Predicting Accident Severity in Seattle using Weather, Road conditions and Light conditions.

Elijah Chipato

September 6, 2020

1 Introduction/ Business Problem

In this project the Seattle accident Data will be used to predict the accident severity based on the weather, road and light conditions. Four models are then used to find the best classifier namely a decision tree, a support vector machine, K-Nearest Neighbors and Logistic Regression. Ultimately the purpose of this work would be to warn drivers before they embark on a journey of their odds of running into problems on the road.

2 Data Section

The example dataset for Seattle city is used in this study.

	SEVERITYCODE	X	Y	OBJECTID	INCKEY	COLDETKEY	REPORTNO	STATUS	ADDRTYPE	INTKEY	...	ROADCOND	LIGHTCOND
0	2	-122.323148	47.703140	1	1307	1307	3502005	Matched	Intersection	37475.0	...	Wet	Daylight
1	1	-122.347294	47.647172	2	52200	52200	2607959	Matched	Block	NaN	...	Wet	Dark - Street Lights On
2	1	-122.334540	47.607871	3	26700	26700	1482393	Matched	Block	NaN	...	Dry	Daylight
3	1	-122.334803	47.604803	4	1144	1144	3503937	Matched	Block	NaN	...	Dry	Daylight
4	2	-122.306426	47.545739	5	17700	17700	1807429	Matched	Intersection	34387.0	...	Wet	Daylight

5 rows × 38 columns

Figure 1: First 5 rows of the dataset and a couple of columns.

Figure 1 shows the first 5 rows of the original dataset which has about 38 features/columns. However, in this study only a couple of features will be used for the modelling. These features are shown on Figure 2 below.

	INCDATE	SEVERITYCODE	SPEEDING	WEATHER	ROADCOND	LIGHTCOND
0	2013/03/27 00:00:00+00	2	NaN	Overcast	Wet	Daylight
1	2006/12/20 00:00:00+00	1	NaN	Raining	Wet	Dark - Street Lights On
2	2004/11/18 00:00:00+00	1	NaN	Overcast	Dry	Daylight
3	2013/03/29 00:00:00+00	1	NaN	Clear	Dry	Daylight
4	2004/01/28 00:00:00+00	2	NaN	Raining	Wet	Daylight

Figure 2: First 5 rows of the selected feature to be used in this study.

The severity code feature will be used as the label of the dataset to train the models. In Figure 2 there are columns of incident date (INCDATE) and speeding (SPEEDING), these will be used during feature engineering to assess whether the day of the week is important in predicting accident severity. If observed that it has no impact these two will be dropped.

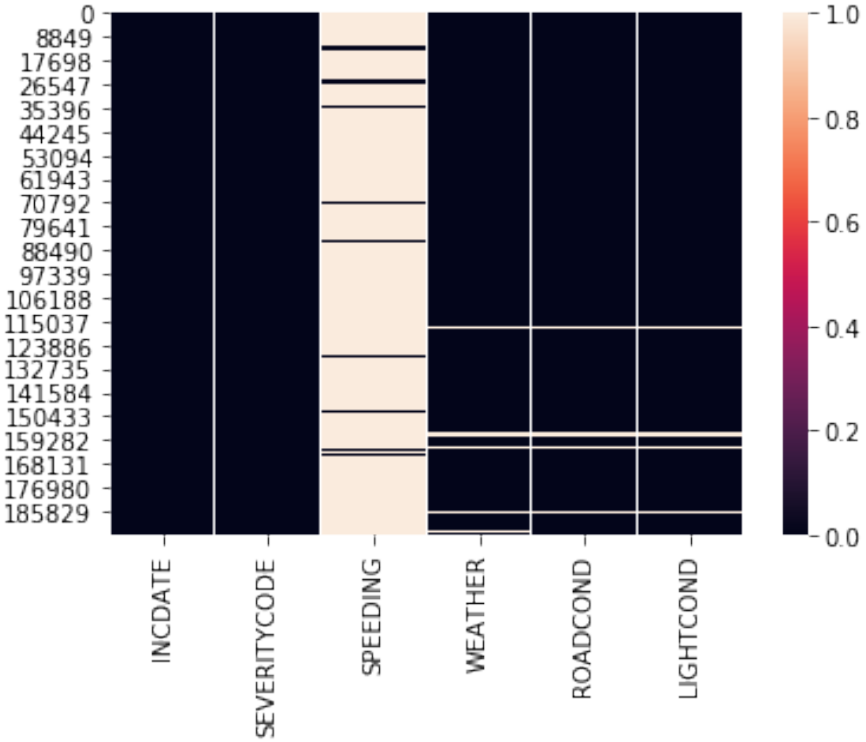

Figure 3: Heatmap showing null values in the dataset.

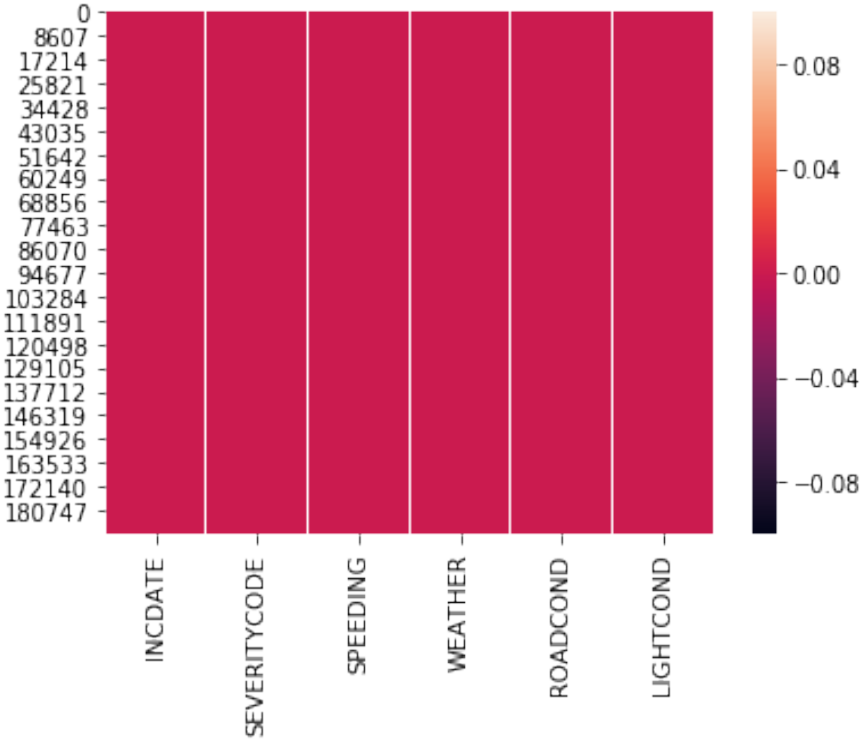

Figure 4: Heatmap showing absence of null values in the dataset.

To assess the quality of the data a heat-map is used to reveal the location or concentration of null values in our dataset. Figure 3 shows a heatmap which reveals that the speeding column has the most null values in the selected features for modelling. WEATHER, ROADCOND and LIGHTCOND also have similar number of null values. For the speeding column the attributes are Y for yes and NaN for No meaning the vehicle was

not speeding. The NaN values are replaced by N to represent No. The null values for the rest of the features namely WEATHER, ROADCOND and LIGHTCOND are dropped. Figure 4 shows the new heatmap after dropping all null values and it shows that the datset is now clean and free of null values.

3 Methodology

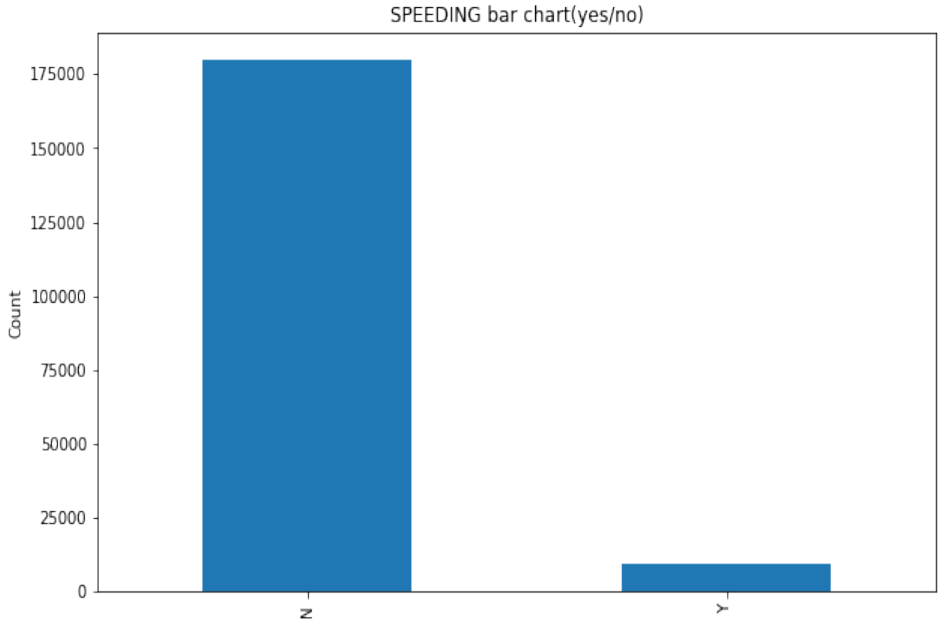

Figure 5: Bar Chart showing speeding data.

After this manipulation a bar chart is plotted to observe the speeding data. Figure 5 shows the bar chart for the speeding column. It is apparent that only a small proportion of drivers are speeding as compared to

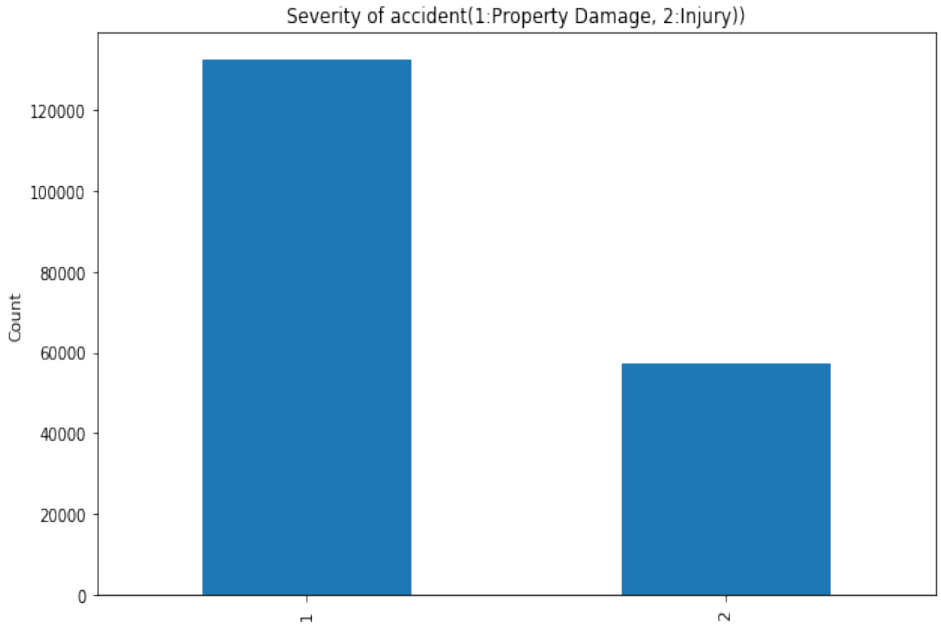

Figure 6: Bar Chart showing count on severity code.

those that are not. Figure 6 shows a bar chart for the label used for modelling that is to say the severity code.

Figure 7: Bar Chart showing count of accidents across the entire week and whether the vehicles are speeding or not.

	SEVERITYCODE	WEATHER	ROADCOND	LIGHTCOND
0	2	Overcast	Wet	Daylight
1	1	Raining	Wet	Dark - Street Lights On
2	1	Overcast	Dry	Daylight
3	1	Clear	Dry	Daylight
4	2	Raining	Wet	Daylight

Figure 8: First 5 rows of the features for the new dataset to be used in modelling.

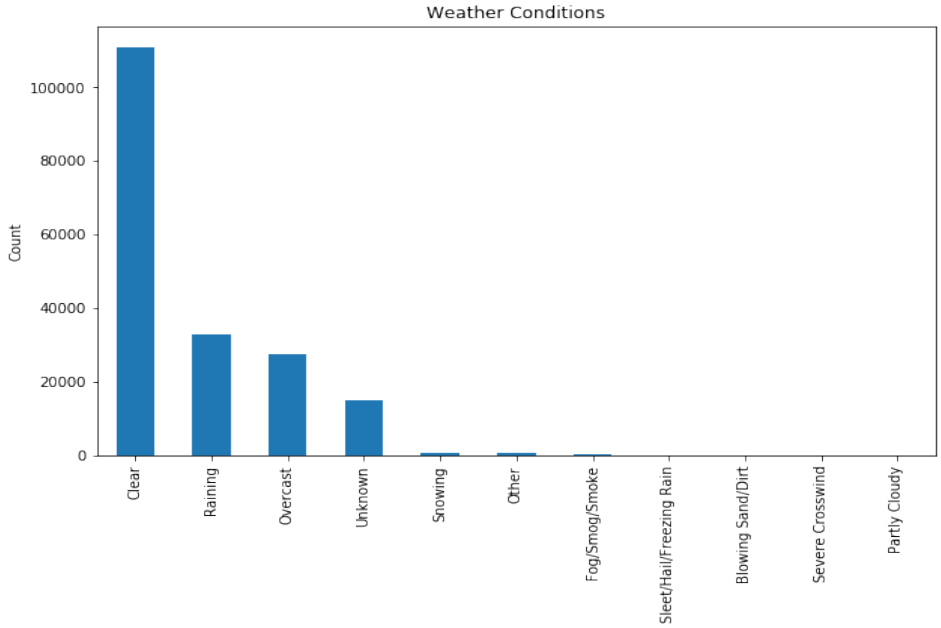

Figure 9: Bar Chart showing count on weather conditions.

It is evident that the data is not balanced and therefore the data will have to be resampled before training

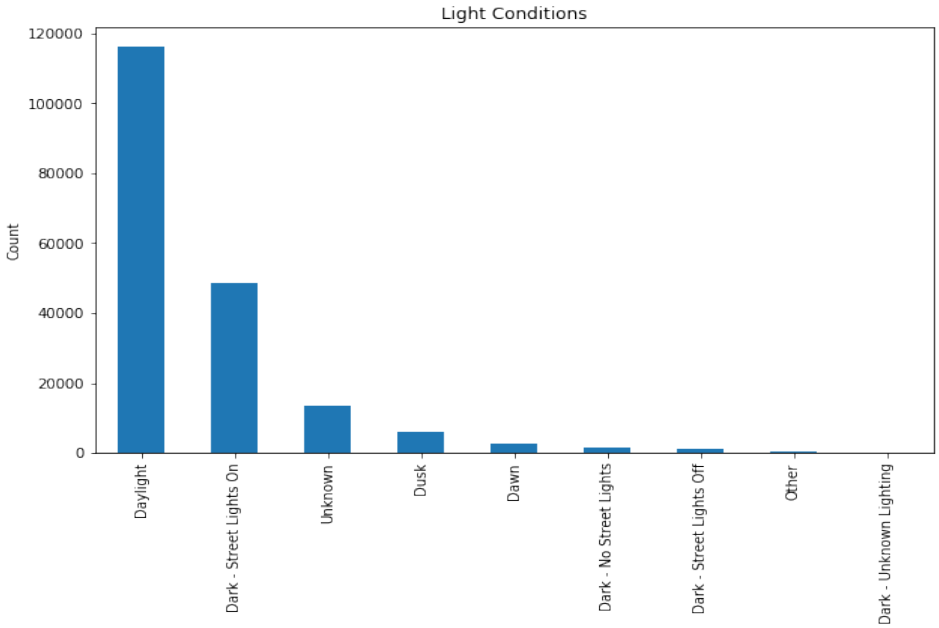

Figure 10: Bar Chart showing count on light conditions.

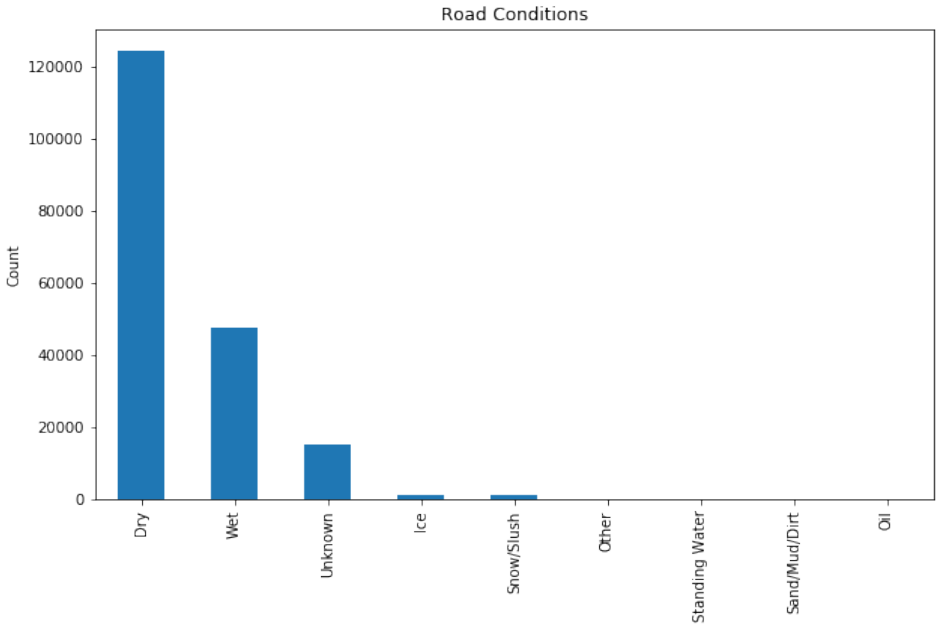

Figure 11: Bar Chart showing count on road conditions.

the models.In addition, another bar chart is also plotted to asses the impact of day of the week on accident severity. Figure 7 shows that for both severity codes that 1 for property damage and 2 for injury, across the entire week the accidents are uniformly distributed therefore for this reason the INCDATE and SPEEDING features are dropped from the data set. Figure 8 shows the new dataset. Figure 9, 10 and 11 shows the bar charts showing the attributes of the WEATHER, LIGHTCOND column. The bar charts also shows that the features include "Unknown" and "Other" for all three columns. These are dropped because these are not specific to any given weather, road or light condition.It is apparent that the most count of accidents happens when there is clear weather therefore bad weather is not directly linked to the accidents.As observed in Figure 8 the features with the exception of the severity code are all categorical. Therefore, the one hot encoding technique is used to convert the data to binary, this process also create new features. Figure 12 shows the new dataset to be used in the modelling after the one hot encoding technique.

Finally, the balanced dataset is then assessed on weather it makes sense. This can be done by finding

	Blowing Sand/Dirt	Clear	Fog/Smog/Smoke	Overcast	Partly Cloudy	Raining	Severe Crosswind	Sleet/Hail/Freezing Rain	Snowing	Dry	...	Snow/Slush	Standing Water	Wet
49258	0	1	0	0	0	0	0	0	0	1	...	0	0	0
184569	0	0	0	0	0	1	0	0	0	0	...	0	0	1
119375	0	1	0	0	0	0	0	0	0	1	...	0	0	0
128465	0	1	0	0	0	0	0	0	0	1	...	0	0	0
48015	0	1	0	0	0	0	0	0	0	1	...	0	0	0

5 rows × 23 columns

Figure 12: First 5 rows of the features for the new dataset to be used in modelling.

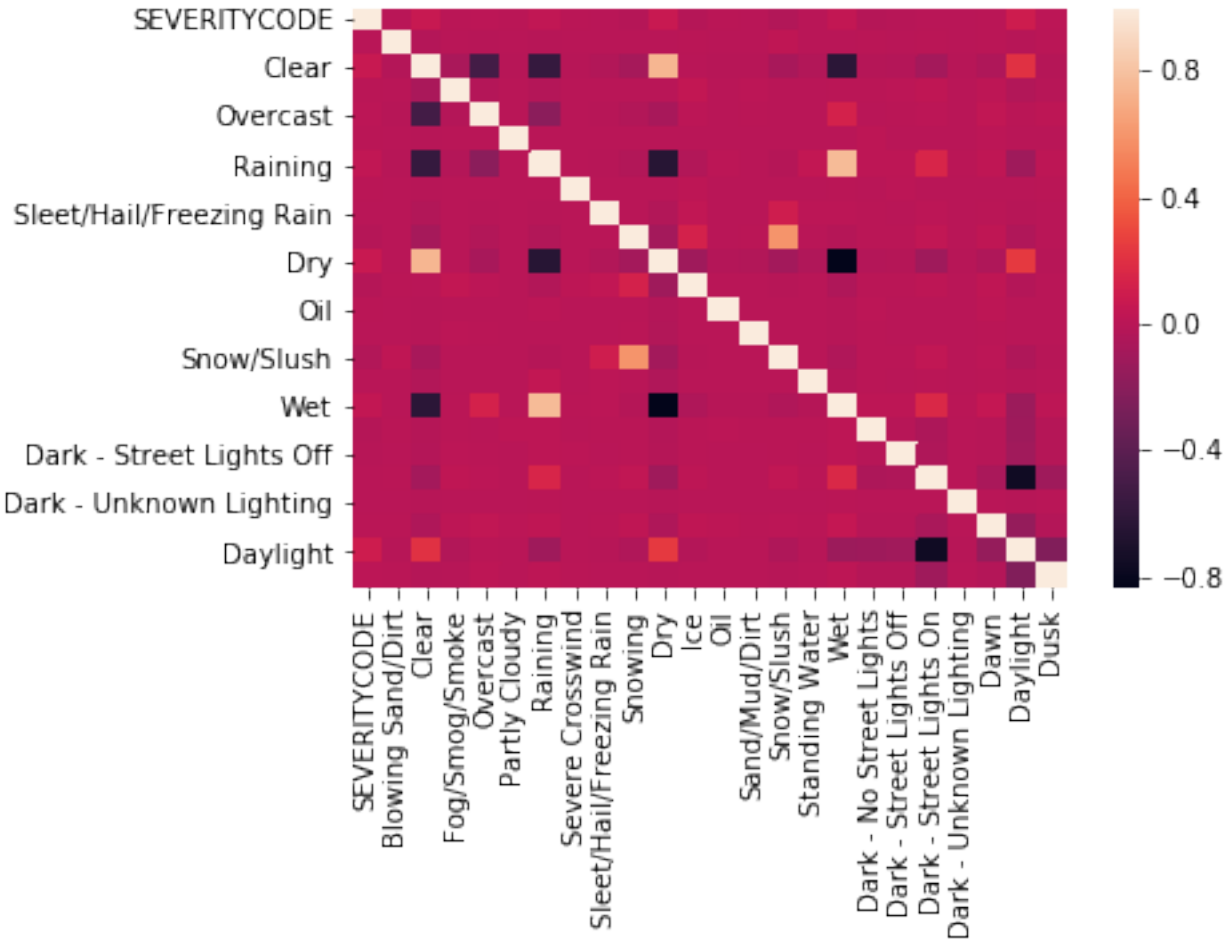

Figure 13: Heatmap showing correlation between different features.

the correlationship between different features. A heatmap is used for this. Figure 13 shows a heatmap of the correlationship. It is obvious that the diagonal in brown is showing full correlationship because features are correlated to them selves. Other features in brown are Clear and dry conditions and wet and Raining conditions. Some features are not correlated and these are in black for example dark-street lights on is not correlated to daylight and wet and clear conditions is not correlated at all. Therefore this dataset makes sense and therefore the modelling should go ahead.

3.1 Modelling

The dataset shown in Figure 12 is first standardized/ normalized to give it zero mean and unit variance. Normalizing the data will generally speed up learning and leads to faster convergence. The data is then split into training and testing sets. In this work 20% of the data was used for testing and 80% for training the

models. Four different models were trained using K-Nearest Neighbors, Decision tree, Support vector machines and Logistic regression algorithms

4 Results and Discussions

The performance of the models was assessed using various mainly the F1 and the Jaccard score.

Table 1: Nonlinear Model Results

ML Algorithm	F1-score	Jaccard	LogLoss
SVM	0.5316	0.5606	NA
Decision Tree	0.5323	0.5609	NA
KNN	0.5415	0.5569	NA
Logistic Regression	0.5312	0.5609	0.6655

Table 1 shows all four models and how well they have performed. The F1 and Jaccard score show that the

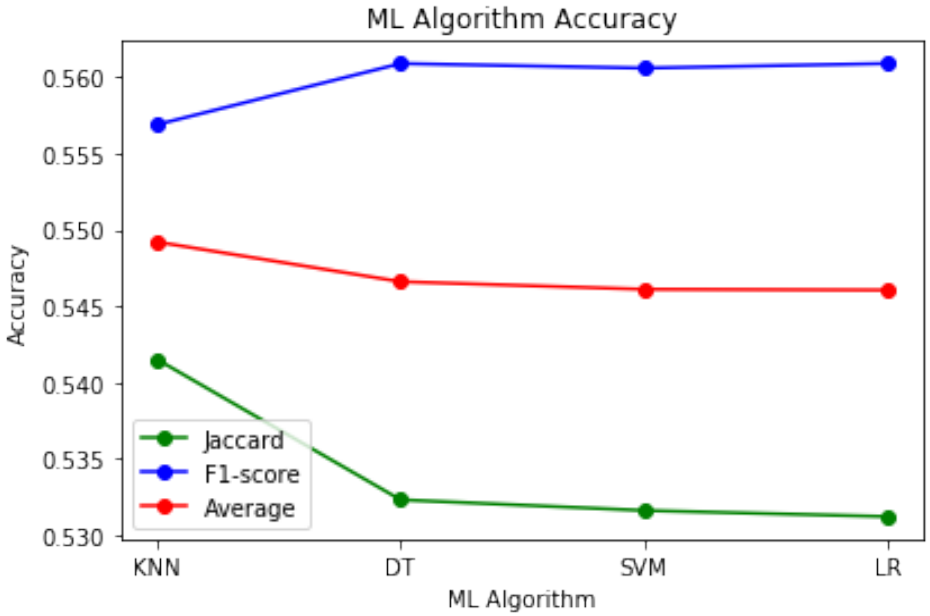

Figure 14: Bar Chart showing count on road conditions.

models have all accurately predicted over 50% of the test data correct. This is a somewhat good result but not satisfactory. Figure 14 shows a line chart for the 2 metrics as well as an average of the two. It is evident from this figure that even though all metrics predict over 50% of the test set correct by judging by the average accuracy, the K-Neareest Neighbors algorithm is the best classifier.

5 Conclusion

In this study, accident severity was predicted using weather, road and light conditions. Models were built, trained and tested using KNN, decision trees, support vector machines and logistic regression models. The performance of the models was good but not excellent. It was also found that the KNN model performed better in predicting that the other three models.