Accident Severity Report

Introduction

Currently there are many accidents occurring due to various factors which cause a loss of property and life itself. This report is produced with an objective to reduce such occurrences by providing the stakeholders with prediction and likelihood of accident given the environmental factors such as weather and lighting. This will help deter accidents which can be avoided if correct information is available at the correct time. This would be helpful to the police department, road authorities and mainly drivers.

Data

Data is used from Seattle's City police department from 2004 to 2020 with over 194,000 records. The main fields which will be observed to achieve our objective would be the location of the accident, severity and the lighting conditions. This will help us predict a severity code which could indicate the level of danger of traveling at a given time. The code ranges from 0 to 4 with following explanation :

0: Little to no Probability (Clear Conditions)
1: Very Low Probability — Chance or Property Damage
2: Low Probability — Chance of Injury
3: Mild Probability — Chance of Serious Injury
4: High Probability — Chance of Fatality

Methodology

To start with, I downloaded the csv file available in the portal. Data cleansing was vital as I deleted out the unwanted columns which would reduce the overall data size and increase the relevancy of. Data in columns was edited to fit the format. Records with missing fields were deleted as these may tamper with the results reduce the accurateness. Jupyter notebook using Python was used to carry out data analysis with packages such as Pandas, Numpy and Sklearn. We used different machine learning models such as K-nearest neighbor, Decision tree & Logistic regression.

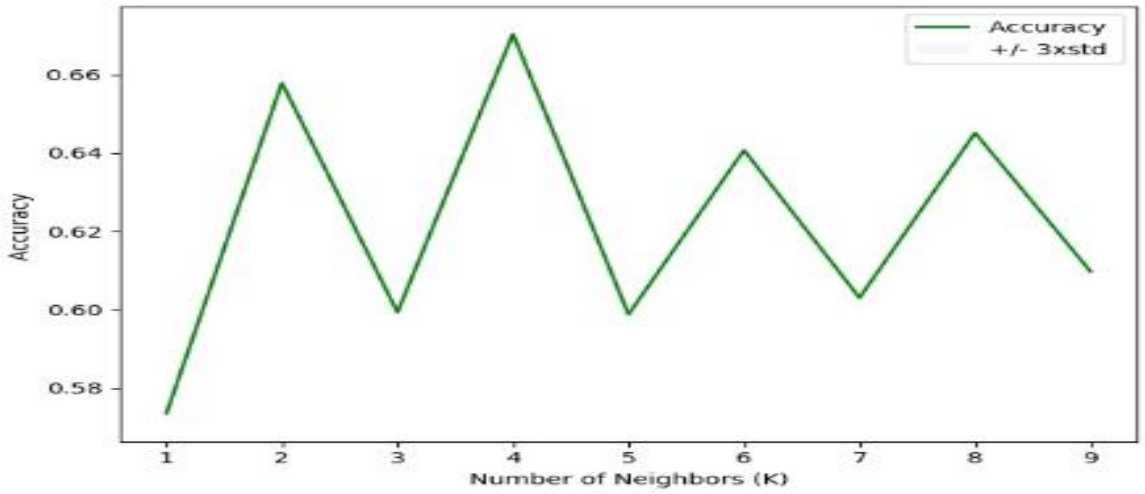
- **k-Nearest Neighbor:** K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (based on distance)

- **Decision Tree Analysis:** The Decision Tree Analysis breaks down a data set into smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.
- **Logistic Regression:** Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable

Results

1) KNN

The best K appears to be 4 with below results



	Precision	Recall	f1-score
0	0.93	0.70	0.80
1	0.08	0.32	0.13
Accuracy	0.67		
Macro Avg	0.50	0.51	0.46
Weighted Avg	0.86	0.67	0.75

2) Decision Tree

	Precision	Recall	f1-score
0	0.64	0.72	0.68
1	0.44	0.34	0.39
Accuracy	0.58		
Macro Avg	0.54	0.53	0.53
Weighted Avg	0.56	0.58	0.56

3) Logistic regression

	Precision	Recall	f1-score
0	0.72	0.67	0.69
1	0.35	0.41	0.38
Accuracy	0.59		
Macro Avg	0.53	0.54	0.53
Weighted Avg	0.61	0.59	0.60
Log Loss	0.68		

Conclusion

When comparing all the models for accuracy we can see that the f1-score is highest for k-Nearest Neighbor at *0.75*. However, when we look at the precision and recall we observe that the k-Nearest Neighbor model performs poorly in the precision as well as the recall.

When looking at decision tree and logistic regression, it is observed that the decision tree has better whereas, the logistic regression is more balanced for recall. The average f1-score for these are similar.

It can be concluded that the both the models can be used side by side for the best performance.

Algorithm	Average f1-Score	Property Damage (0) vs Injury (1)	Precision	Recall
Decision Tree	0.56	0	0.64	0.72
		1	0.44	0.34
Logistic Regression	0.60	0	0.72	0.67
		1	0.35	0.41
k-Nearest Neighbor	0.75	0	0.93	0.70
		1	0.08	0.32

Conclusion

It is to be noted that results are derived from the data available. Hence the accuracy of the results are directly related to the data used.

In future, such reports should be evaluated regularly to help deter the increase in accidents and help save loss of property and life.