: float64
```

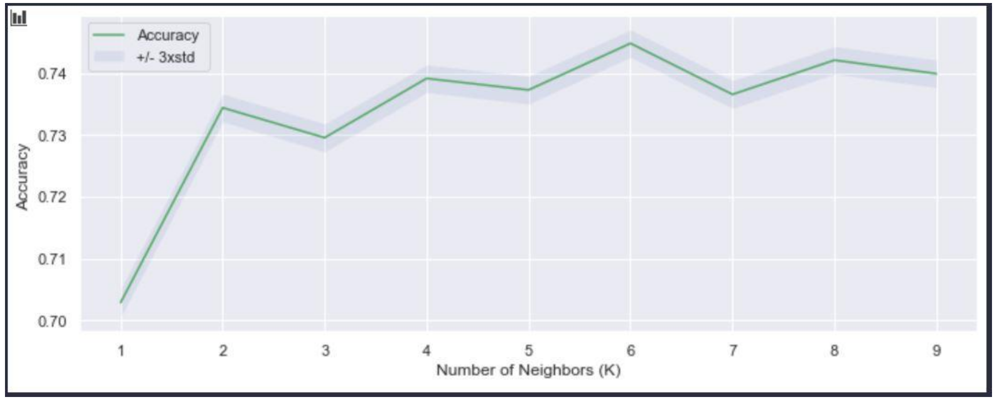
Correlation indexes

Predictive Modeling

1. Logistic Regression: C = 0.001

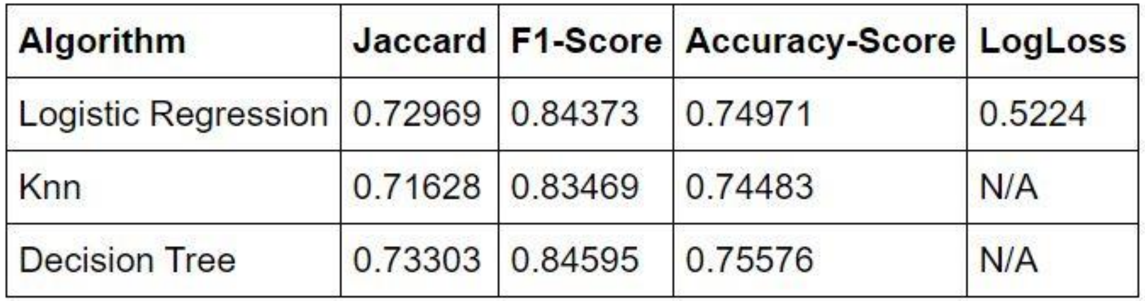
2. K-Nearest Neighbors: K = 6

3. Decision Tree



Best value of K

Results



Algorithm	Jaccard	F1-Score	Accuracy-Score	LogLoss
Logistic Regression	0.72969	0.84373	0.74971	0.5224
Knn	0.71628	0.83469	0.74483	N/A
Decision Tree	0.73303	0.84595	0.75576	N/A

Metrics result report

- The metrics enlisted before are used to compare the results and the quality for each of them between the three models analized in this project.

Discussion

- It can be observed that the results generated by the Decision tree model were the more accurate generating the best results but analyzing the Logistic regression model results we observe that the accuracy scores were very similar to those obtained by the decision tree.
- The logistic regression algorithm presented a very good approximation generating good results having the smallest run-time of all the three analyzed methods.

Conclusion

- Traffic Accidents are an important cause of deaths worldwide.
- This is the reason of the analysis on this project, to predict the fatality of accidents due to several conditions can help saving lives that can be lost on the roads.
- Even though the logistic regression model did not present the best accuracy scores, the quality vs. time results were a good metric to validate the quality of the results generated by the logistic regression model.