Prediction of Collision Severity

Author: CHALLA K S N M SANKAR

IBM Capstone Project Report

1. Introduction

1.1 Background

According to the 2017 WSDOT data, a car accident occurs every 4 minutes and a person dies due to a car crash every 20 hours. The number of car accidents with suspected serious injuries altogether 1,932 for the year 2017. Suspected serious injury collisions went from 1,896 in 2016 to 1,932 in 2017, resulting in 2,229 serious injuries. Fatal crashes went from 508 in 2016 to 525 in 2017, resulting in the death of 555 people. This number has stayed relatively steady for the past decade. The highest number of crashes occur on Saturdays. The month of November held the highest number of crashes throughout the year. People in the 16-25 age group were the most at risk for being killed in a car accident.
(https://www.injurytriallawyer.com/library/car-accident-statistics-seattle-washington-state.cfm)

1.2 Problem

The project aims at prediction of collision in Seattle area by using supervised learning and develop a model to predict the severity of an accident for given attributes like weather, road conditions

etc. The stakeholder can alert the driver by prediction of collision and reduce the severity of collision or avoid the collision.

2. Data Understanding

2.1 Data source

In this project, we will use the data provided by the SPD about accidents occurred in Seattle area. The data is obtained from the following link.
https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

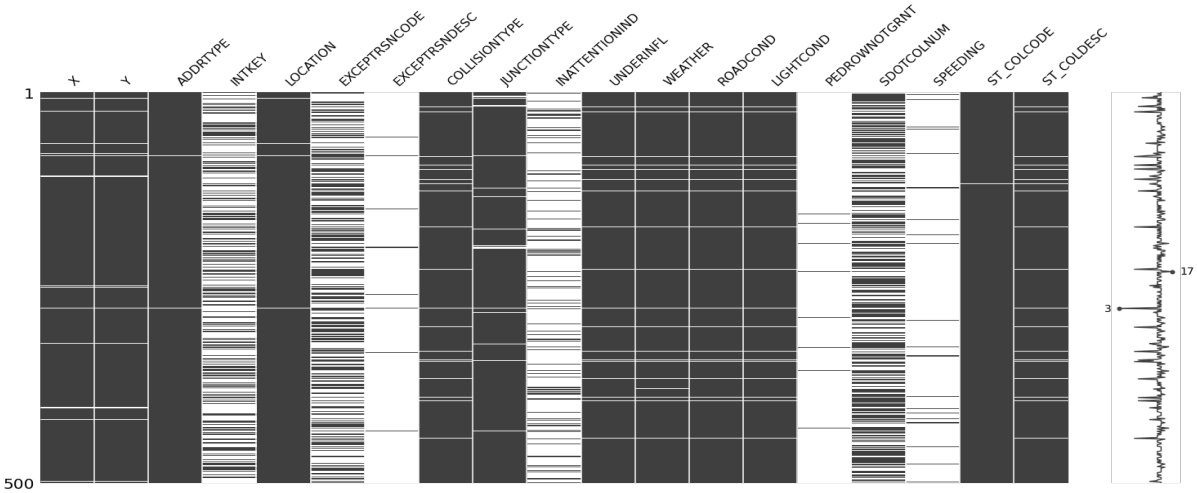
2.2 Data understanding

The data set contains 37 attributes and 194,673 incidents happened. The provided data is an imbalanced data.
The missing data is provided by the following table.

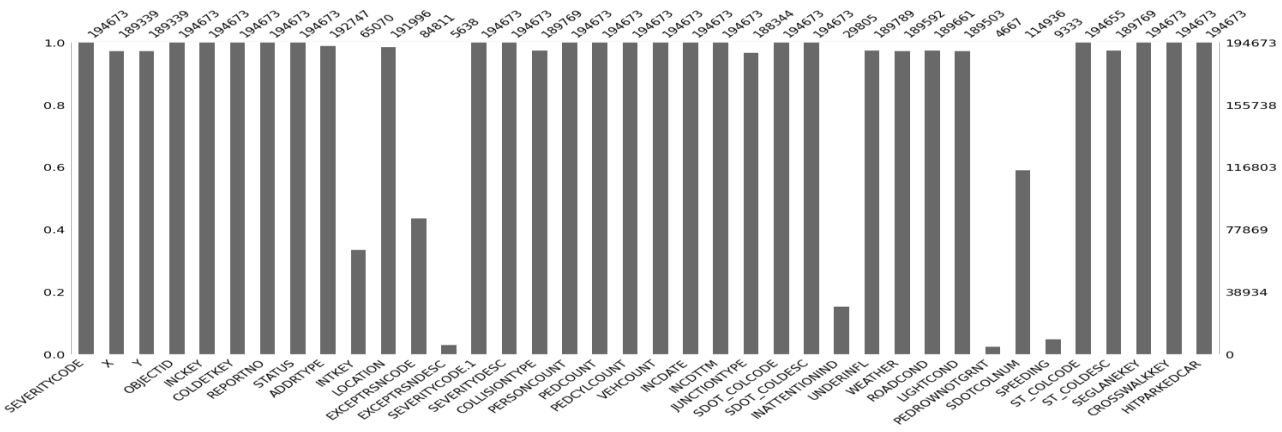
	count_missing	*perc_missing*
PEDROWNOTGRNT	190006	97.602646
EXCEPTRSNDESC	189035	97.103861
SPEEDING	185340	95.205807
INATTENTIONIND	164868	84.689710
INTKEY	129603	66.574718
EXCEPTRSNCODE	109862	56.434123
SDOTCOLNUM	79737	40.959455
JUNCTIONTYPE	6329	3.251093
X	5334	2.739979
Y	5334	2.739979
LIGHTCOND	5170	2.655736
WEATHER	5081	2.610018
ROADCOND	5012	2.574574
ST_COLDESC	4904	2.519096

COLLISIONTYPE	4904	2.519096
UNDERINFL	4884	2.508822
LOCATION	2677	1.375126
ADDRTYPE	1926	0.989351
ST_COLCODE	18	0.009246

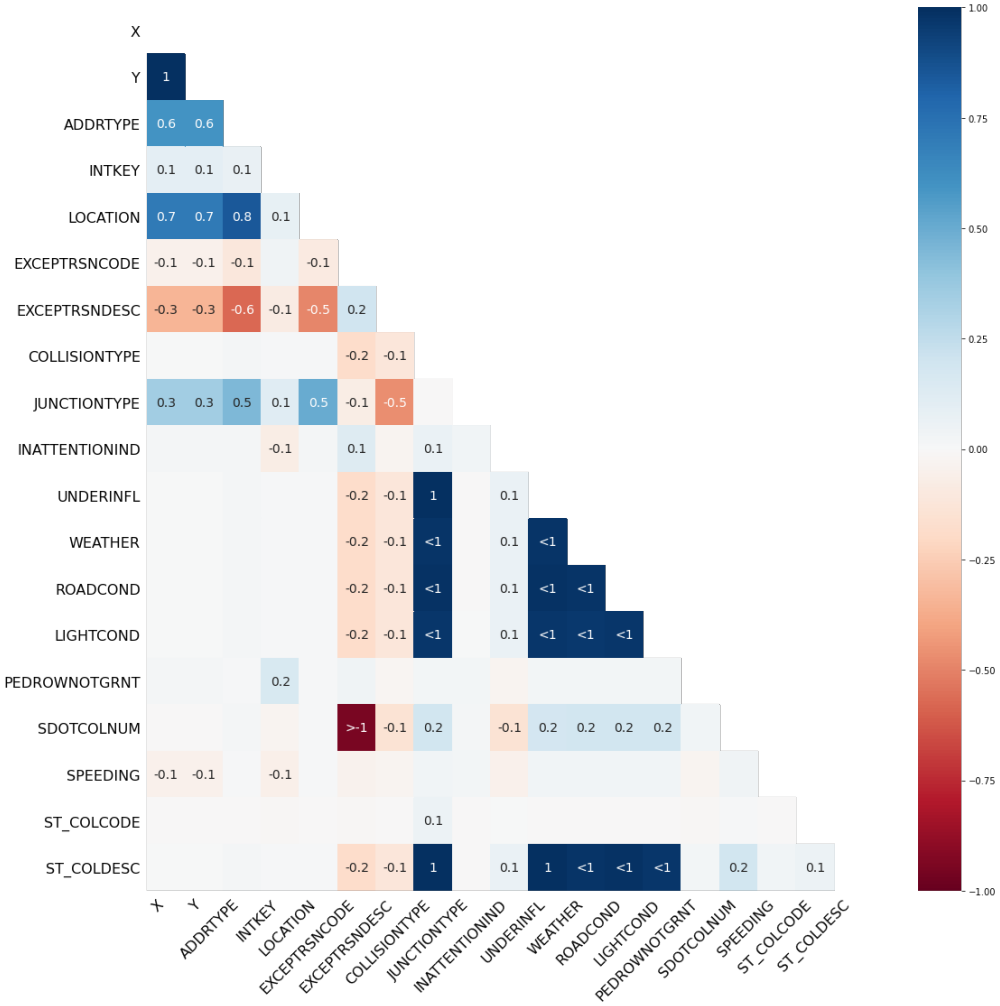
The nullity matrix of the data is as follows.



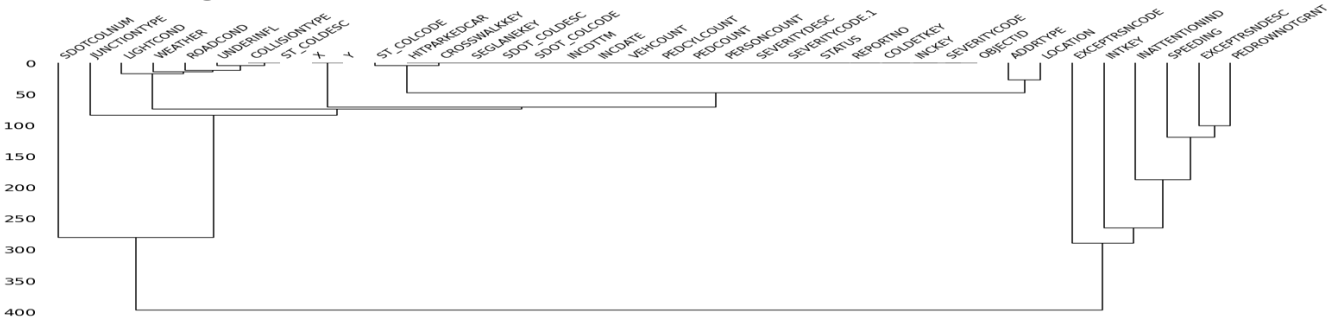
The bar graph of the data is as follows



The Heat map of the data as follows



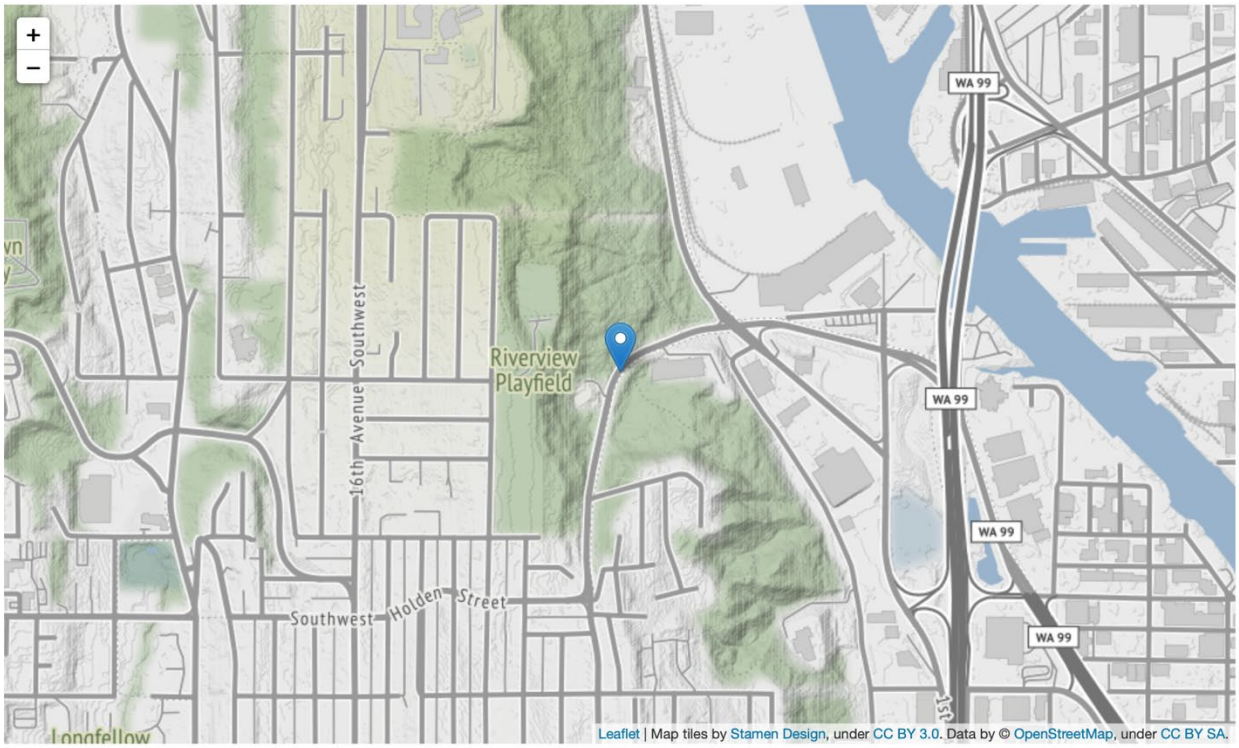
The dendogram of the data is as follows



The missing values in data as shown below.

Attribute	Number of missing values
SEVERITYCODE	0
X	5334
Y	5334
OBJECTID	0
INCKEY	0
COLDETKEY	0
REPORTNO	0
STATUS	0
ADDRTYPE	1926
INTKEY	129603
LOCATION	2677
EXCEPTRSNCODE	109862
EXCEPTRSNDESC	189035
SEVERITYCODE.1	0
SEVERITYDESC	0
COLLISIONTYPE	4904
PERSONCOUNT	0
PEDCOUNT	0
PEDCYLCOUNT	0
VEHCOUNT	0
INCDATE	0
INCDTTM	0
JUNCTIONTYPE	6329
SDOT_COLCODE	0
SDOT_COLDESC	0
INATTENTIONIND	164868
UNDERINFL	4884
WEATHER	5081
ROADCOND	5012
LIGHTCOND	5170
PEDROWNOTGRNT	190006
SDOTCOLNUM	79737
SPEEDING	185340
ST_COLCODE	18
ST_COLDESC	4904
SEGLANEKEY	0
CROSSWALKKEY	0
HITPARKEDCAR	0

The columns 'EXCEPTRSNCODE', 'PEDROWNOTGRNT', 'EXCEPTRSNDESC', 'INATTENTIONIND' and 'INTKEY' are dropped from the data frame. Though with the help of above information and looking to dendograph we can drop the column ‘SPEEDING’ but studied the data firstly without dropping it. It is observed that when the vehicle is speeding there are more chances to have a collision at a particular location -122.346085, 47.539248. With the help of Folium maps, the place is identified as sharp curved road at Riverview Playfield.



Now the column ‘SPEEDING’ is also dropped from the data frame as studied from dendograph.

3. Exploratory Data Analysis

3.1 Data Cleaning

The new data frame is selected choosing the columns ‘X’, ‘Y’, ‘SE VERITYCODE’, ‘ROADCOND’, ‘LIGHTCOND’, ‘WEATHER’ an d ‘ADDRTYPE’.

The columns ‘X’ and ‘Y’ gives the longitude and latitude of the location where collision occurred. The column ‘SEVERITYCODE’ specifies whether collision is only property loss or injury collision. The column ‘LIGHTCOND’ specifies the lighting conditions at the moment of collision. The column ‘WEATHER’ specifies the weather conditions at the instant collision happened. The new count of the data after dropping nulls is 184167.

The count of values of ‘WEATHER’ as follows.

Clear	108833
Raining	31987
Overcast	27105
Unknown	13846
Snowing	888
Other	765
Fog/Smog/Smoke	553
Sleet/Hail/Freezing Rain	112
Blowing Sand/Dirt	49
Severe Crosswind	24
Partly Cloudy	5

The count of values of ‘ROADCOND’ as follows.

Dry	121871
Wet	46009
Unknown	13795
Ice	1174
Snow/Slush	984
Other	116
Standing Water	102
Sand/Mud/Dirt	63
Oil	53

The count of values of ‘LIGHTCOND’ as follows.

Daylight	113522
Dark - Street Lights On	47250
Unknown	12416

```
Dusk                            5763
Dawn                            2422
Dark - No Street Lights         1450
Dark - Street Lights Off        1145
Other                            188
Dark - Unknown Lighting           11
```

The count of values of ‘ SEVERITYCODE’ as follows.

```
1    128154
2     56013
```

The rows containing categorical values of ‘WEATHER’ column ‘ Unknown’, ‘Snowing’, ‘Other’, ‘Fog/Smog/Smoke’ , ‘Sleet/Hail/ Freezing Rain’, ‘Blowing Sand/Dirt’, ‘Severe Crosswind’, ‘Partly Cloudy’ are removed.
The rows containing categorical values of ‘ROADCOND’ column ‘Unknown’, ‘Ice’, ''Snow/Slush', 'Other', 'Standing Water', 'Sand /Mud/Dirt', 'Oil' are removed. The rows containing categorical values of ‘LIGHTCOND’ column ‘Unknown’ , ‘Dust’, ‘Dawn’ , ‘D ark - No Street Lights’, ‘Dark – Street Lights Off’, ‘Other’, ‘Dark – Unknown Lighting’ are removed. After dropping, the new co unt of the data is 161033. The first five lines of the
data are

	X	Y	SEVERITYCODE	ROADCOND	LIGHTCOND	WEATHER	ADDRTYPE
0	-122.323148	47.703140	2	Wet	Daylight	Overcast	Intersection
1	-122.347294	47.647172	1	Wet	Dark - Street Lights On	Raining	Block
2	-122.334540	47.607871	1	Dry	Daylight	Overcast	Block
3	-122.334803	47.604803	1	Dry	Daylight	Clear	Block
4	-122.306426	47.545739	2	Wet	Daylight	Raining	Intersection

The count of values of ‘SEVERITYCODE’ as follows.

```
1    107634
2     53399
```

The best choice is to choose ‘SEVERITYCODE’ as target variable whose values are either 1 or 2 and other selected columns are c onsidered as attributes. The target variable data is imbalanced a nd it is fixed by

downsampling of major class. After downsampling of major clas s, the count of each class of ‘SEVERITYCODE’ is 53399.

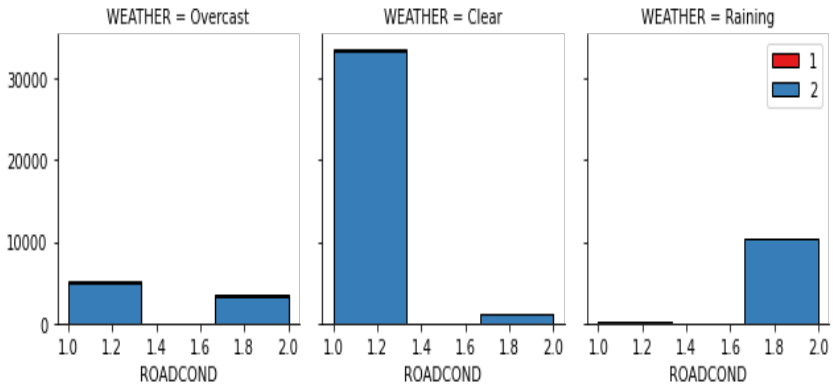
```
2    53399
1    53399
Name: SEVERITYCODE, dtype: int64
```

The new count of data after downsampling is 106798.

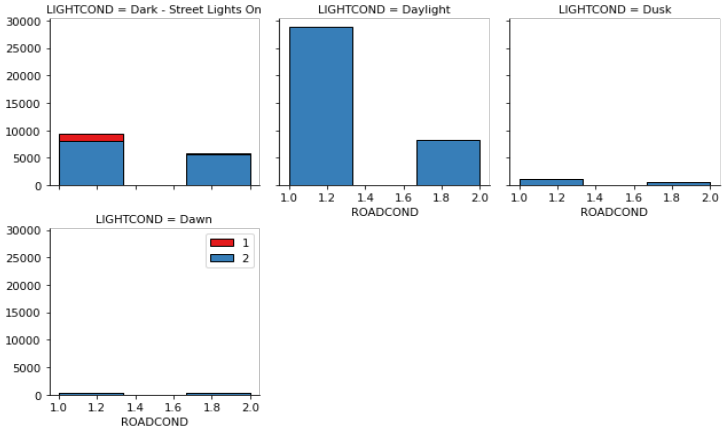
For further analysis, the ‘ROADCOND’ column categorical values 'Dry', 'Wet', are replaced with integers 1 and 2 respectively. Also ‘ADDRTYPE’ column categorical values are replaced with integers 0 and 1 respectively.

3.2 Data Visualization

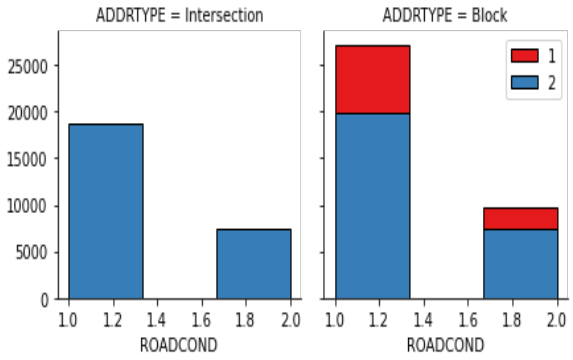
The histograms of count of severity of collision vs road conditions are drawn for various weather conditions are as follows.



The histograms of count of severity of collision vs road conditions are drawn for various lighting conditions are as follows.



The histograms of count of severity of collision vs road conditions are drawn for various address types are as follows.



4. Data Modelling

The data is ready for modelling. The categorical values of 'WEATHER' column and 'LIGHTCOND' are encoded with binary values. The feature data after encoding is as follows.

	X	Y	ROADCOND	ADDRTYPE	Clear	Overcast	Raining	Dark - Street Lights On	Dawn	Daylight	Dusk
91327	-122.379987	47.679521	1	0	0	1	0	1	0	0	0
74418	-122.374857	47.663663	1	1	1	0	0	0	0	1	0
37046	-122.376223	47.656989	2	1	0	0	1	0	0	1	0
93018	-122.345047	47.729585	1	1	1	0	0	0	0	1	0
126042	-122.382526	47.573156	2	0	0	1	0	0	0	0	1

Afterwards the Data is normalised. The normalised data as follows.

```
array([[-1.67253057,  1.05020178, -0.61506543, -1.22541887, -1.36003473,
         2.30800687, -0.48881647,  1.63773678, -0.11641411, -1.46123785,
        -0.18763101],
       [-1.49946977,  0.77182281, -0.61506543,  0.8160475 ,  0.73527534,
        -0.43327427, -0.48881647, -0.61059873, -0.11641411,  0.68435128,
        -0.18763101],
       [-1.54554556,  0.65466922,  1.6258433 ,  0.8160475 , -1.36003473,
        -0.43327427,  2.04575757, -0.61059873, -0.11641411,  0.68435128,
        -0.18763101],
       [-0.49397868,  1.92901777, -0.61506543,  0.8160475 ,  0.73527534,
        -0.43327427, -0.48881647, -0.61059873, -0.11641411,  0.68435128,
        -0.18763101],
       [-1.75816205, -0.81694954,  1.6258433 , -1.22541887, -1.36003473,
         2.30800687, -0.48881647, -0.61059873, -0.11641411, -1.46123785,
         5.32960934]])
```

5. Classification models

The following Supervised machine learning classification models are build. The following models are trained by choosing testing data size 20%.

- K Nearest Neighbor(KNN)
- Decision Tree
- Support Vector Machine
- Logistic Regression

5.1 K Nearest Neighbor(KNN)

The best value of K =59 to build KNN model.

```
K = 59 mean_accuracy = 0.59573970
K = 59 std_accuracy = 0.00335783
```

5.2 Decision Tree

The Decision Tree Classifier is build with criterion="entropy", max_depth = 4.

5.3 Support Vector Machine

Support vector classifier is build with criteria, kernel='rbf' and gamma = 'auto'.

5.4 Logistic Regression

The Logistic Regression model is build with criteria C=0.01, solver='liblinear'.

6. Model Evaluation Using Test Set

The models are evaluated using test set. The features of test set are as follows.

	X	Y	ROADCOND	ADDRTYPE	Clear	Overcast	Raining	Dark - Street Lights On	Dawn	Daylight	Dusk
0	-122.323148	47.703140	2	0	0	1	0	0	0	1	0
1	-122.347294	47.647172	2	1	0	0	1	1	0	0	0
2	-122.334540	47.607871	1	1	0	1	0	0	0	1	0
3	-122.334803	47.604803	1	1	1	0	0	0	0	1	0
4	-122.306426	47.545739	2	0	0	0	1	0	0	1	0

After fitting and transforming , the first five values of test set are

```
array([[ 0.24317482,  1.48632102,  1.63899828, -1.30401153, -1.36348759,
         2.29884866, -0.48533542, -0.61633185, -0.11788007,  0.69066248,
        -0.1874992 ],
       [-0.56971069,  0.49635289,  1.63899828,  0.76686439, -1.36348759,
        -0.43500036,  2.06043068,  1.62250254, -0.11788007, -1.44788523,
        -0.1874992 ],
       [-0.1403299 , -0.19881069, -0.61012877,  0.76686439, -1.36348759,
         2.29884866, -0.48533542, -0.61633185, -0.11788007,  0.69066248,
        -0.1874992 ],
       [-0.14919404, -0.25308752, -0.61012877,  0.76686439,  0.73341335,
        -0.43500036, -0.48533542, -0.61633185, -0.11788007,  0.69066248,
        -0.1874992 ],
       [ 0.80613386, -1.29781114,  1.63899828, -1.30401153, -1.36348759,
        -0.43500036,  2.06043068, -0.61633185, -0.11788007,  0.69066248,
        -0.1874992 ]])
```

The first five values of testing target are
`array([2, 1, 1, 1, 2]).`

7. Results and Evaluation

Jaccard indices of classification models are

```
Jaccard index for KNN Classification = 0.5
Jaccard index for Decision Tree Classification = 0.52
Jaccard index for SVM Classification = 0.55
Jaccard index for LogisticRegression = 0.55
```

F1 scores of classification models are

```
F1_score for KNN Classification = 0.61
F1_score for Decision Tree  Classification = 0.62
F1_score for SVM Classification = 0.63
F1_score for LogisticRegression = 0.63
```

Log Loss of Logistic Regression model is 0.68.

The overall performance of models in tabular form:

	Algorithm	Jaccard	F1-score	LogLoss
0	KNN	0.50	0.61	NA
1	Decision Tree	0.52	0.62	NA
2	SVM	0.55	0.63	NA
3	LogisticRegression	0.55	0.63	0.68

Logistic Regression model is more appropriate as the problem is fundamentally on binary classification.

8. Conclusion

In the data provided for this capstone, we can conclude that the severity of the collision is dependent on road conditions, weathe r conditions, light conditions and address type. Logistic Regress ion model is the best for classification of given data and to predi ct severity of collision.