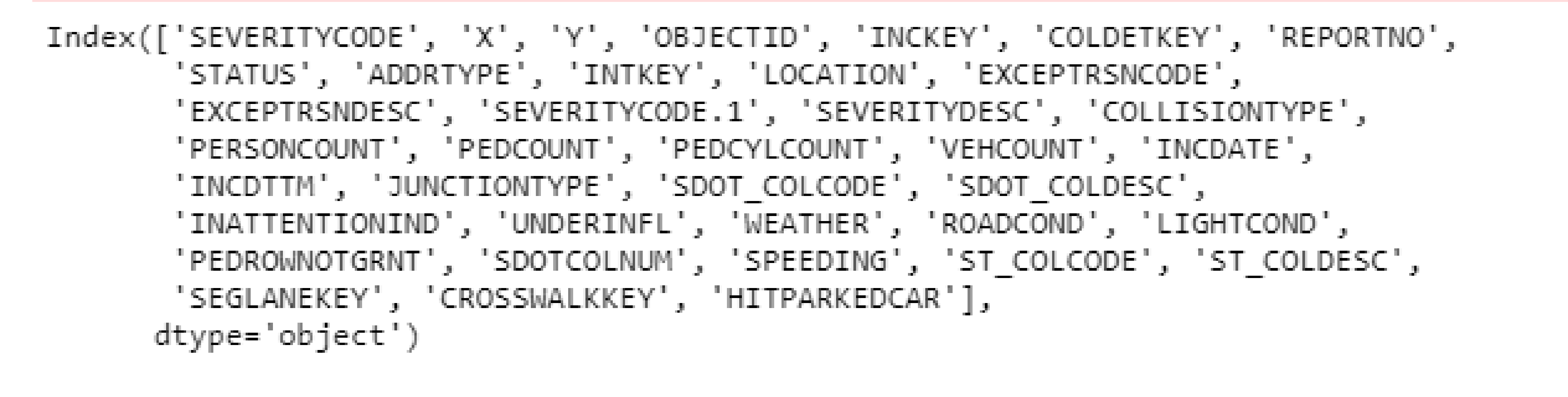
Predicting the severity of an Accident

Problem Statement

- In traffic situations, passengers at Seattle are prone to accidents on the roads. This can be due to different factors such as the weather conditions, the road conditions, the light conditions amongst other factors. It is highly recommended to be able to predict the severity of an accident based on the factors available to prepare for the casualty before the accident occurs.

Data Set

- The dataset provided for the Seattle city contains a total of 194673 observations and 37 attributes (relating to the accidents that occur on the road) with the labelled data (SEVERITYCODE) which describes the fatality of an incident. Given this dataset, the aim of this project is to select the necessary attributes that will be used to build a model that will help to predict the severity of an accident.
- The attributes are shown below



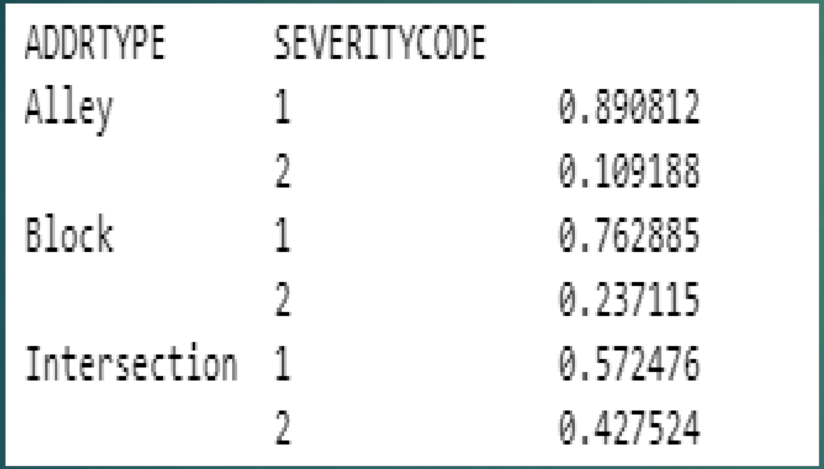
```
Index(['SEVERITYCODE', 'X', 'Y', 'OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO',
       'STATUS', 'ADDRTYPE', 'INTKEY', 'LOCATION', 'EXCEPTRSNCODE',
       'EXCEPTRSNDESC', 'SEVERITYCODE.1', 'SEVERITYDESC', 'COLLISIONTYPE',
       'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'INCDATE',
       'INCDTTM', 'JUNCTIONTYPE', 'SDOT_COLCODE', 'SDOT_COLDESC',
       'INATTENTIONIND', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND',
       'PEDROWNOTGRNT', 'SDOTCOLNUM', 'SPEEDING', 'ST_COLCODE', 'ST_COLDESC',
       'SEGLANEKEY', 'CROSSWALKKEY', 'HITPARKEDCAR'],
      dtype='object')
```

Data Preprocessing and Cleaning

- The selectd attributes are: 'ADDRTYPE', 'COLLISIONTYPE', 'JUNCTIONTYPE', 'WEATHER','ROADCOND','LIGHTCOND', 'PERSONCOUNT','PEDCOUNT', 'VEHCOUNT', 'HITPARKEDCAR’.
- The other attributes were dropped either because they do not relate to the target variable or because they have a lot of missing values.
- The following attributes are categorical values and needed to be changed to numerical values using the one hot encoding; 'ADDRTYPE', 'COLLISIONTYPE', 'JUNCTIONTYPE', 'WEATHER','ROADCOND','LIGHTCOND' while ‘HITPARKEDCAR’ was replace with 0 and 1 to represent its categorical values.

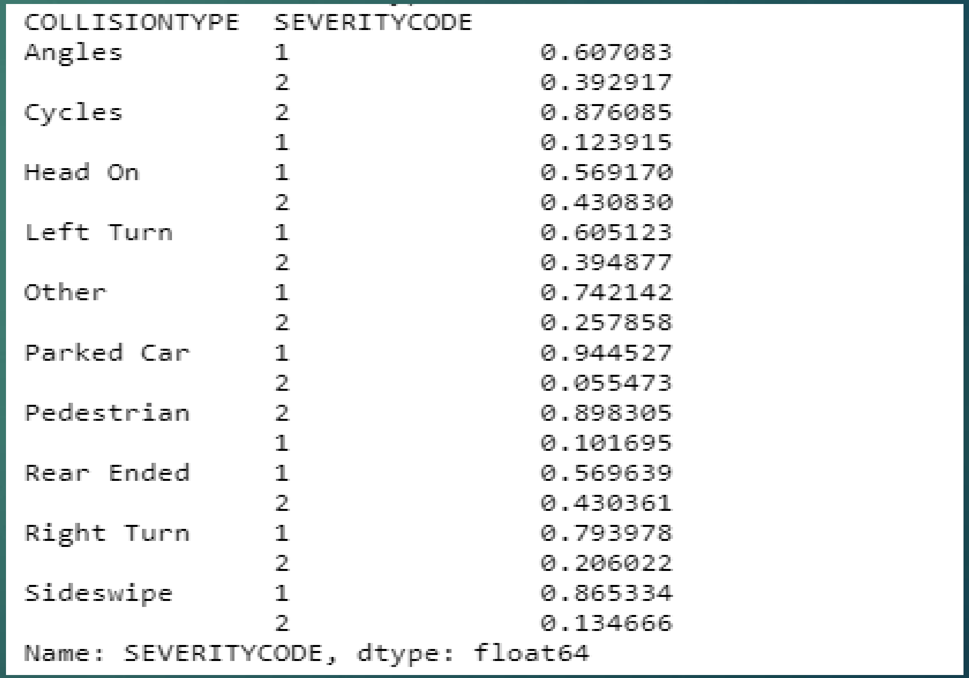
Exploratory Analysis

- **Relationship between ADDRTYPE and SEVERITYCODE**



```
ADDRTYPE      SEVERITYCODE
Alley         1               0.890812
              2               0.109188
Block         1               0.762885
              2               0.237115
Intersection  1               0.572476
              2               0.427524
```

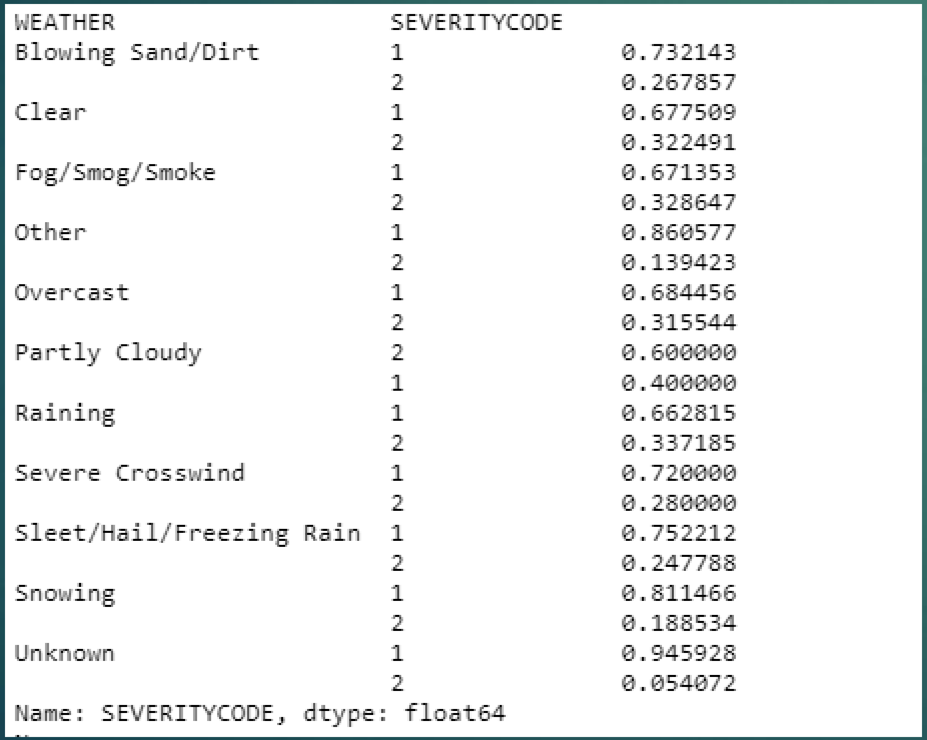
- **Relationship between COLLISIONTYPE and SEVERITYCODE**



```
COLLISIONTYPE  SEVERITYCODE
Angles         1               0.607083
               2               0.392917
Cycles         2               0.876085
               1               0.123915
Head On        1               0.569170
               2               0.430830
Left Turn      1               0.605123
               2               0.394877
Other          1               0.742142
               2               0.257858
Parked Car     1               0.944527
               2               0.055473
Pedestrian     2               0.898305
               1               0.101695
Rear Ended     1               0.569639
               2               0.430361
Right Turn     1               0.793978
               2               0.206022
Sideswipe      1               0.865334
               2               0.134666
Name: SEVERITYCODE, dtype: float64
```

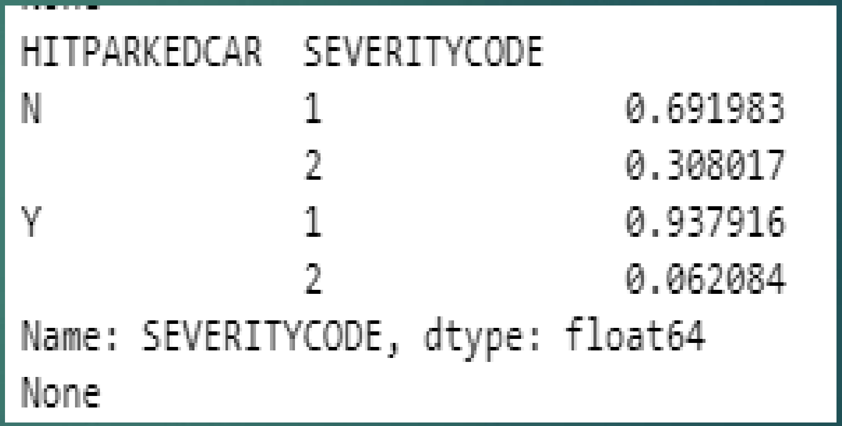
Exploratory Analysis

- **Relationship between WEATHER and SEVERITYCODE**



```
WEATHER                   SEVERITYCODE
Blowing Sand/Dirt         1                 0.732143
                          2                 0.267857
Clear                     1                 0.677509
                          2                 0.322491
Fog/Smog/Smoke            1                 0.671353
                          2                 0.328647
Other                     1                 0.860577
                          2                 0.139423
Overcast                  1                 0.684456
                          2                 0.315544
Partly Cloudy             2                 0.600000
                          1                 0.400000
Raining                   1                 0.662815
                          2                 0.337185
Severe Crosswind          1                 0.720000
                          2                 0.280000
Sleet/Hail/Freezing Rain  1                 0.752212
                          2                 0.247788
Snowing                   1                 0.811466
                          2                 0.188534
Unknown                   1                 0.945928
                          2                 0.054072
Name: SEVERITYCODE, dtype: float64
```

- **Relationship between HITPARKEDCAR and SEVERITYCODE**

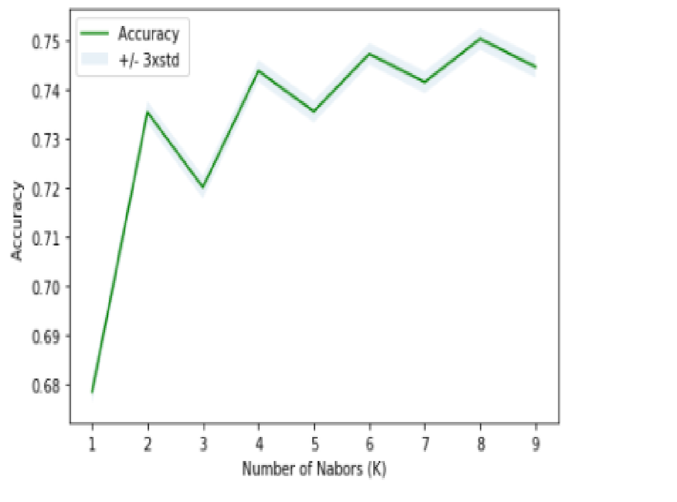


```
HITPARKEDCAR  SEVERITYCODE
N             1               0.691983
              2               0.308017
Y             1               0.937916
              2               0.062084
Name: SEVERITYCODE, dtype: float64
None
```

Predictive Modelling

The following algorithms were used:

1. **KNN**: The optimal number of neighbours that gave the best accuracy was 8. This was then used to build the model.



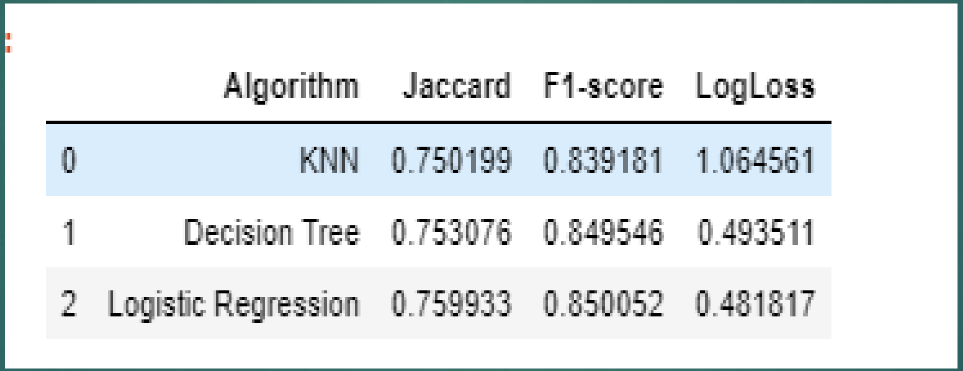
2. **Decision Tree**: The max depth used to build this model was 4 while the criterion selected was entropy as it is the most common.
3. **Logistic Regression**

- The following metrics were used to calculate the accuracy of each of the models used.

1. **Jaccard**: This indicates the similarity between two datasets, the predicted severity values dataset, and the actual severity value dataset. The higher the Jaccard value, the more accurate the model is said to be
2. **F1-score**: This metric conveys the balance between the recall and the precision. The higher the value, the more accurate the model is said to be.
3. **LogLoss**: This indicates the performance of a classifier where the predicted output is a probability value between 0 and 1. The classifier with a lower log loss has better accuracy.

Results

- From the table above it is observed that logistic regression has the highest Jaccard value, the highest F1-score as well as the lowest log loss value out of the other three machine learning algorithms. This implies that the logistic regression is the best model to use to predict the severity of the accident using the data provided.



	Algorithm	Jaccard	F1-score	LogLoss
0	KNN	0.750199	0.839181	1.064561
1	Decision Tree	0.753076	0.849546	0.493511
2	Logistic Regression	0.759933	0.850052	0.481817

Conclusion

- In this project, I outlined the attributes that tend to affect the severity code of an incident such as weather, road condition, address type just to mention a few.
- I developed different classification models to predict the severity code of an incident based on the attributes provided.
- The logistic regression model proved to be the best model in making this prediction.
- This prediction will be helpful for residents as well traffic attendants and paramedics to predict the severity of incidents and plan ahead in terms of providing medical attention and safety guidelines.