An investigation into weather effects on road accident likelihood and severity in Seattle

Rosany Antonyvincent

Data Science Professional Specialisation
IBM

September 2020

Table of contents

Chapter 1

Introduction

1.1 Background

Seattle is the largest city in the state of Washington and consists of roughly 710000 residents[1][2]. The best way to travel within the city is by car, and therefore evaluating the likelihood and severity of road accidents is of great importance. Not only for the safety of the civilians, but also for the economy. Reducing road accidents, improves the safety of the residents, reduces potential time off work for busy commuters and less demand for road repairs. This will allows for less demand for paid sick leave and for employees to continue working[3][4]. The most common job type in the area, as of 2018, is a software developer which receives an average income of $91000[5]. This overtook retail salesperson in 2016, which had an average salary of $28000[6]. This jump in average wage, shows the necessity for preserving and furthering the economy.

1.2 Problem

Although across the years the number of car accidents have been decreasing across Seattle, and across Washington, the number of vehicles involved in fatal accidents are slowly increasing as are the fatalities[7]. This could be due to drastic weather patterns[8]. Equally the number of pedestrians involved in fatal accidents are increasing. The most likely time of accident takes place on Friday afternoons.

The aim of this project is to analyse the likelihood and the severity of a road accident at different times across the year and predict whether the weather[9] is a significant contributor to the incidents.

Chapter 2

Data acquisition and cleaning

2.1 Data sources

The labelled data used will be the Data Collision records in Seattle across all years provided by SeattleGeoData[10]. Severity will be the predominant dependent variable, with varying attributes. Attributes include location, collision type, incident time and date, weather, road conditions and light conditions. To this end, all other attributes will be discarded. significance to the time and date of accidents and fatality. This will be done using histograms and classification techniques. It will also look at whether the weather has a significant impact on accidents and to what extent.

If a correlation is established, the focus will then be placed on degree of severity. The attributes will then be converted to binary form, which allows for test/train predictions. The accuracy of the test/train prediction will be qualified using f1-score and Jaccuard index.

2.2 Cleaning for Analysis

The first quality to check is to make sure there is at least one attribute to use as a universal identifier. This attribute needs to comprise of unique values of length comparable to the amount of rows in the dataframe. a Boolean check was performed on OBJECTID to verify this. The other unique attributes (INCKEY, COLDETKEY, REPORTNO, INTKEY) can be redacted from the dataframe as some have missing values or have different types of formats within the column.

Since this report will focus predominantly on the date and time of the collision, the date and time format will need to be altered. The dataframe itself has fairly inconsistent values for these columns.

The datetime (INCDTTM) attribute was split into two separate columns, and then put into d/m/Y and H/M/S string format. The year could then be easily extracted by taking the last four letters. The date (INCDATE) attribute was cleaned by using the pandas function strftime to split the date into day, month, year and day of year. A 00:00 time appears for cells with date and no time after applying string format to INCDTTM. This was created by first checking how many cells had incidents at midnight (this turned out to be zero), then 00:00 was replaced with an undefined value (NaN). For the purpose of consistent dataframes, the rows with NaN values were removed. This was around 2% of the dataset. The hour was extracted from the time the same way the year was extracted from date.

2.3 Cleaning for Modeling

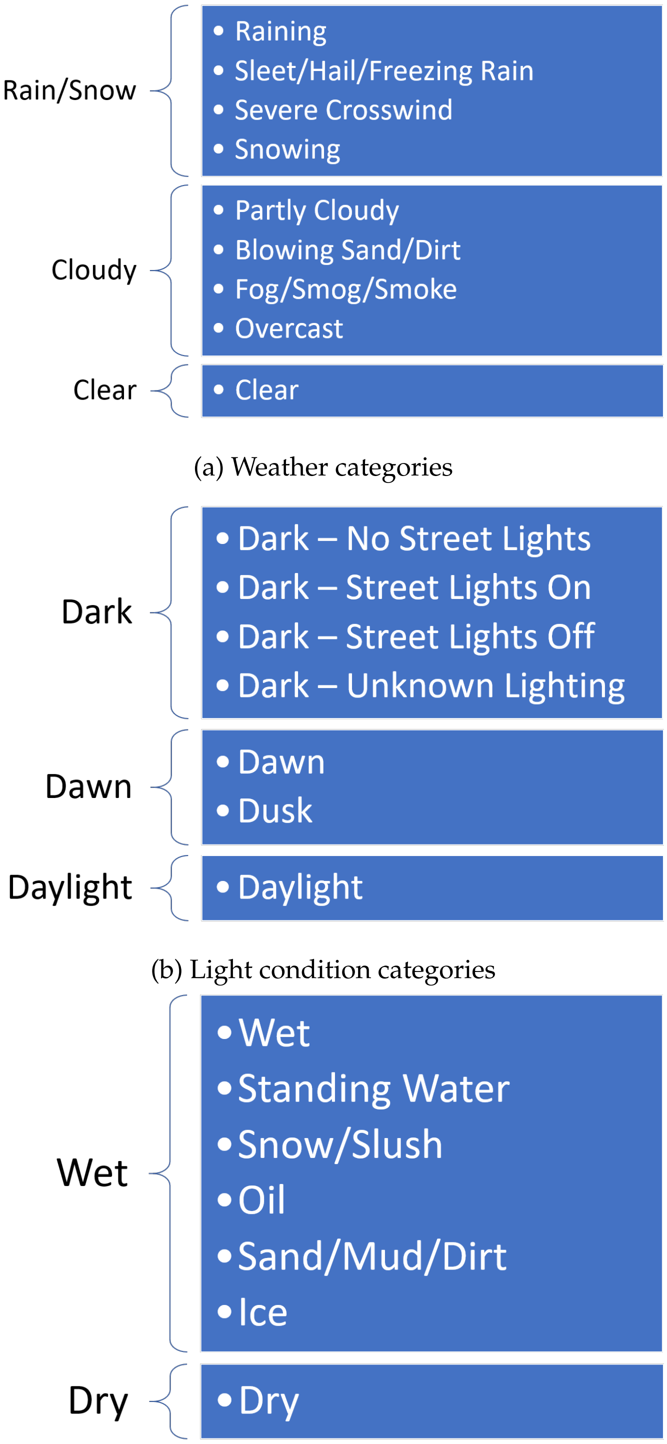
2.3.1 One Hot Encoding

Before using classification methods to predict the outcome of severity, the attributes were organised into numerical form. The attributes being:

- Weather
- Light condition
- Road condition
- Day

Firstly, the a new dataframe is created merging day number and month number from the previous date dataframe. All rows with 'Unknown' or 'Other' values were removed to keep all data entries consistent. This resulted in 7% drop in data. Again, 'OBJECTID' was kept as a consistent attribute.

Many of the attributes have multiple unique values, some of which are repeated. In order to employ the best possible outcome of severity but to also keep runtime fairly low, these values have been grouped as follows 2.1. One hot encoding will take these categorical variables and transform them into numerical ones under binary variables.



(a) Weather categories

(b) Light condition categories

(c) Road condition categories

Fig. 2.1 Figure to show the different categories for predictive modeling

Chapter 3

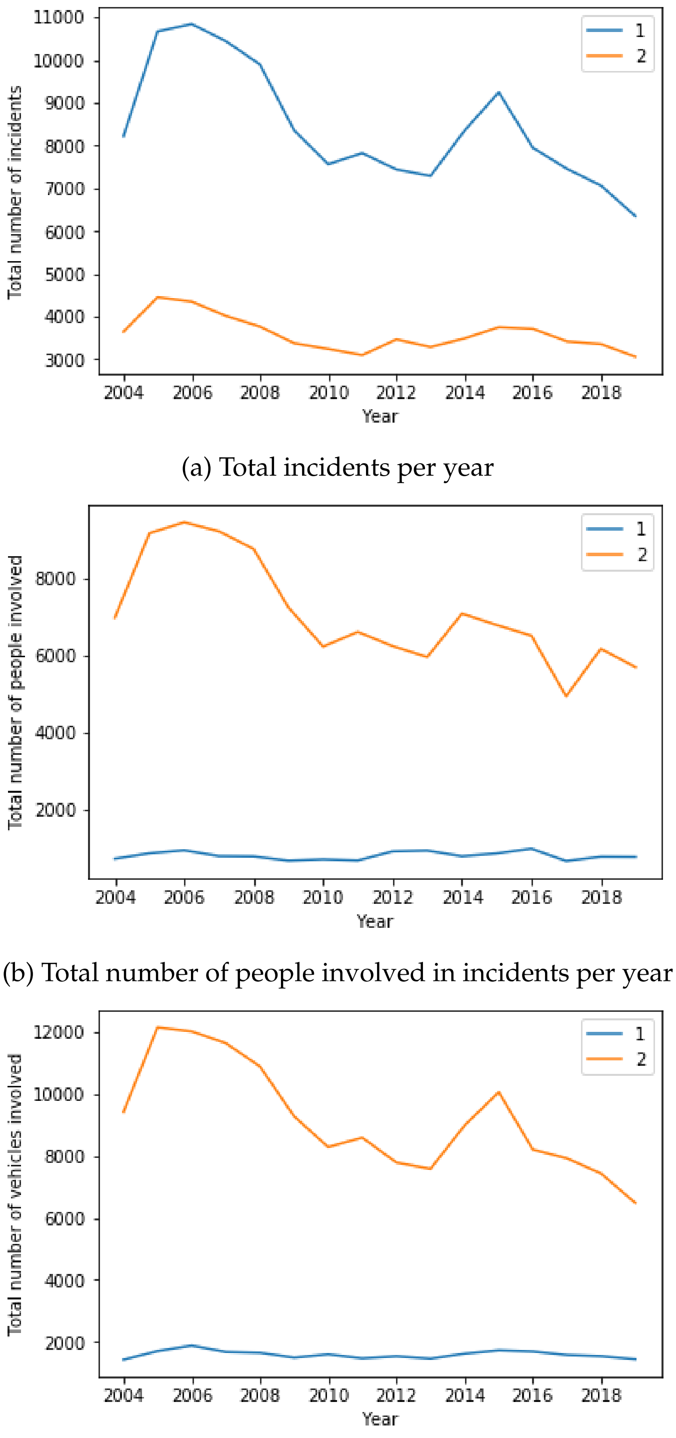
Exploratory Data Analysis

The first comparison made is the severity against the new date attributes. The severity code obeys the following,

- 3 - Fatalilty
- 2b - Serious injury
- 2 - Injury
- 1 - Prop damage
- 0 - Unknown

The column was analysed to see the count of each severity. The only severity codes within the dateset are 1 and 2. This allowed for optimising the function of the code and improve the runtime.

3.0.1 Total per year



(a) Total incidents per year

(b) Total number of people involved in incidents per year

(c) Total number of vehicles involved in incidents per year

Fig. 3.1 Figure to show total incidents, people involved and vehicles involved per year.

First, let's look at the total number of incidents across the years. This can be seen in 3.1a. This shows a peak at 2006 but then continues to decrease beyond this point. The number of incidents have effectively halved across the 16 years (neglecting the peak at 2015). This figure also shows that incidents including injury are a lot less likely than those of property damage. Suggesting that injuries are very minimal in Seattle, and decreasing.

Equally, the total amount of vehicles involved peaked at the same time (as expected), and is also decreasing. This will be important to note, as the weather conditions will affect the amount of vehicles damaged. An interesting observation is the amount of vehicles involved in incidents is more than that of people involved, suggesting there are a noticeable amount of stationary, empty vehicles involved in collisions. An example being a car driving at night in January on black ice, could swerve and hit multiple parked cars.

3.0.2 Total per month

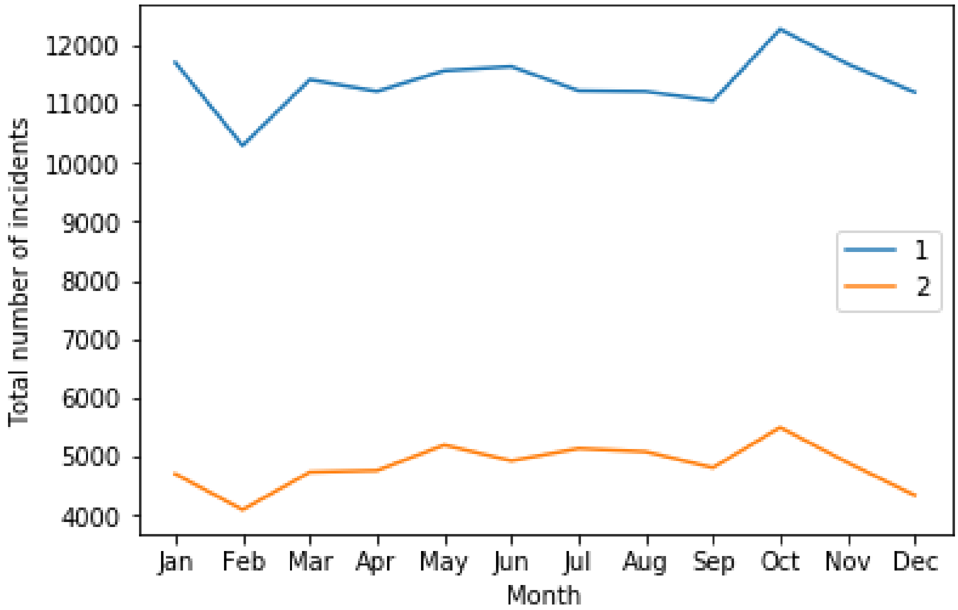


Fig. 3.2 Graph showing incidents across each month

Having analysed the total amounts across the year, it's important to also analyse incidents across the months. From 3.2 we can see that the amount of incidents reported are fairly consistent across the months. The drop in incidents occurs in February, potentially due to fewer drivers on the road. A spike in property damage incidents occurs in October, which could be due to weather, similar to January. Otherwise, in the Spring/Summer/Autumn months, the number of incidents are consistent.

3.0.3 Total per day

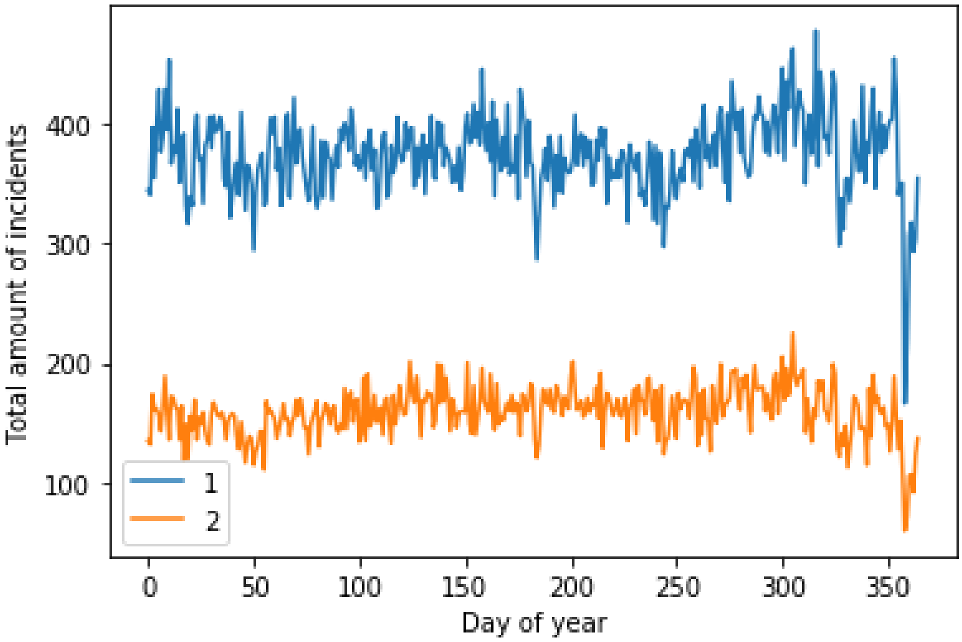
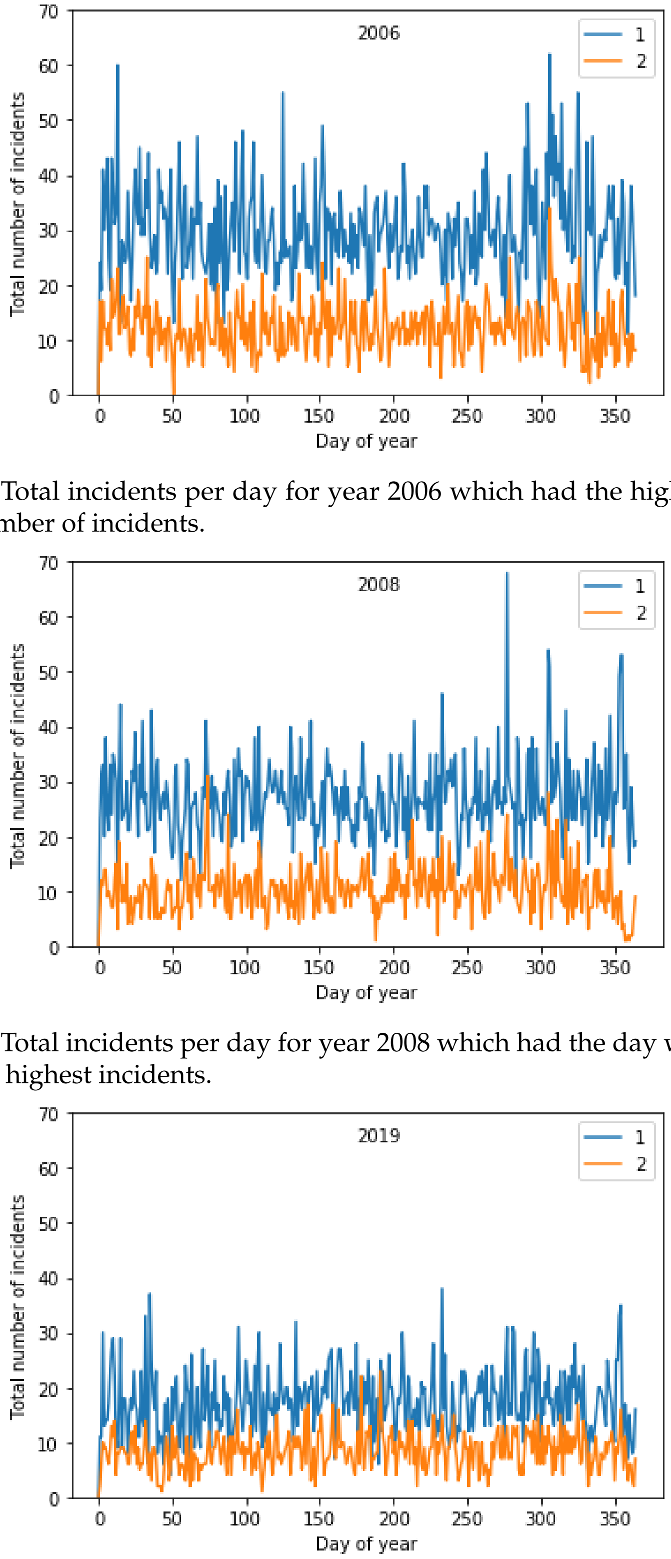


Fig. 3.3 Graph showing incidents across each day of the year

As can be seen in 3.3, property damage incidents occur twice as often as injuries. What is noticeable is that Christmas day has the least amount of incidents, which is expected as far fewer people travel on that day. It is noticeable also that days with a drop in incidents occur around public holidays [11].

What is noticeable across 3.4 is that although the number of incidents involving property damage reduces across the year, the amount of injuries stays fairly consistent. This could be due to people being more careful on the road to avoid cars being damaged. 2008 had the highest number of incidents per day as shown in 3.4b. What was seen in the data however that after this period, the incidents decreased dramatically. 3.4c has dropped to have more consistent amounts of incidents.



(a) Total incidents per day for year 2006 which had the highest number of incidents.

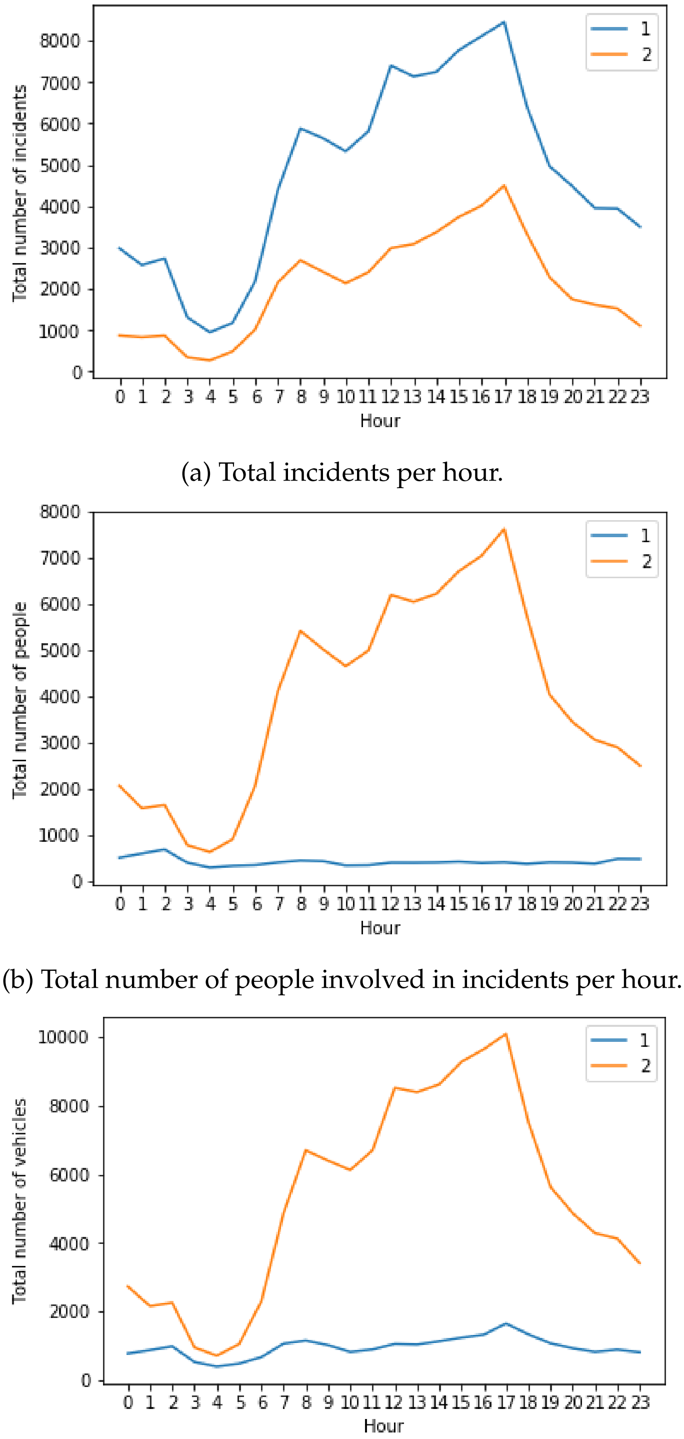
(b) Total incidents per day for year 2008 which had the day with the highest incidents.

(c) Total incidents per day for year 2019 which had the lowest incidents.

Fig. 3.4 Figure to show the total incidents across different years.

3.1 Total per hour

Finally, let's look at the total number of incidents across time. As can be seen in 3.5a, the amount of incidents drops between 2am and 6am. This is expected as most people will be sleeping during these hours. The incidents peak at 5pm, which would be rush hour. This suggests that travelling later in the day could reduce the likelihood of an incident occurring. In 3.5b and 3.5c, property damage has no correlation to time. Otherwise, these graphs suggest that there is a correlation in number of incidents and number of people involved at peak times.



(a) Total incidents per hour.

(b) Total number of people involved in incidents per hour.

(c) Total number of vehicles involved in incidents per hour.

Fig. 3.5 Figure to show the total incidents across different hours.

Chapter 4

Prediction Modeling

As stated in chapter 2, in order to use classification modelling, the attributes to split into fewer categories (ex. dry/wet). With the stated attributes being split, the dataframe consisted of 10 columns. This required a long runtime, especially for the decision tree. To that end, multiple accuracy values were drawn for different sets of features, predicting the severity of accidents in each case:

- Weather binary variables
- Weather and light condition binary variables
- Weather, light condition and road condition binary variables

The severity of an accident could take two values, either a 1 (less severe) or a 2 (more severe). In this sense it is a binary classification problem. Importantly, there was a large majority of data labelled with a 1 (75%).

All data used is labelled, so the data was split into a test/train model of ratio 1:5. The models I chose were k-Nearest Neighbours, Decision Tree and Support Vector Machine.

The accuracy was measured through:

- Jaccuard index - Measures similarity of predicted values and true values
- F1 - score - Measures accuracy based on true/false positives and true/false negatives

4.1 KNN

K-Nearest Neighbours classifies a new data point, based on the majority vote of the classes of it's nearest neighbours. The K value describes how many nearest neighbours should be taken into account in order to classify a new data point. The training process involves simply finding what value of K allows for the highest classification score on the training data. This model is then used on the test data and the Jaccard and F1-Scores are calculated. The results of training are shown in figure 4.1. One sees that the score significantly increases at even numbers of nearest neighbours. This highlights an issue with the way the experiment was carried out, which is that when K is an even number, there may not be a majority vote (the nearest neighbour's may be labelled with equal numbers of 1's and 2's). In this case, the function that was used to carry out this research automatically predicted that the new data point had a label of 1. This improved accuracy due to the fact that the majority of the data was labelled with a 1. The best accuracy was found at k=10, and this is what was used in the remainder of the results.

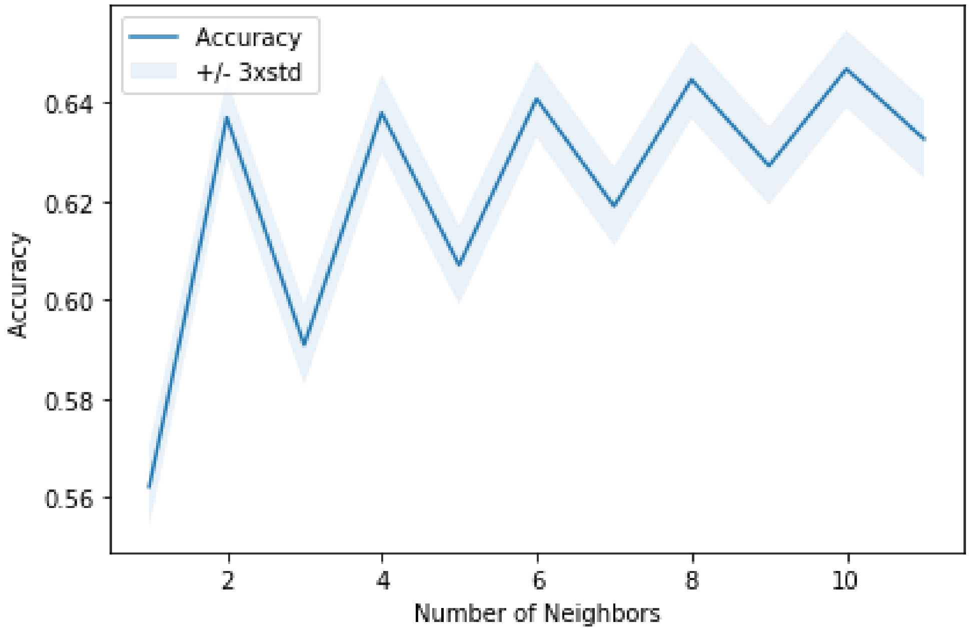


Fig. 4.1 Graph to show the accuracy as K is varied with standard deviation. This graph is obtained from using attributes 'WEATHER', 'ROADCOND' and 'LIGHTCOND'.

4.2 Decision Tree

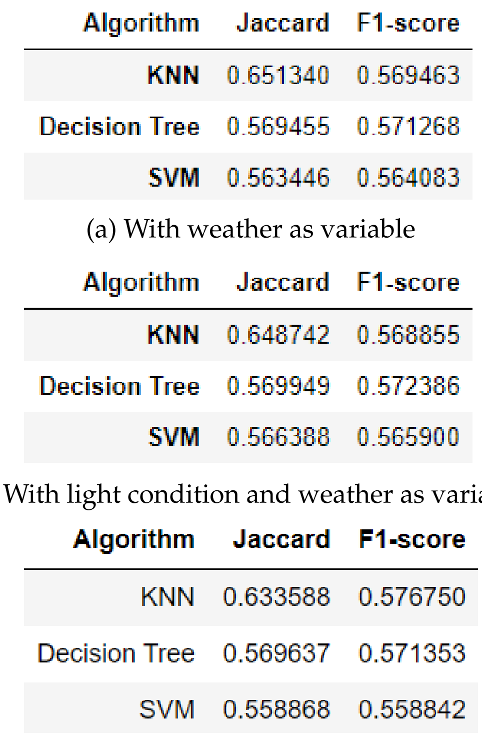
Within a decision tree, a series of boolean questions are used to classify data points. The training process involves fine tuning these questions, in order to maximize the accuracy of the classification on the training data set. Finally, data that is previously unseen by the model is classified, and the Jaccard and F1-Scores are calculated.

4.3 Support Vector Machine

Support vector machine classifies cases by finding a decision boundary in feature space that splits the two classes. The training process involves finding the best location for this separator.

4.4 Results & Discussion

As seen in 4.2, all accuracy values sit between 0.5 and 0.7. The highest value is Jaccard index = 0.65 for kNN model. kNN model is has the highest Jaccard index. The Jaccuard index decreases as the attribute increases in kNN modelling, but increases otherwise. The f1-score decreases when the weather and light condition attributes are used for each model.



Algorithm	Jaccard	F1-score
KNN	0.651340	0.569463
Decision Tree	0.569455	0.571268
SVM	0.563446	0.564083

(a) With weather as variable

Algorithm	Jaccard	F1-score
KNN	0.648742	0.568855
Decision Tree	0.569949	0.572386
SVM	0.566388	0.565900

(b) With light condition and weather as variables

Algorithm	Jaccard	F1-score
KNN	0.633588	0.576750
Decision Tree	0.569637	0.571353
SVM	0.558868	0.558842

(c) With road condition, light condition and weather as variables.

Fig. 4.2 Table showing the accuracy of different means of classification.

Most of the Jaccard scores are around 0.6, and there is a simple reason for this. If a model learned to place the decision boundary a long way from the data (so that all new data points were classified as a 1), it would classify 75% of data points correctly. Labelling the true labels with A, and the predictions with B, the Jaccard score in this case would be calculated using:

$$J(A,B) = \frac{|A \cap B|}{|A| + |B| + |A \cap B|} = \frac{0.75}{1+1-0.75} = 0.6 \tag{4.1}$$

The fact that the decision tree and SVM often obtained a lower score than this, implies that they overfit on the training data.

Chapter 5

Conclusion

This report has explored the severity of incidents in Seattle. It has looked at the total number of incidents, people involved and vehicles involved as part of data analysis. Then it has used the weather to see whether the type of incident can be predicted. The results weren't definitive but did suggest overfitting and therefore the models do need to be looked at again.

5.1 Future improvements

This analysis requires a higher runtime and memory and therefore with more time the decision tree can be a very effective model to predict severity. With that in mind, the attributes used in this analysis wouldn't need to be grouped. It would also be interesting to look at whether the location has any bearing on the type of incident.

References

[1] 2019 washington state car accident statistics reports. Feb 2018.

[2] Bryan Jones, Lester Janssen, and Fred Mannering. Analysis of the frequency and duration of freeway accidents in seattle. *Accident Analysis & Prevention*, 23(4):239–255, 1991.

[3] Most common job in seattle isn't in retail anymore, and 4 out of 5 of these workers are men. June 2018.

[4] Terje Assum and Michael Sørensen. Safety performance indicator for alcohol in road accidents—international comparison, validity and data quality. *Accident Analysis & Prevention*, 42(2):595–603, 2010.

[5] Average software developer salary in seattle, washington. 2016.

[6] Average retail sales associate hourly pay in seattle, washington. 2016.

[7] Fatal car crashes and road traffic accidents in seattle, washington. 2016.

[8] Athanasios Theofilatos and George Yannis. A review of the effect of traffic and weather characteristics on road safety. *Accident Analysis & Prevention*, 72:244–256, 2014.

[9] Venkataraman Shankar, Fred Mannering, and Woodrow Barfield. Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. *Accident Analysis & Prevention*, 27(3):371–389, 1995.

[10] Collisions. May 2018.

[11] Seattle public and national holidays 2020. 2020.