ACCIDENT PREDICTION USING SDOT COLLISION DATASET

INTRODUCTION

Using the dataset provided form SDot Traffic Management Department, we have developed a simple model using KNN and SVM for predicting the type of accident that has taken place under various weather condition, light condition, etc.

Predicting the type of accident that will take place, is a form of knowing which type of accident have higher probability of occurring on a certain day, and help the Traffic management department prepare accordingly beforehand. Data Science and predictive analysis can be utilised efficiently to predict the type of collision that might take place on a certain day, based on the characteristics of the day.

Data

Data Understanding

For the predictive we are going to use the dataset provided to us by the Seattle Traffic Management Department, that is weekly updated database of collisions or traffic accidents that occurred in the city of Seattle.

The data set contains a total of 38 columns detailing everything from exact location of the accident to weather conditions and the type of collision that takes place. Out of all this data we create another data frame that helps us determine the type of collisions that takes place. We choose the attributes, like- the key of incident, address type, severity code, collision type, number of people involved, number of vehicles involved, the type of junction where the accident takes place, speeding information, information of if driver was intoxicated, weather, road condition, light condition, and day of the week to make a prediction.

Data Preparation

Before modelling our data, we have to make the data prepared by cleaning the data, and normalizing it. We begin preparing our data by first removing nan values and changing the types of various attributes. Also, attributes like collision type with string values have to label encoded into float form.

# Methodology

Data Visualization

The first graph shows the relation between the day of the week and the number of total accidents that take place on that day. It gives us an idea that there are more accidents taking place on the weekends than on the weekdays. With the highest number of accidents taking place on Fridays.
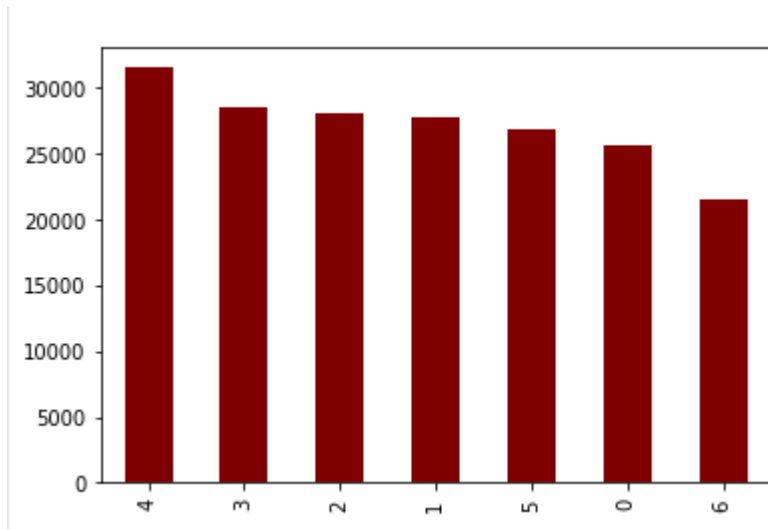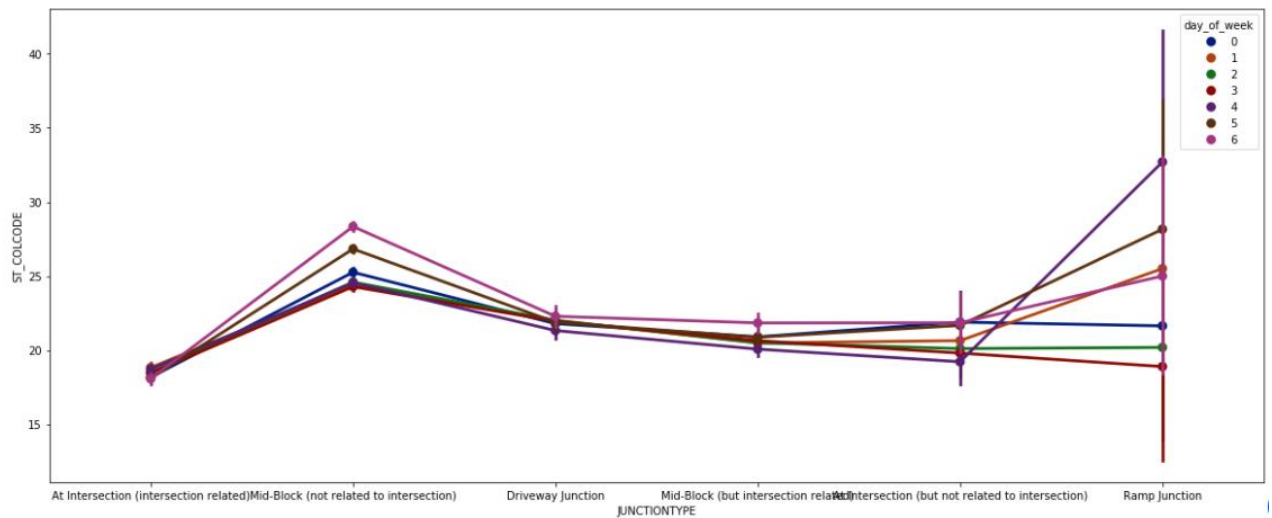
The second graph shows us the junction types and the count of accidents related.
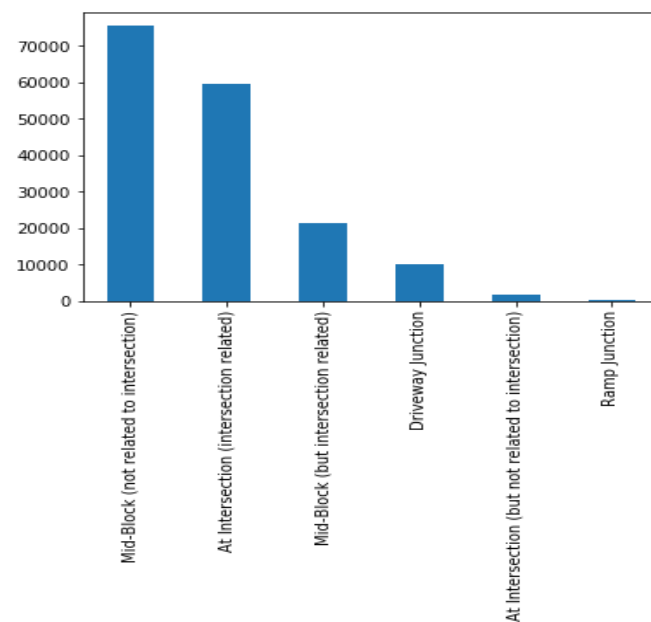
The third graph displays how many accidents that did take place in a junction took place while speeding.
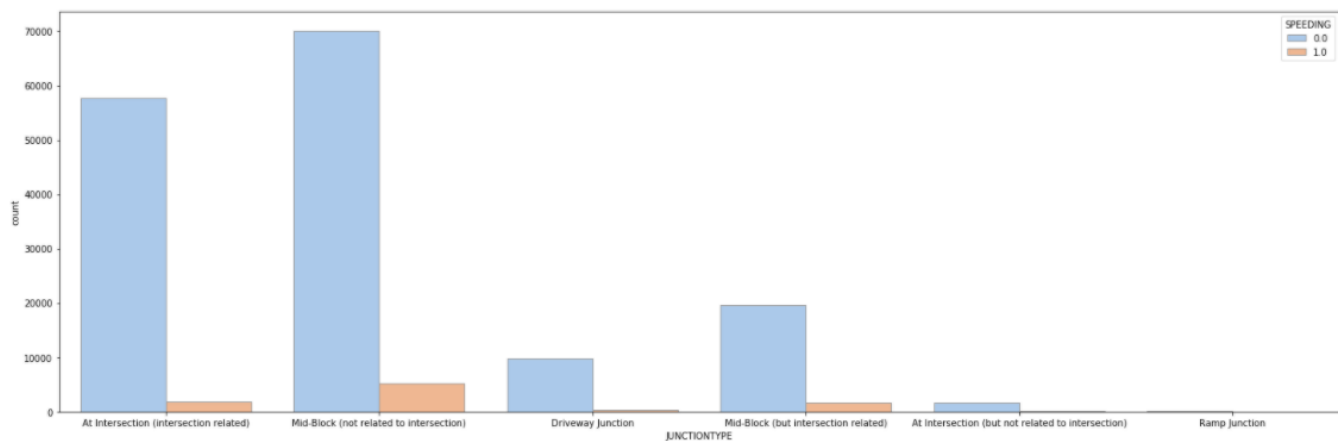
The fourth graph shows us which are the most common kind of accident that takes place in a junction on a particular day of the week.


The fifth graph displays the count of the accident that takes place under a particular lighting condition.
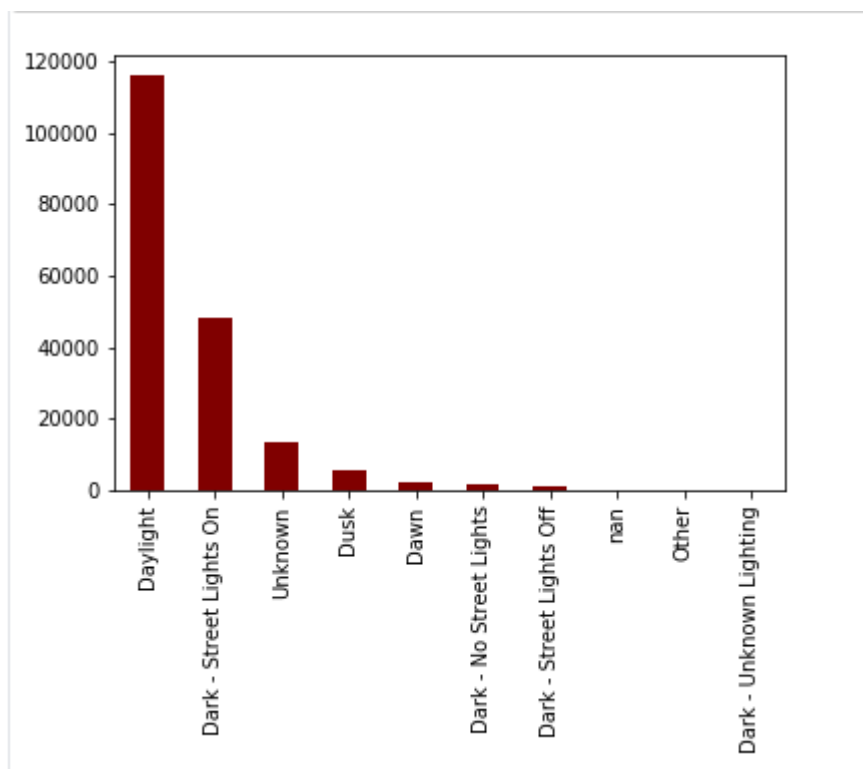

The Sixth graph helps us to find to see a which type of accidents take place under a particular light condition, divided into two categories whether speeding was involved or not. It shows how ST_COLCODE 40-60 where the most common among the speeding drivers, which include vehicles overturning, and vehicles striking a fixed object. Also, we see under daylight we see less number of accidents involving speeding.
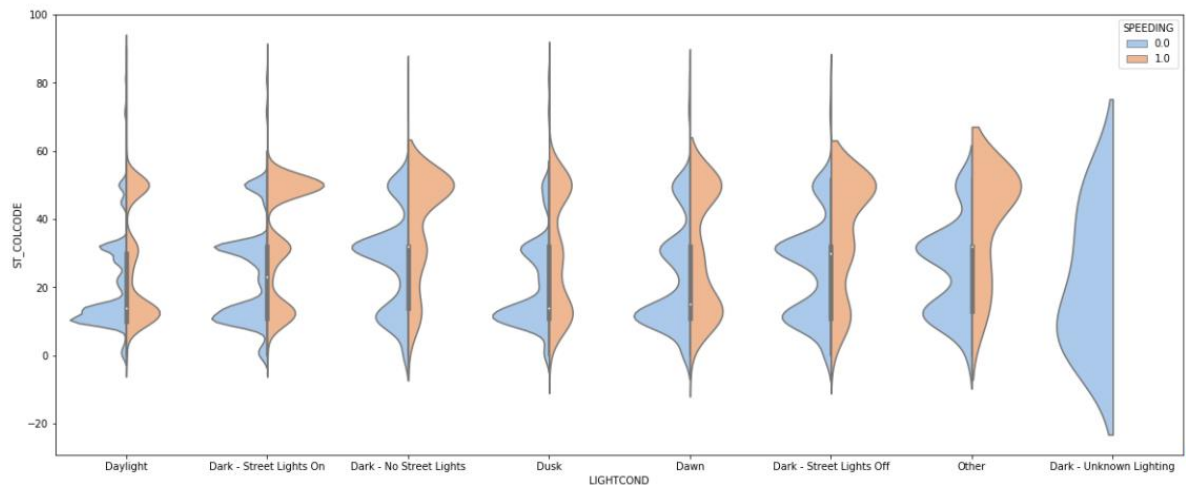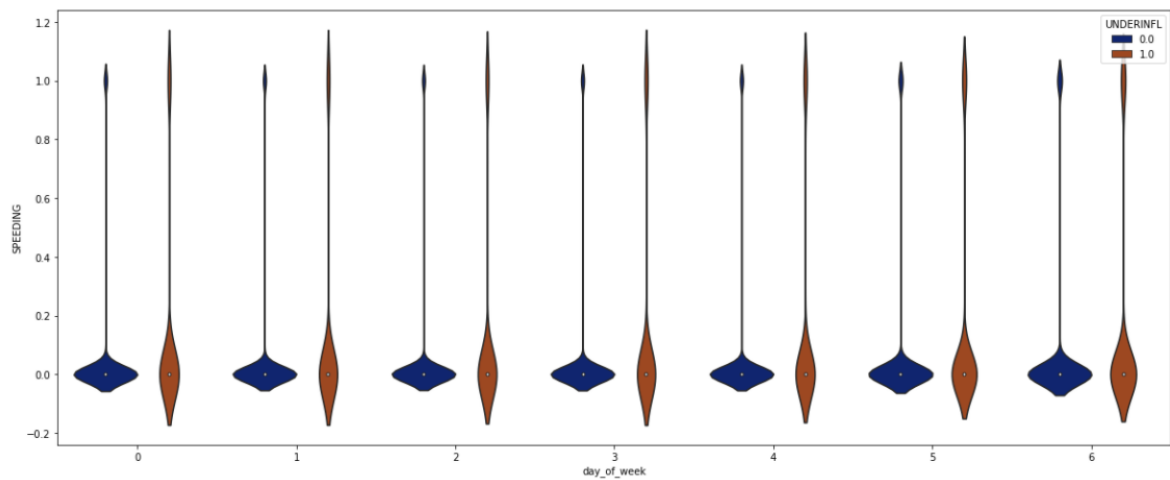
The seventh graph shows us which type of accidents take place under various road condition, divided by whether or not speeding was involved. Under wet, and dry climate they observe minimum number of accidents, while the sand or dirt causes maximum number of accidents. Also, there a lot of accidents in ST_COLCODE 0-20 that takes place in oily road conditions, specially when speeding is involved.
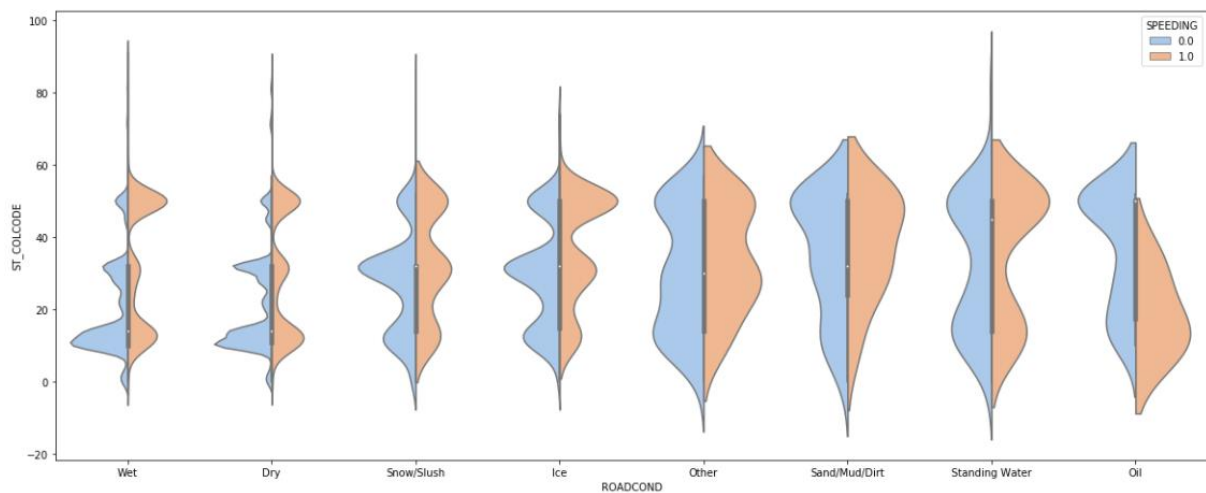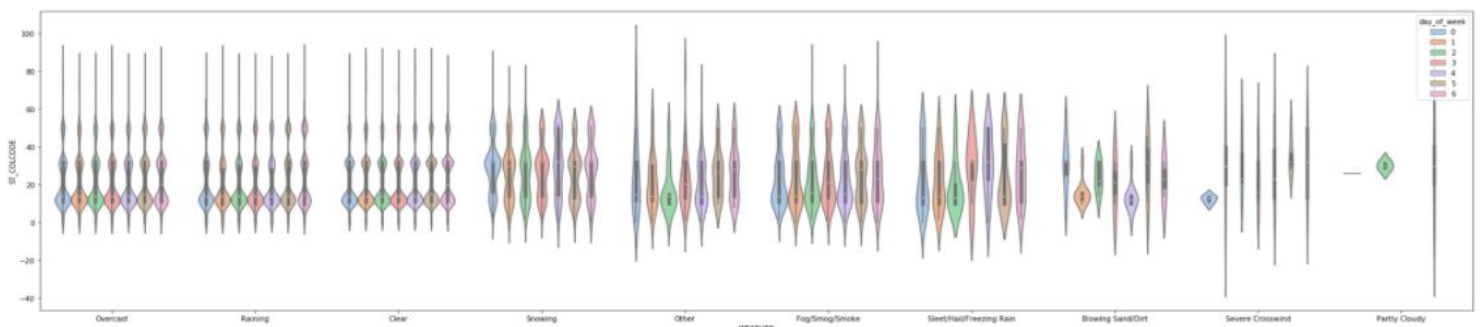

The eighth graph shows us which type of accidents takes place under various road conditions, divided by whether or not driver was under the influence. Under these circumstances we observe a decrease in accidents caused under oily road condition as compared to prior graph.
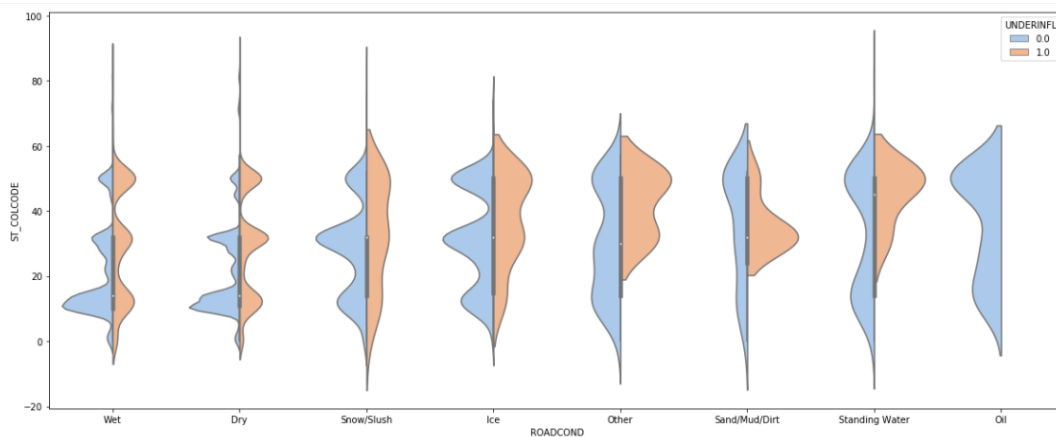
The ninth graph shows how the speeding and day of the week are related, divided by whether the collisions took place under influence.


The tenth graph shows the most common types of accidents that take place under certain weather conditions split into days of the week.

The Eleventh graph shows which types of accidents takes place under various weather conditions and if speeding was involved in the incident.
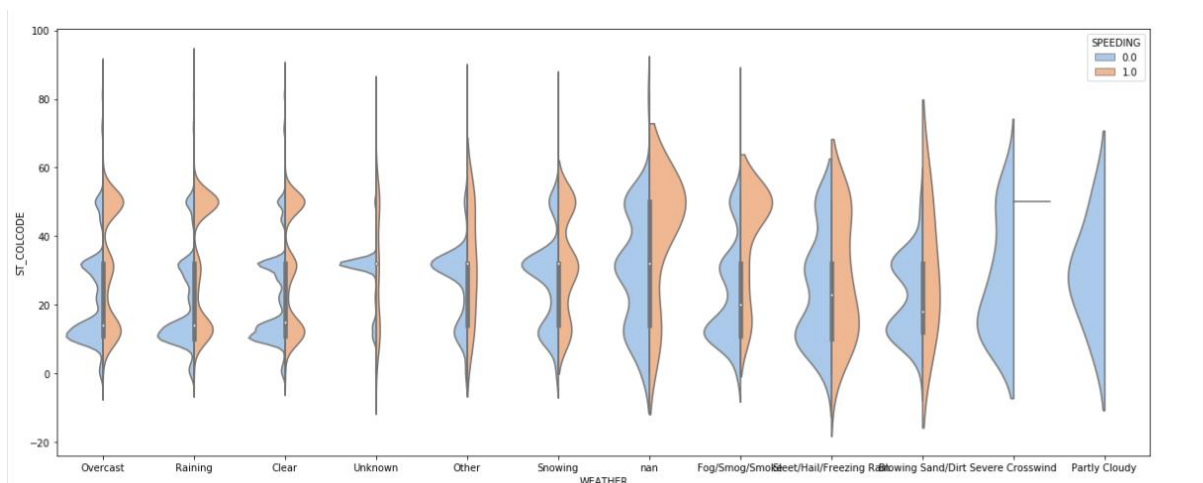
The Twelfth graph shows which types of accidents takes place under various weather conditions and if driver was driving under the influence.
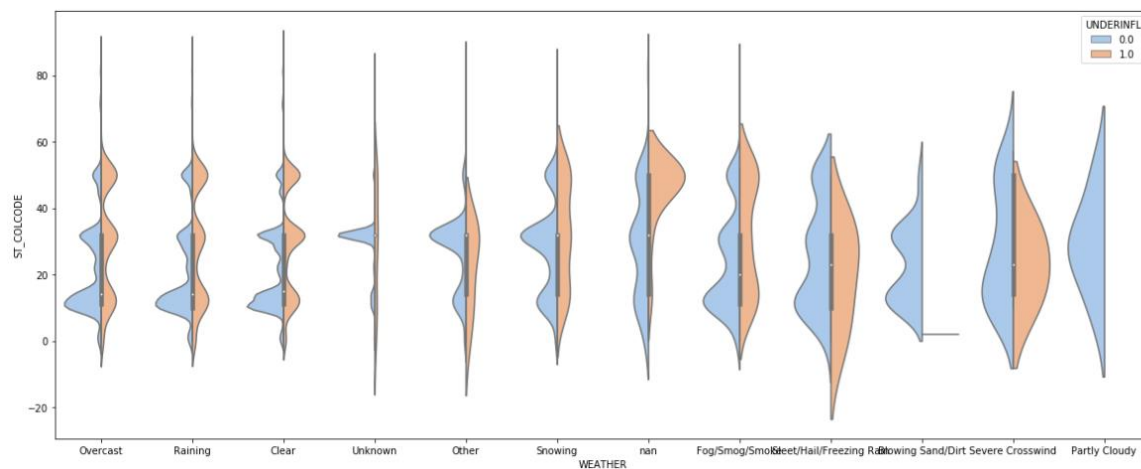
Model Selection

There are three types of classification models we have utilised to make a prediction of types of accidents.

First, we created a separate smaller data frame with only the attributes directly effecting the type of accident. For K- neighbour classification modelling we use the attributes- Collision type, speeding, under influence, weather, road conditions, light conditions, and days of the week.

For the SVM and decision tree we used the attributes-collision type, junction type, speeding, under influence, weather, road conditions, light conditions and days of the week.

Then to measure the accuracy of each model we use F1 Score, to get an idea about how correct our predictive model is by running it on test data.

Results

The model that we have created have highest accuracy of 71% with kNN modelling, when we assume k=8. The k vs f1_score graph when plotted is shown below.
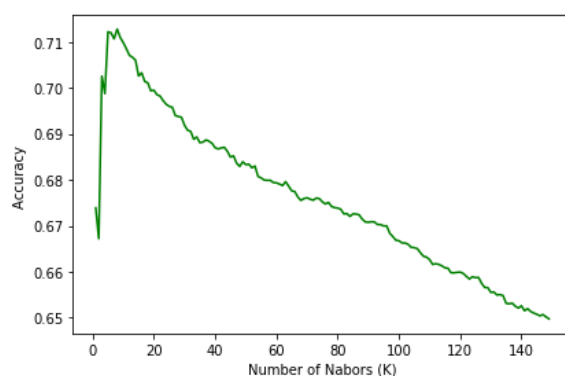
Similarly, the decision tree modelling gives us an accuracy of 74.92% when depth is considered to be 12. The depth vs f1_score graph is as below.
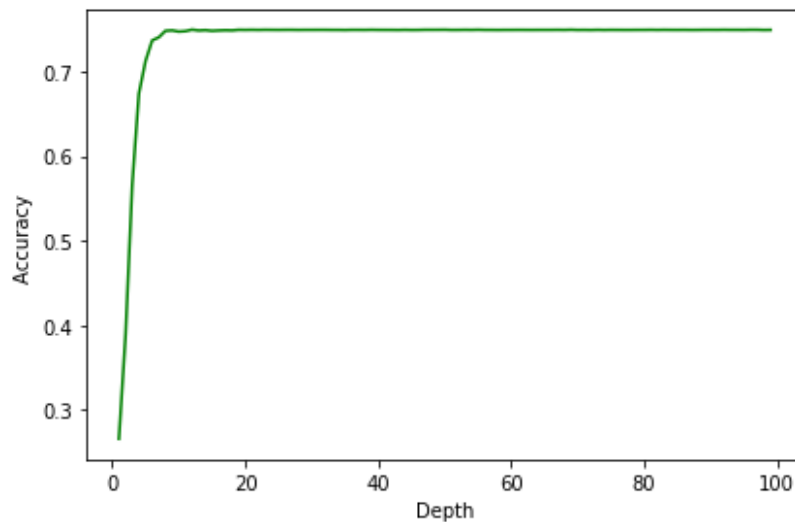
Finally, the SVM modelling, gives us an accuracy of 73.8%, using kernel='rbf'.

Conclusion

In this study, we analysed various relationships between the type of accidents that occur in Seattle and physical and conditional characteristics/conditions that may or may not have led to the accidents. We identified that weather, light conditions, road conditions, and junction types to be some of key aspects that influence the types of accidents that occur. We built classification models like KNN, SVM and decision tree to predict which kinds of accidents that can take place under various circumstances. These models can really help the traffic management department to prepare before hand for any accidents, by predicting which type of accidents have an higher probability of occurring on any given day based of on what day of the week it is, what is the weather conditions, how is the road conditions, etc.