ANALYZING AND PREDICTING TRAFFIC ACCIDENT SEVERITY IN SEATTLE, WASHINGTON

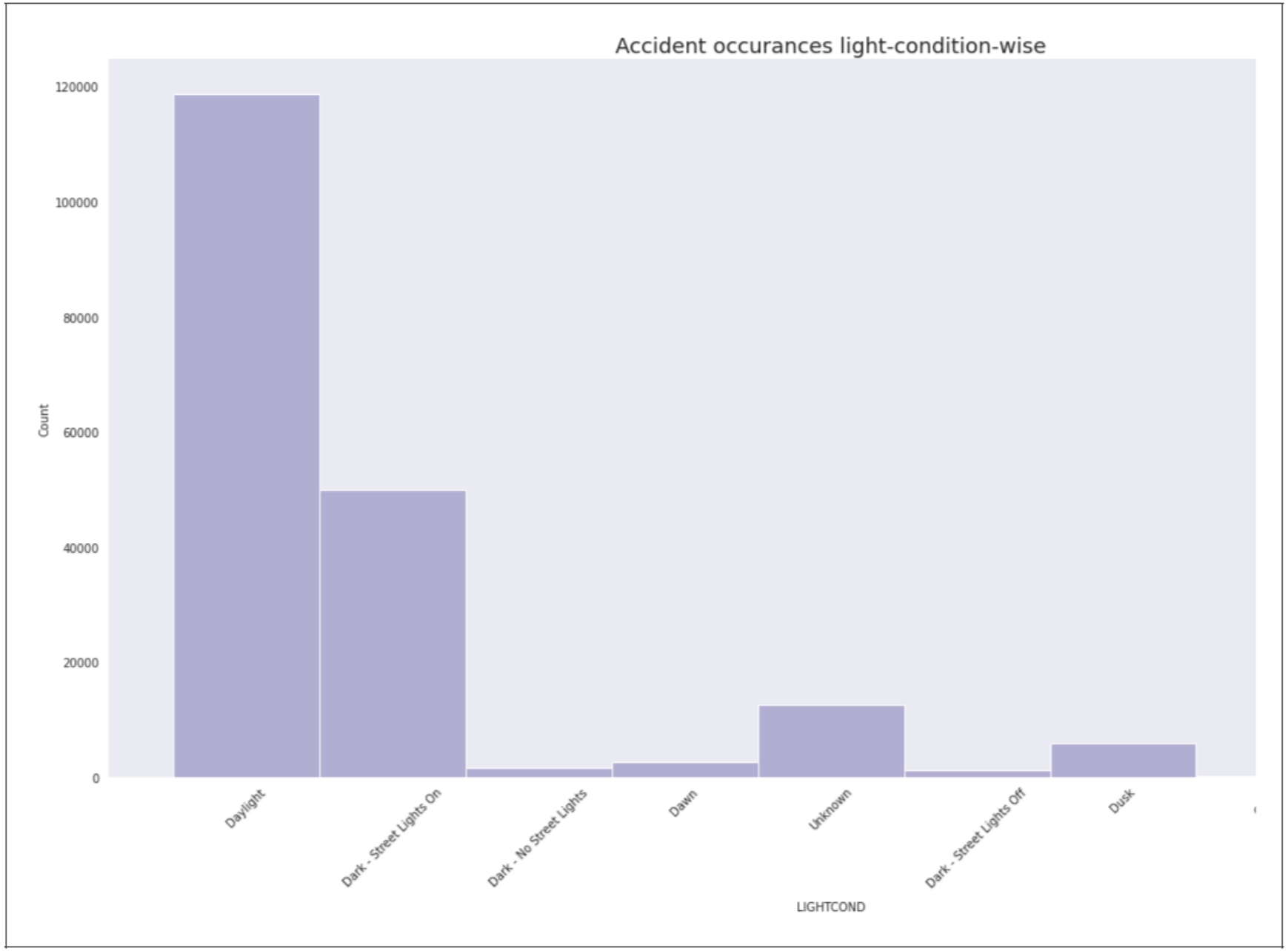
Rohit Rajesh Dube
October 5, 2020

INTRODUCTION

- According to the World Health Organization (WHO), every year approximately 1.35 million people die from traffic accidents. Out of which 93% of the world's fatalities happen in low and middle-income countries which only posses 60% of world's vehicles. On top of it, traffic accidents are the leading causes of death among children and young adults aged 5-29 years. The seaport city of Seattle is the largest city in the state of Washington, as well as the largest in the Pacific Northwest. As of the latest census, there were 713,700 people living in Seattle. Seattle residents get around by car, trolley, streetcar, public bus, bicycle, on foot, and by rail. With such bustling streets, it is no surprise that Seattle sees car accidents every day. According to data from the Washington State Department of Transportation (WSDOT), last year, Seattle saw more than 10,315 crashes on the street.

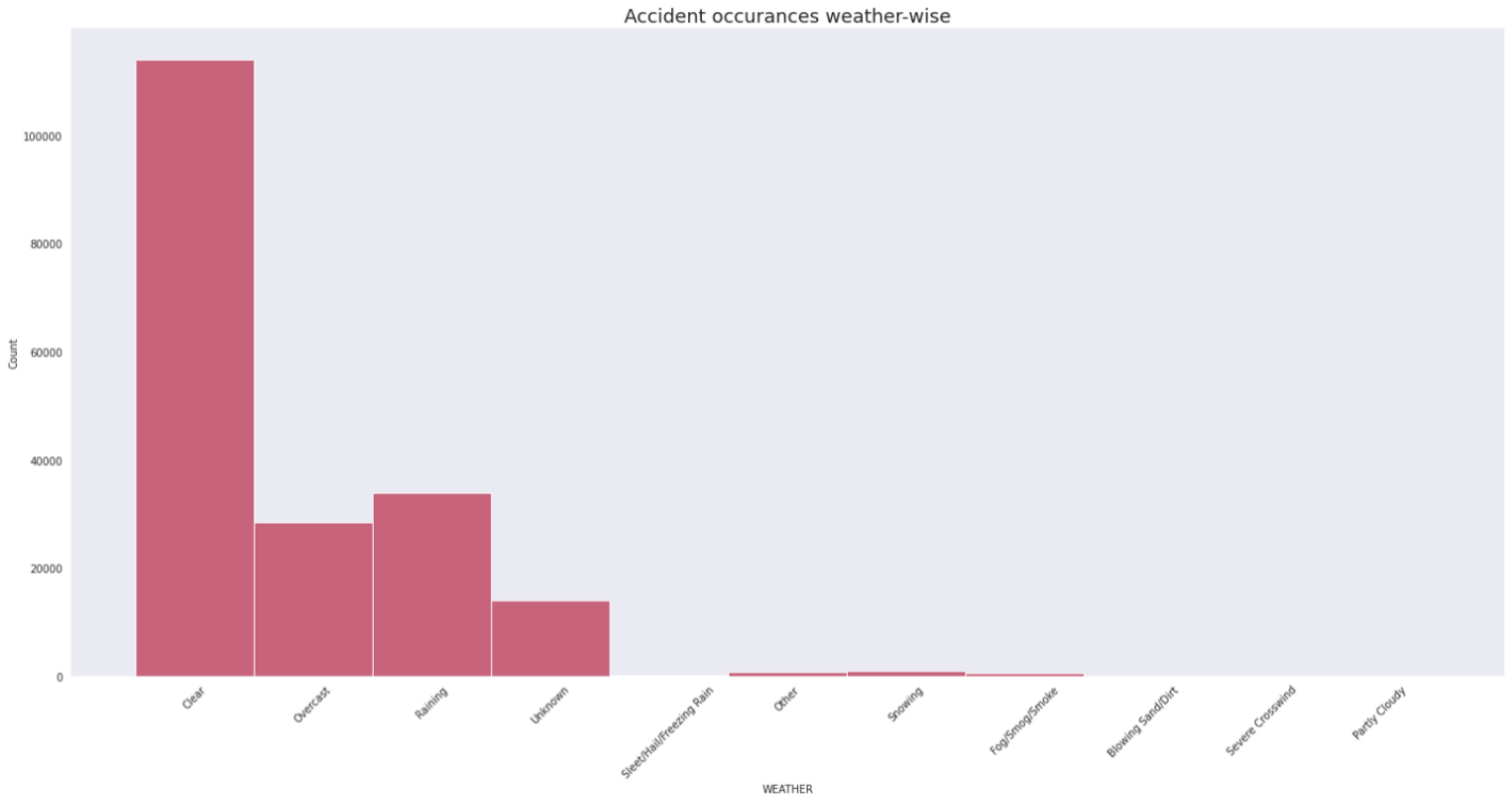
Data acquisition and cleaning

- The dataset, we will be using in this project was downloaded from the City of Seattle Open Data portal website. Seattle Police Department and Accident Traffic Records Department collected and maintained data from 2004 to present. The data includes many columns of details of the accidents and the severity of each car accidents.
- In our quest to predict the severity of an accident, we have come across a dataset which has lots of NaN values in its original form. It has 221,525 rows and 37 columns. After inspecting the dataset carefully using the metadata PDF, I have decided that only 6 columns can help us to make a proper prediction and they are - 'HITPARKEDCAR', 'LIGHTCOND', 'ROADCOND', 'WEATHER', 'UNDERINFL', 'SEVERITYCODE'. Out of which, 'SEVERITYCODE' is the target variable.



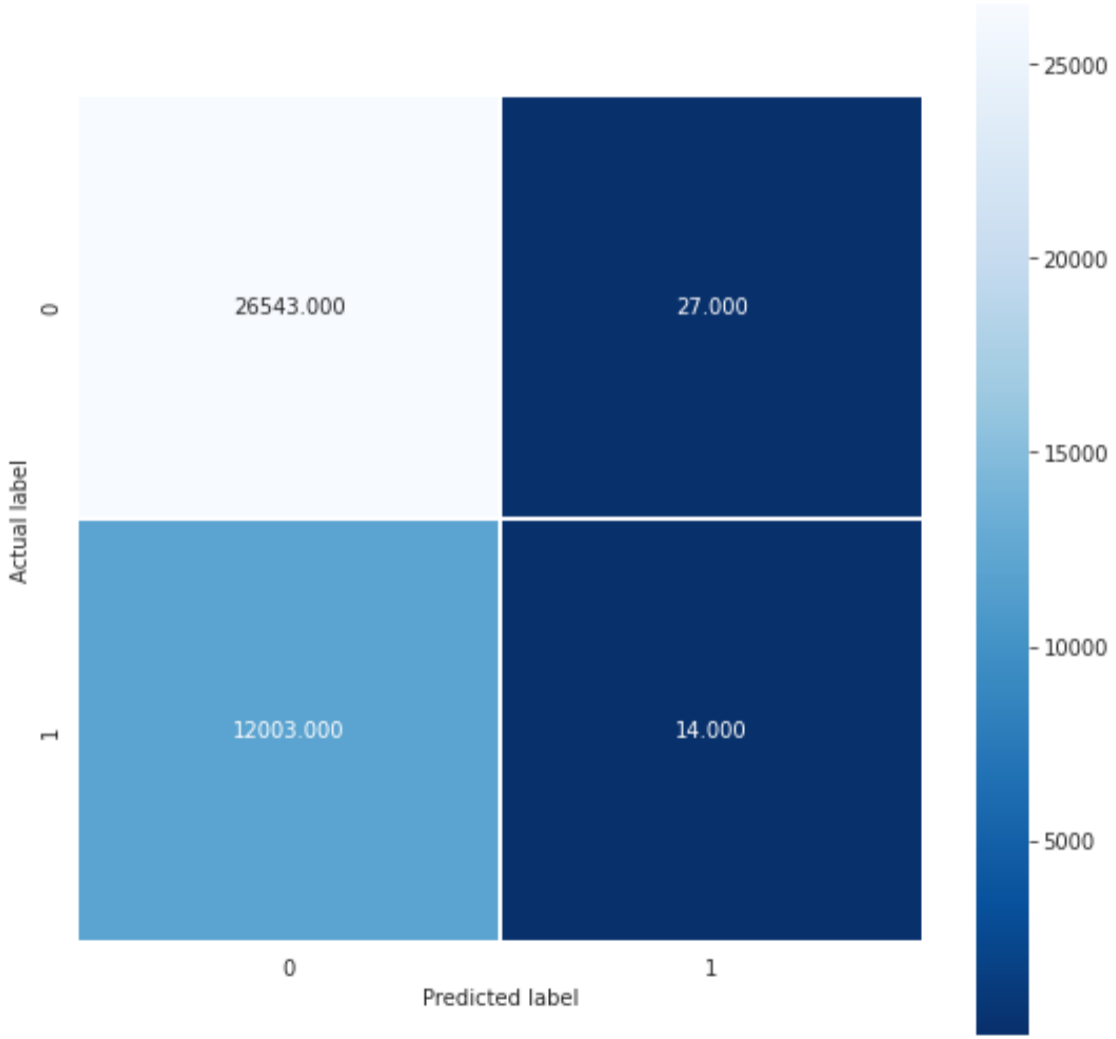
Exploratory Analysis

- In our next visualization, we look at the count of accidents on basis of the time of the day it took place. Majority of the traffic accidents took place at daylight. Less than half of it took place at night, out of which most were in places where there are traffic lights. It rises a question that since most accidents are taking place in well lit areas, does that mean majority accidents are due to not following traffic rules rather than uncontrolled situation?



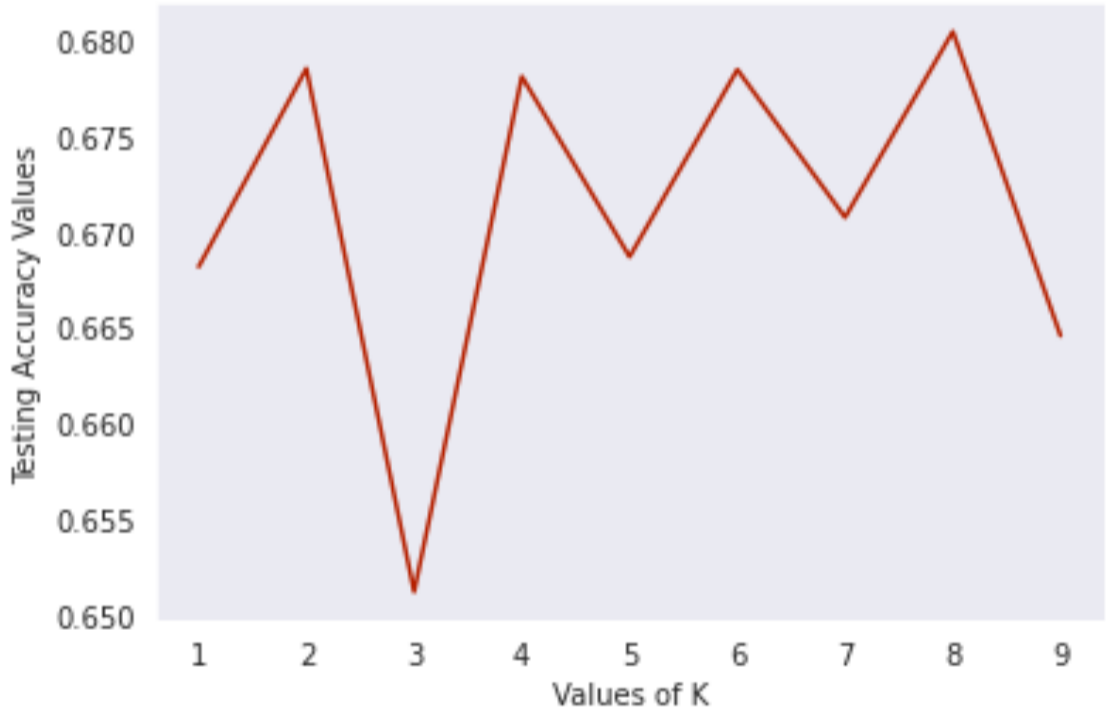
Exploratory Analysis

- Then in the next graph, we look at the weather during the accidents. Majority of the traffic accidents took place on clear days. Only few cases took place in rainy days. This again rises the question, does that mean majority accidents are due to not following traffic rules rather than uncontrolled situation?



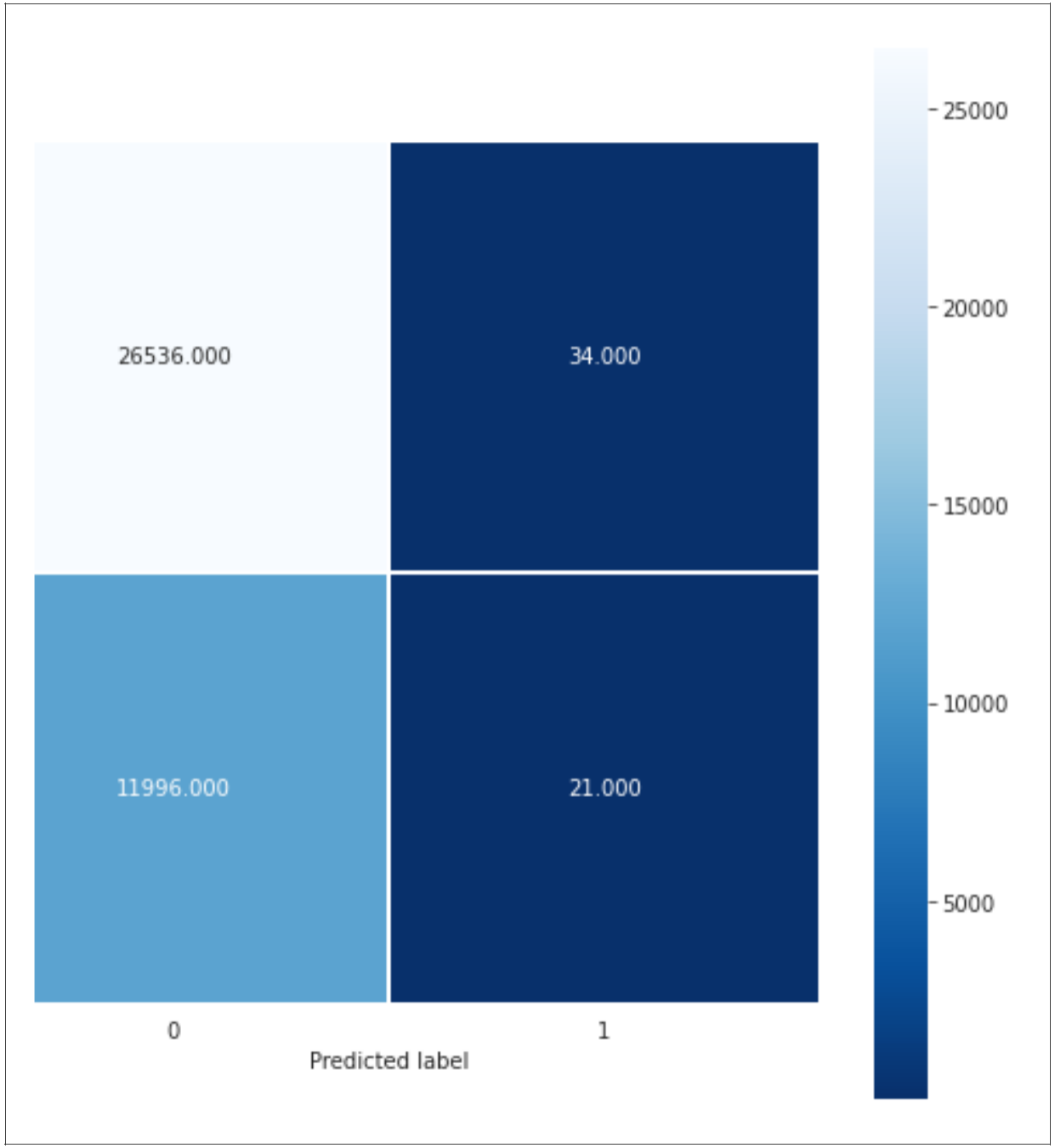
Logistic regression

Jaccard Similarity Score	F1 Score
0.69	0.56



K-Nearest Neighbors

Jaccard Similarity Score	F1 Score
0.68	0.58



Decision Tree

Jaccard Similarity Score	F1 Score
0.69	0.56

Conclusion

- We achieved 68% accuracy using Logistic Regression, KNN, and Decision Tree. However, the performance is not up to the mark. It could have been better if we were able to use more features with more categorized values. Another thing I would say that it is not appropriate to simplify the severity values into binary classes. If we could create a range, then we might have been able to predict severity of traffic cases better using regression models.