Accident Severity Prediction

By

Maniroja M edinburgh

Introduction

- Each year millions of people die in traffic accidents
- If the locations of traffic accidents could be predicted, this could have a huge beneficial impact in potentially helping to reduce their number each year.
- The aim of this project is to predict the severity of road accidents in Seattle, US to help drivers and traffic police department.

Introduction-Contd

- Main cause of accidents and crashes are due to human errors.
- It is also the laxity on part of road users, which cause accidents and crashes.
- They lead to a loss of property and even life.

Different factors of Roads contribute in Accidents

- Drivers
- Pedestrian
- Passengers
- Vehicles
- Road Conditions
- Weather conditions

Preventive measures for accidents

- Education and awareness about road safety
- Strict Enforcement of Law
- Engineering:
  - Vehicle design
  - Road infrastructure

Direct Consequences of Accidents:

- Fatality (Death)
- Injury
- Property Damage

Data Processing

- Jupyter Notebook
- Python Libraries
- Data dropping
  - speeding is an important parameter, we have to drop speeding entirely because it is missing over 180,000 values and this can hamper the results.
- 5 parameters weather, road condition, light condition, Junction type and collision category were considered.

Data Processing -Contd

- Unbalanced severity code
- Down sampling with resample tool
- spliting the data into training data and testing data with a ratio of 80:20
- train data set consists of 93100 samples with 5 parameters and 93100 output labels and the test data consists of 23276 samples with 5 parameters and 23276 output labels.

Methodology

KNN classifier

- KNN can be used for both classification and regression predictive problems.
- We can implement a KNN model by following the below steps:
  - Load the data
  - Initialise the value of k
  - We can import KNN library for implementation

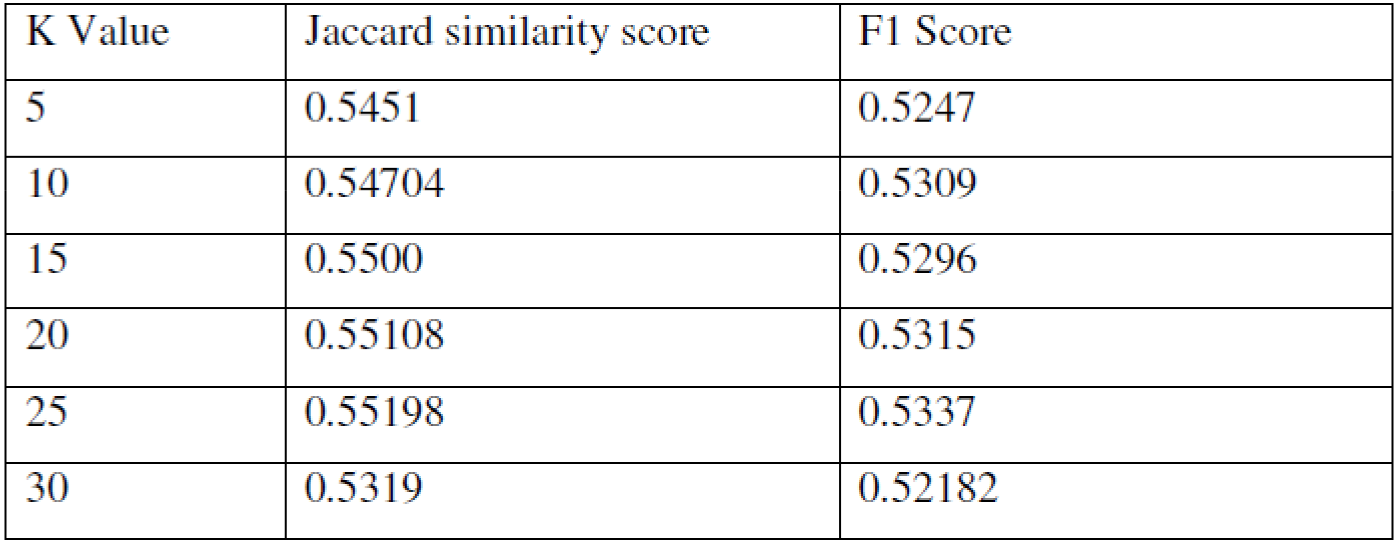
Methodology -Contd

- **Logistic Regression**
  In statics, the logistic model is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick
- Logistic regression is used in various fields, including machine learning, most medical fields, and social sciences

Performance Analysis

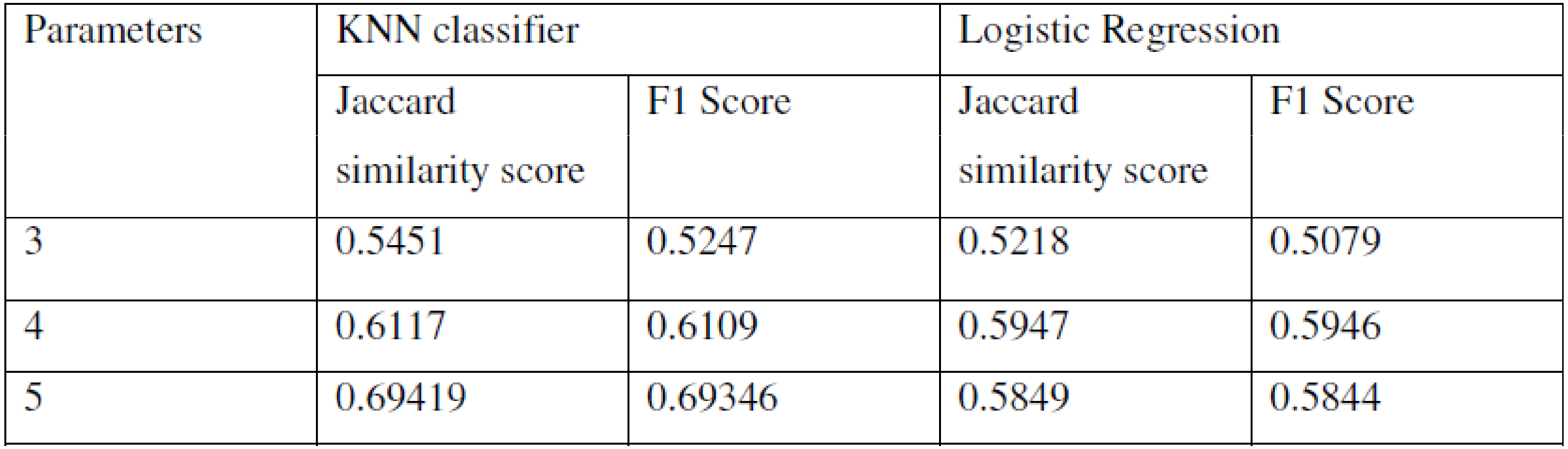
- **Jaccard similarity score**
  - It’s a measure of similarity for the two sets of data, with a range from 0% to 100%.
  - The higher the percentage, the more similar the two populations.
- **F1 score**
  - F_1 score (also F-score or F-measure) is a measure of a test's accuracy
  - It is calculated from the precision and recall of the test

Result Analysis



K Value	Jaccard similarity score	F1 Score
5	0.5451	0.5247
10	0.54704	0.5309
15	0.5500	0.5296
20	0.55108	0.5315
25	0.55198	0.5337
30	0.5319	0.52182

Result Analysis



Parameters	KNN classifier		Logistic Regression	
	Jaccard similarity score	F1 Score	Jaccard similarity score	F1 Score
3	0.5451	0.5247	0.5218	0.5079
4	0.6117	0.6109	0.5947	0.5946
5	0.69419	0.69346	0.5849	0.5844

Discussion

- Even though our data was a good size, there were a number of missing elements and we needed to clean the data in order to get a good result.
- When weather conditions are bad at the junction intersection point, this model can alert drivers to remind them to be more careful.

Conclusion

- Lot of these accidents are minor and avoidable. These findings can be helpful to the Seattle Police Department in enforcing some new measures to prevent future accidents.

Future Work

- Data size can be increased.
- Latest Data can be considered
- Multiple models like decision tree could be trained and then compared.
- More conditions can be included to train the model