Final Capstone Project

Table of content

1. Introduction

Car accidents are happending all the time in the world. According to WSDOT's (2017) data, a car accident occurs every 4 minutes and a person dies in a car crash every 20 hours in the state of Washington, U.S.A. To help with reduction of car accident cases, this project trys analysing the determinants of an accident and sheds light on predicting the severity with those factors.

2. Data

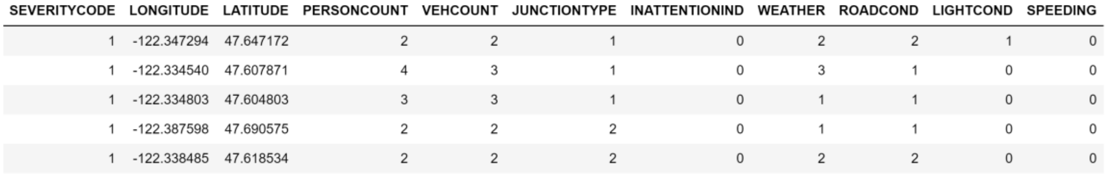
The data of car accidents which have occurred within the city of Seattle, Washington from the year 2004 to 2020 was used. This data is regarding the severity of each car accidents along with the time and conditions under which each accident occurred. The model aims to predict the severity of an accident with other information provided. All useful features were extracted and and the missing values were handled at first, followed by a creation of a balanced dataset with equal number of two severity type cases. Lately classfication methods such as KNN, random forest and decision tree were used.

In the data processing step, the X and Y variables were first renamed to LONGITUDE and LATITUDE for clarification. Columns such as OBJECTID, INCKEY, COLDETKEY, REPORTNO, INTKEY, EXCEPTRSNCODE, SDOT_COLCODE, SDOTCOLNUM, ST_COLCODE, ST_COLDESC, SEGLANEKEY and CROSSWALKKEY were code-based were dropped as they are meaningless. Columns such as STATUS, EXCEPTRSNDESC, INCDATE, INCDTTM, SDOT_COLDESC, PEDROWNOTGRNT, ST_COLDESC, UNDERINFL, PEDCYLCOUNT, HITPARKEDCAR, SEVERITYDESC and ADDRTYPE were dropped as they did not help for the analysis. The column of SEVERITYCODE.1 were dropped as it is a duplicate of SEVERITYCODE.

In the process of dealing with missing values, the rows of missing a WEATHER condition were dropped as weather is an important factor in car accidents, so we can’t simply fill in randomly chosen value. For LONGITUDES and LATITUDES, missing values were filled by mean values. For SPEED and INATTENTIONIND, “N” (negative) were filled in. For other factors such as light condition and road condition, etc, “Unknown” were filled instead.

After all, data were balanced by resample function and encoded with numerical substitutes for categorical string values.

Table 1 below shows the first 5 rows of the data after processing



SEVERITYCODE	LONGITUDE	LATITUDE	PERSONCOUNT	VEHCOUNT	JUNCTIONTYPE	INATTENTIONIND	WEATHER	ROADCOND	LIGHTCOND	SPEEDING
1	-122.347294	47.647172	2	2	1	0	2	2	1	0
1	-122.334540	47.607871	4	3	1	0	3	1	0	0
1	-122.334803	47.604803	3	3	1	0	1	1	0	0
1	-122.387598	47.690575	2	2	2	0	1	1	0	0
1	-122.338485	47.618534	2	2	2	0	2	2	0	0

Table 1: First 5 rows of processed data

3. Methodology

a. Exploratory data analysis

(1) Mapping

Figure 1 below shows the map of all accidents’ location

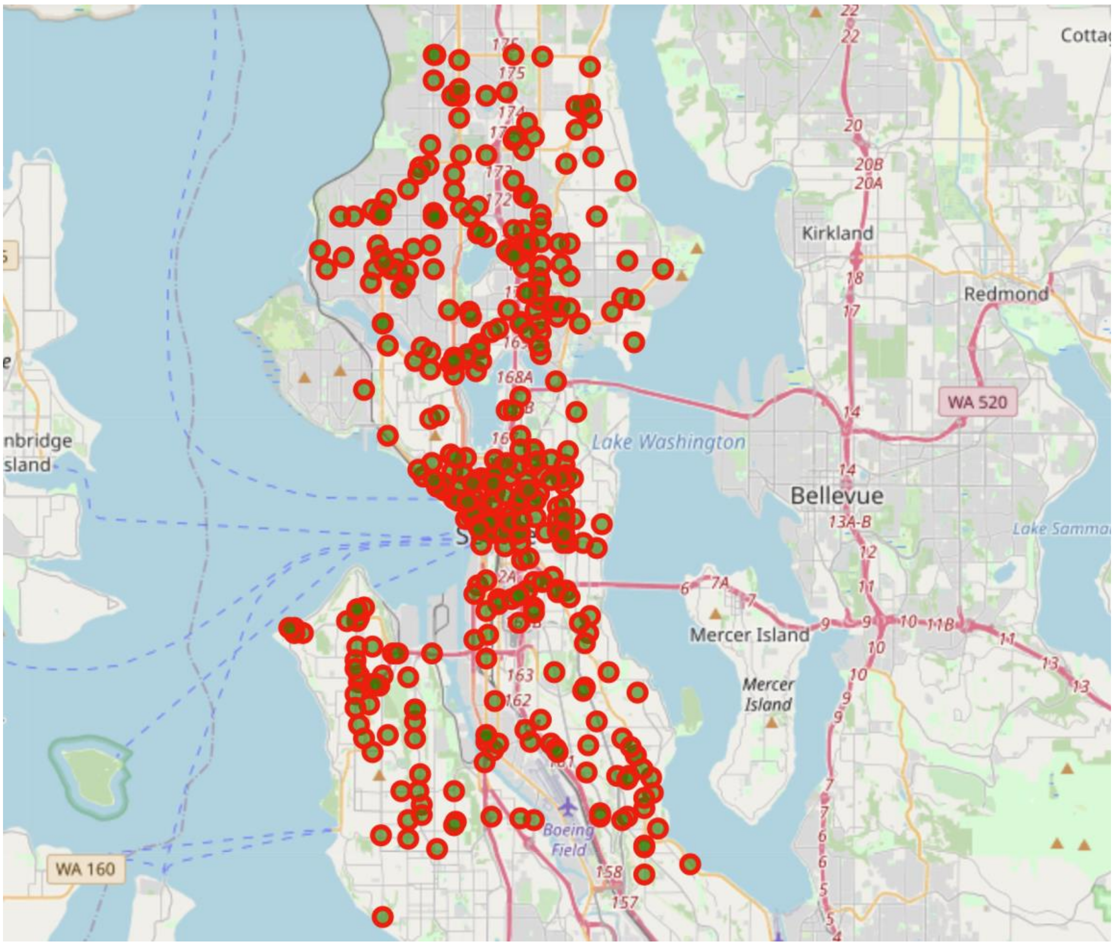


Figure 1: Locations of accidents

(2) Visualizing the types of each feature

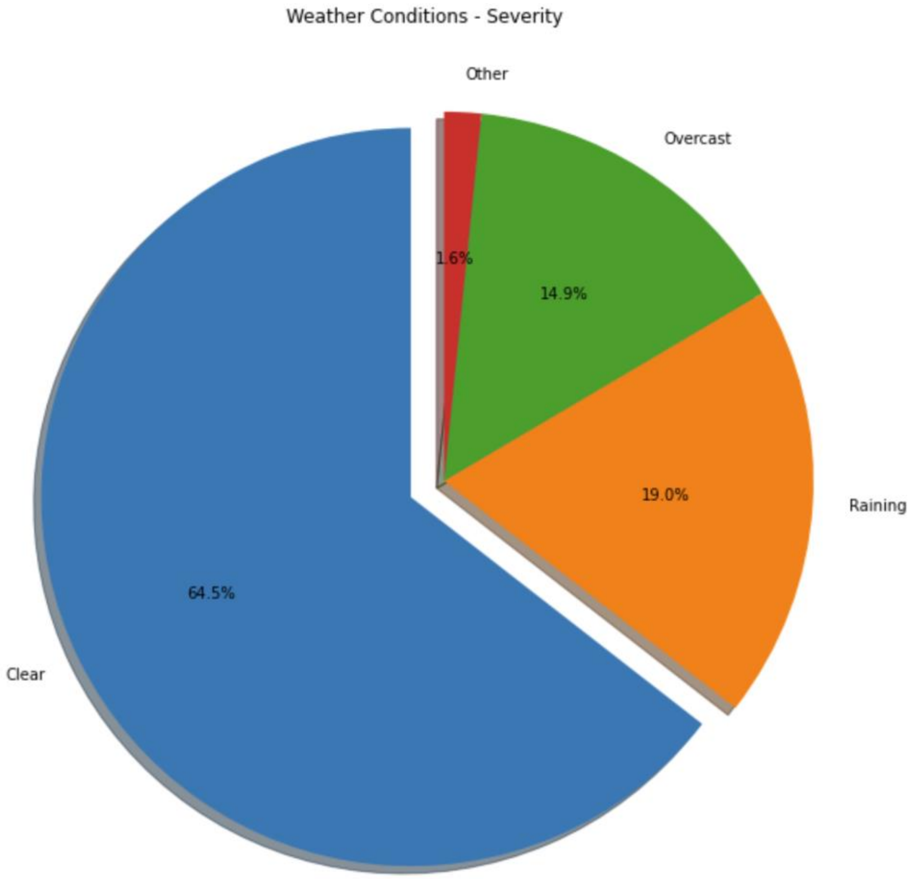


Figure 2: Types of weather conditions

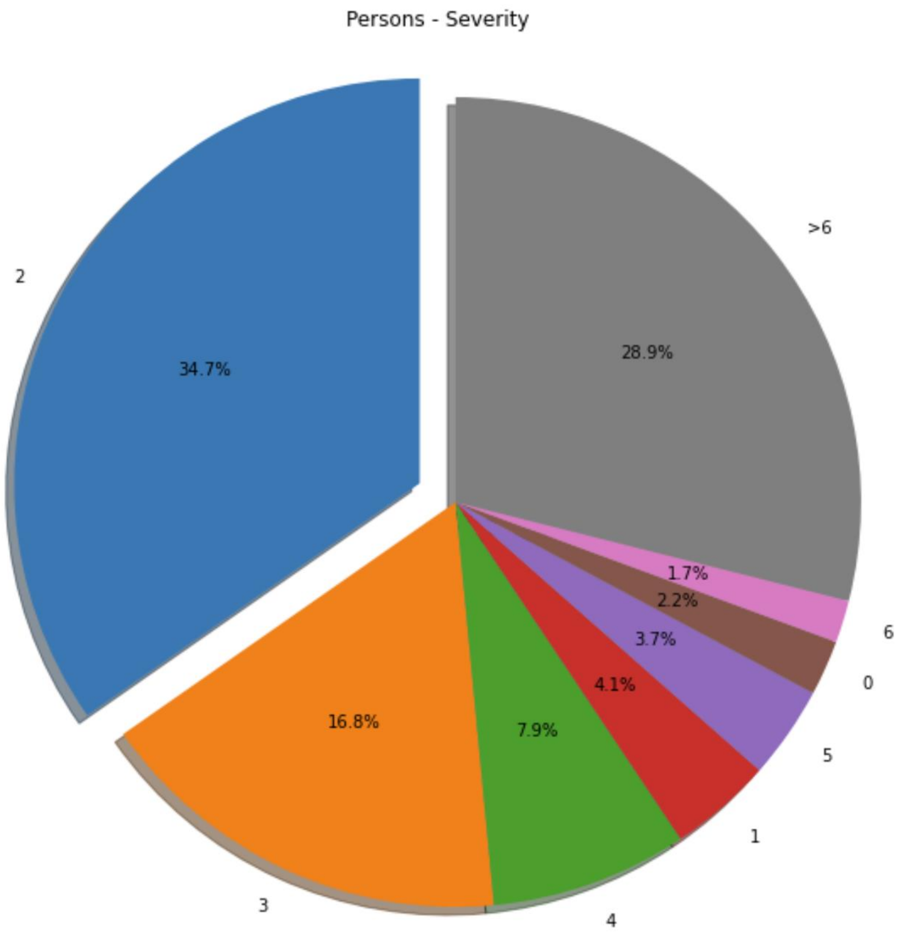


Figure 3: Types of people

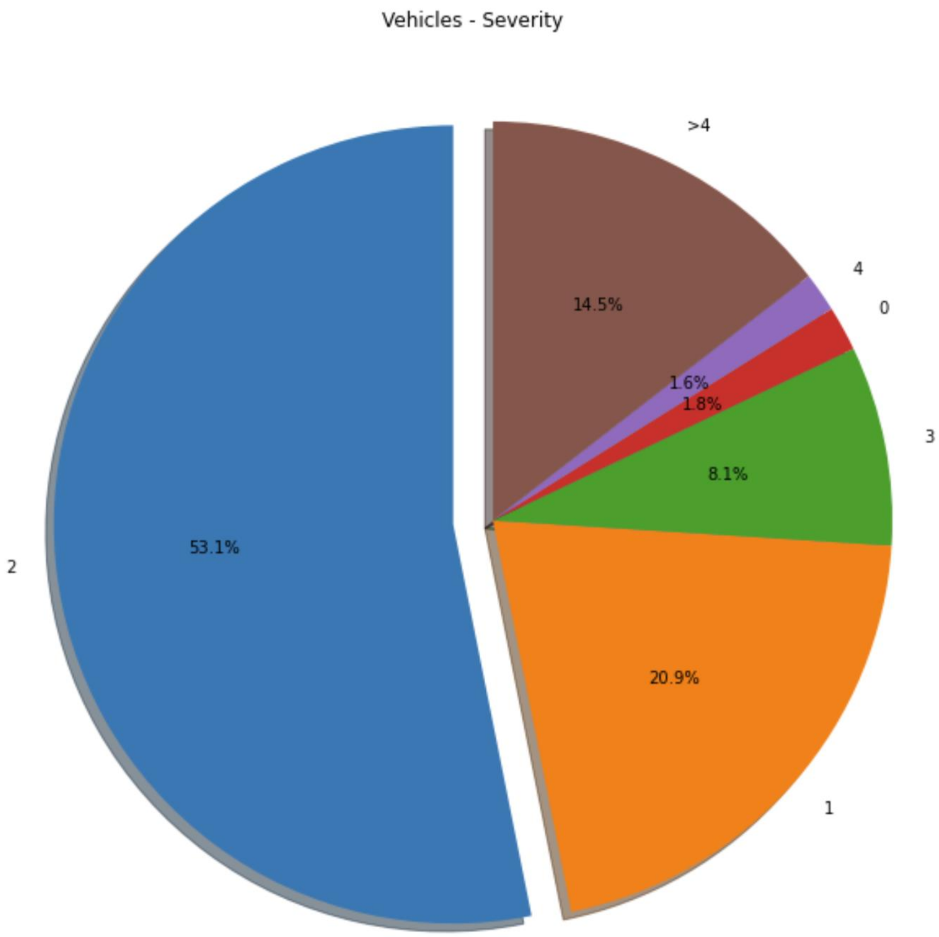


Figure 4: Types of vehicles

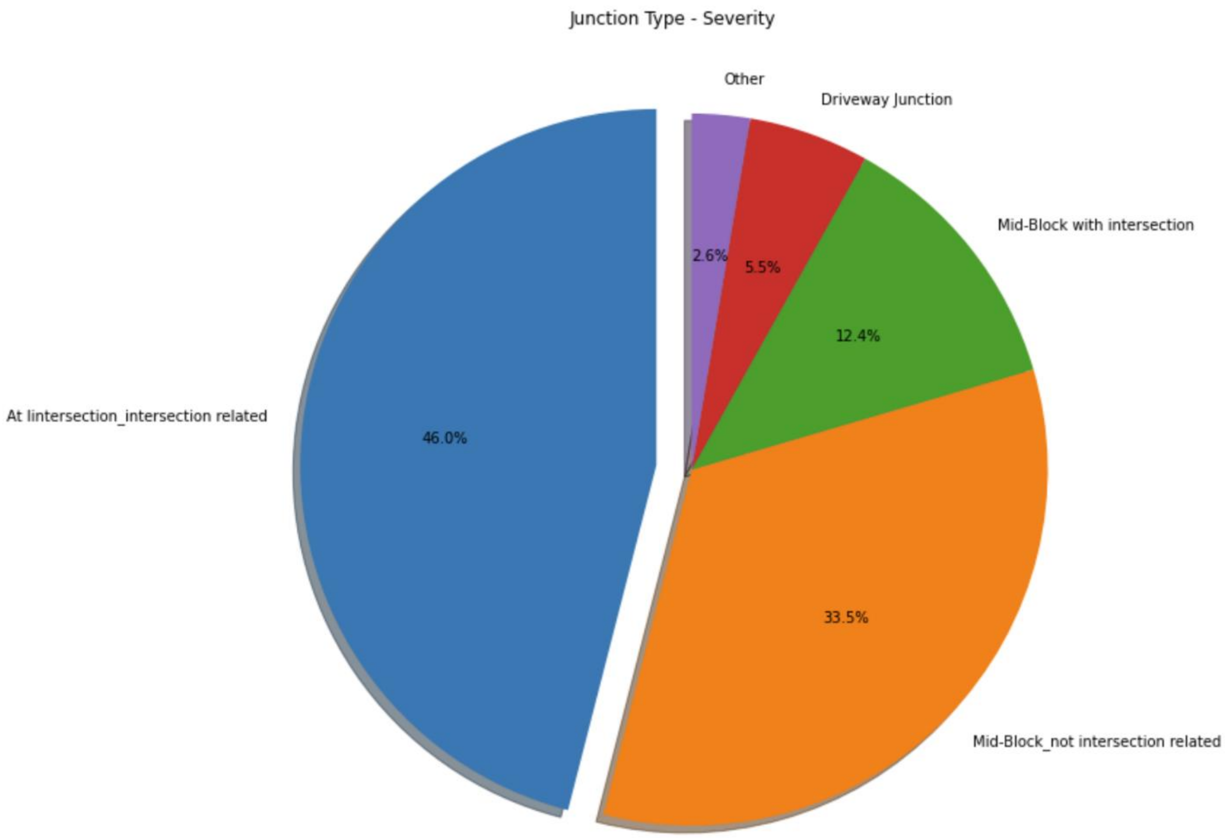


Figure 5: Types of junctions

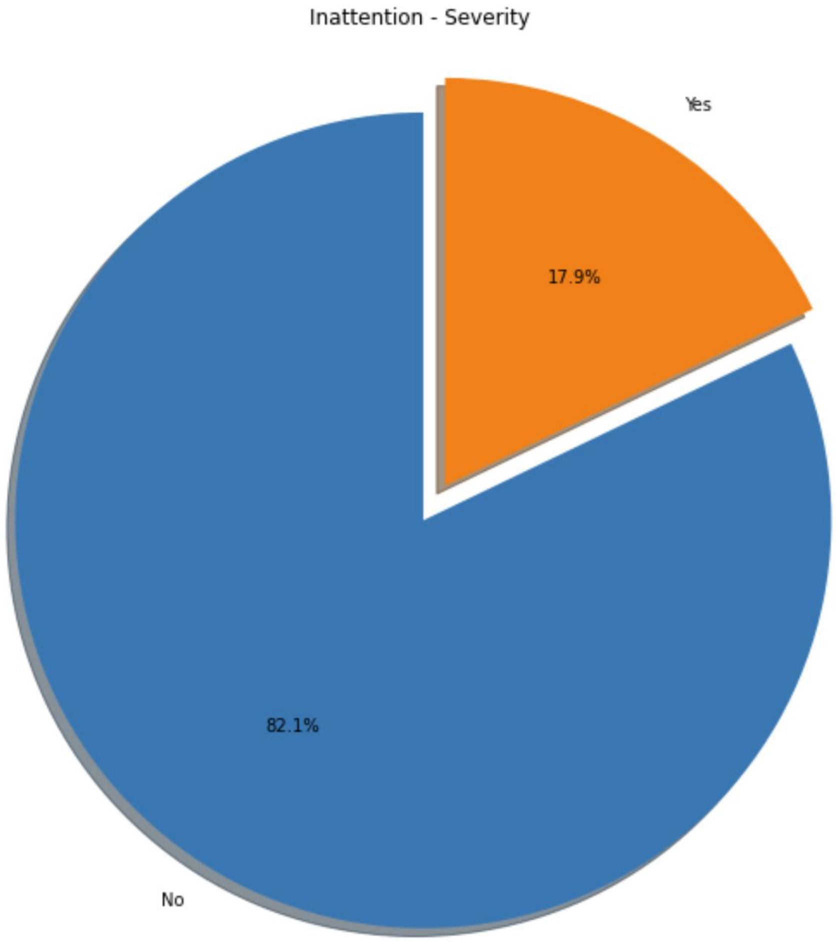


Figure 6: Whether is inattentioned

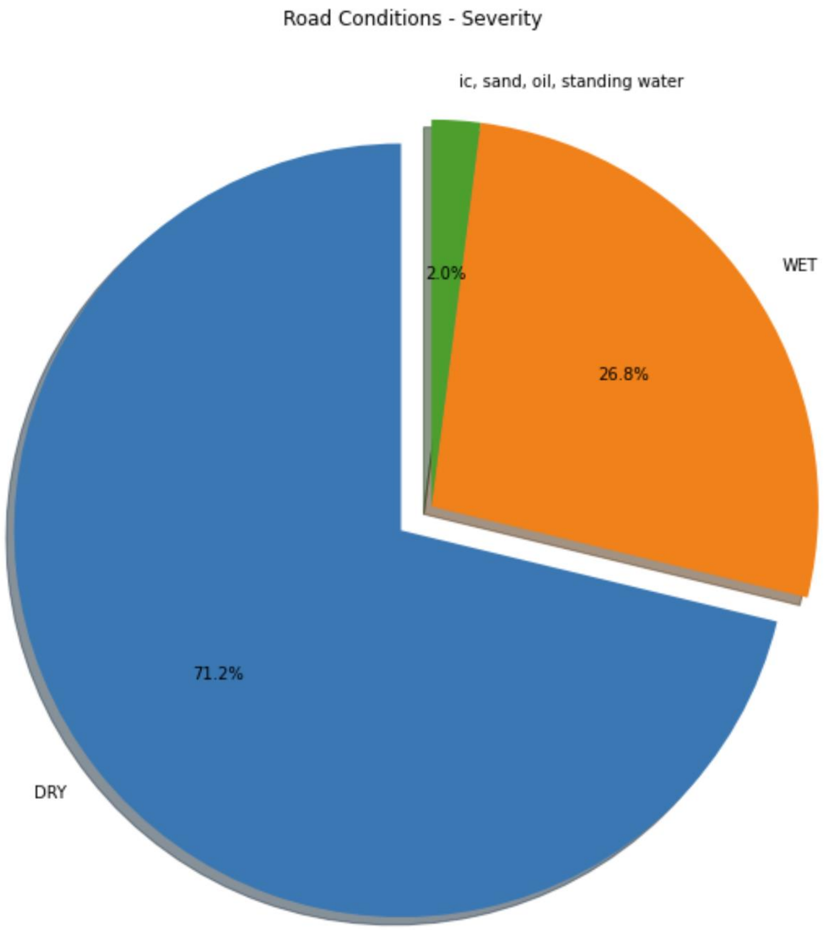


Figure 7: Road conditions

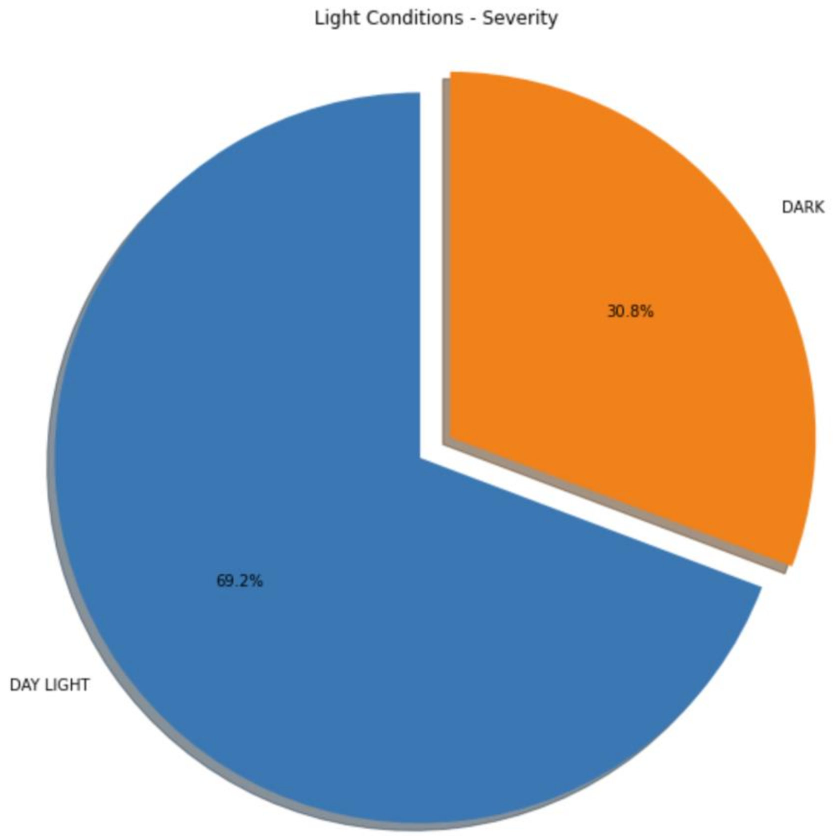


Figure 8: Light conditions

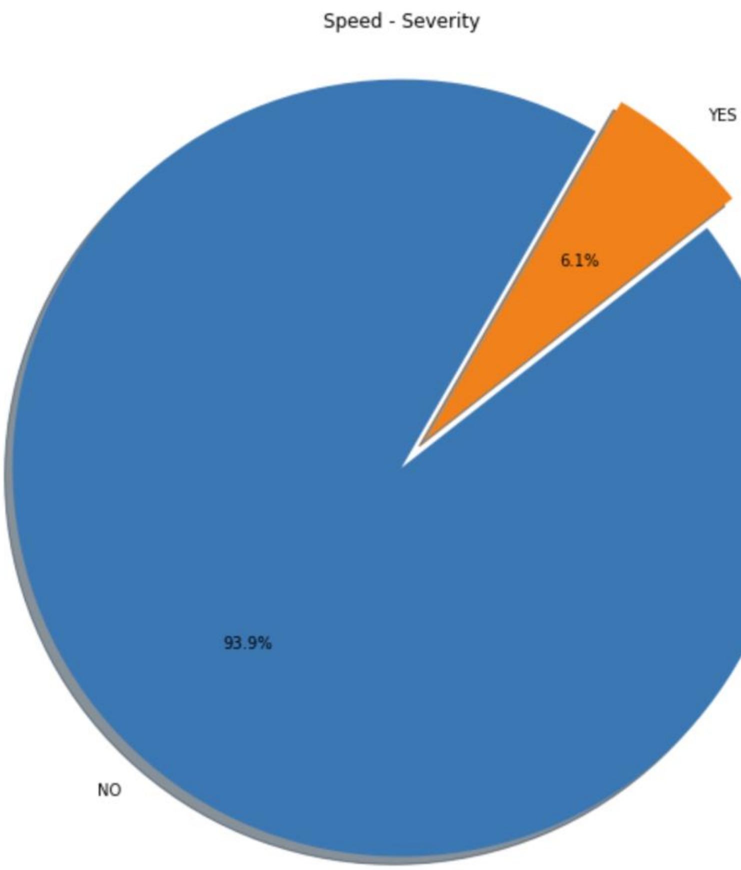


Figure 9: Whether is overspeed

b. Model development

(1) K-Nearest Neighbours
From table 2 below, k with value 1 was selected for K-NN model.

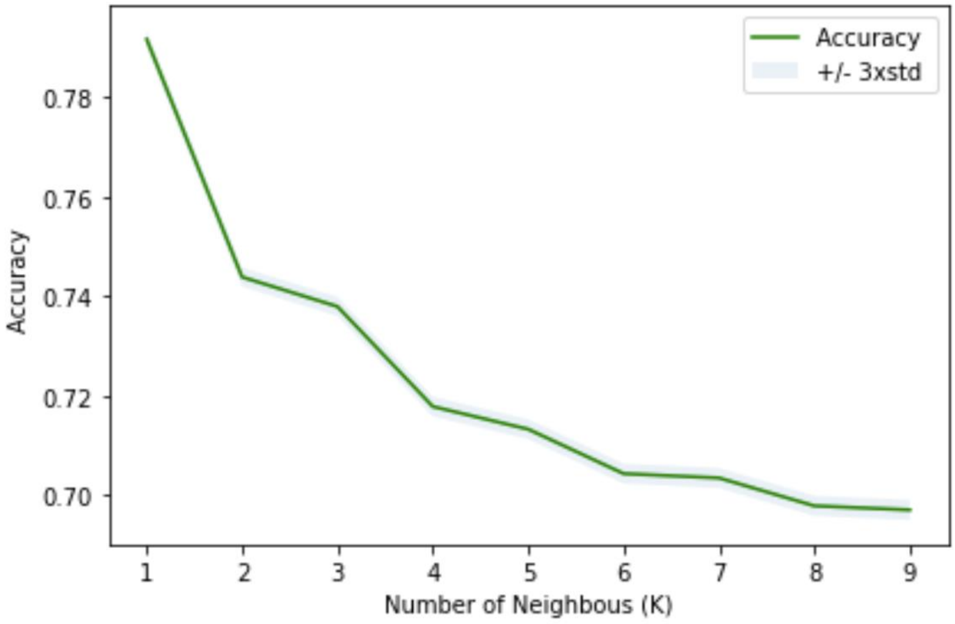
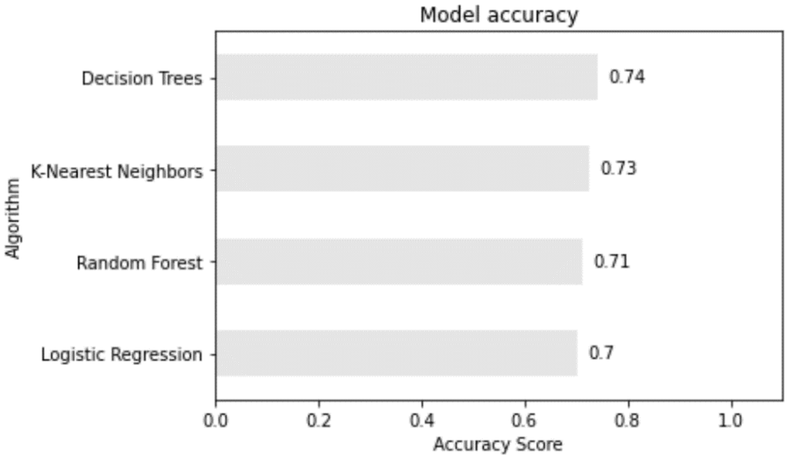


Table 2: Accuracy vs Number of Neighbours

4. Results

Decision tree won the highest accuracy score with 0.74l among all model been tested in the prediction of car accident severity, followed by the K-NN (0.73), random forest (0.71) and logistic regression (0.7).



5. Discussion

The accuracy scores of the four methods used in our test are very close. The score range [0.7, 0.74] indicates that our model neither did an excellent job nor a very bad job. In other words, we can even increase our prediction accuracy to over 80% by further amendment. In our case, missing values occupy a lot, so they might have affected our accuracy.

6. Conclusion

Decision tree is the most accurate model among all model been tested in the prediction of car accident severity.

Future work will be feeding more data into the dataset to increase model accuracy. In our case, age was not considered but it could impact on driving and car accident.