CAR ACCIDENT SEVERITY CAPSTONE PROJECT

Hari Sasongko

INTRODUCTION

- Car accidents in Seattle happen at all times,
- The increase in car ownership rates can lead to higher numbers of accidents on the road because of a simple probability

TARGET AUDIENCE

- The Seattle Administration
- Emergency Services
- Car Insurance Company

DATA

- The data comes from collision and accident reports in Seattle during the years 2004-present.
- focus on only four features, severity, weather conditions, road conditions, and light conditions, among others

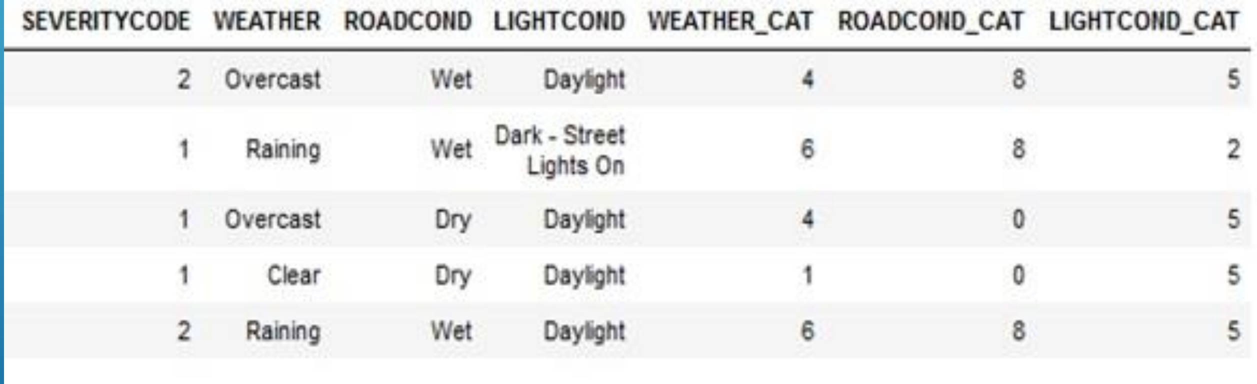
DATA

Severity Code as follow :

No	Description
1	Little to No Probability – Clear Condition
2	Very Low Probability – Chance or Property Damage
3	Low Probability – Chance of Injury
4	Mild Probability – Chance of Serious Injury
5	High Probability – Chance of Fatality

DATA

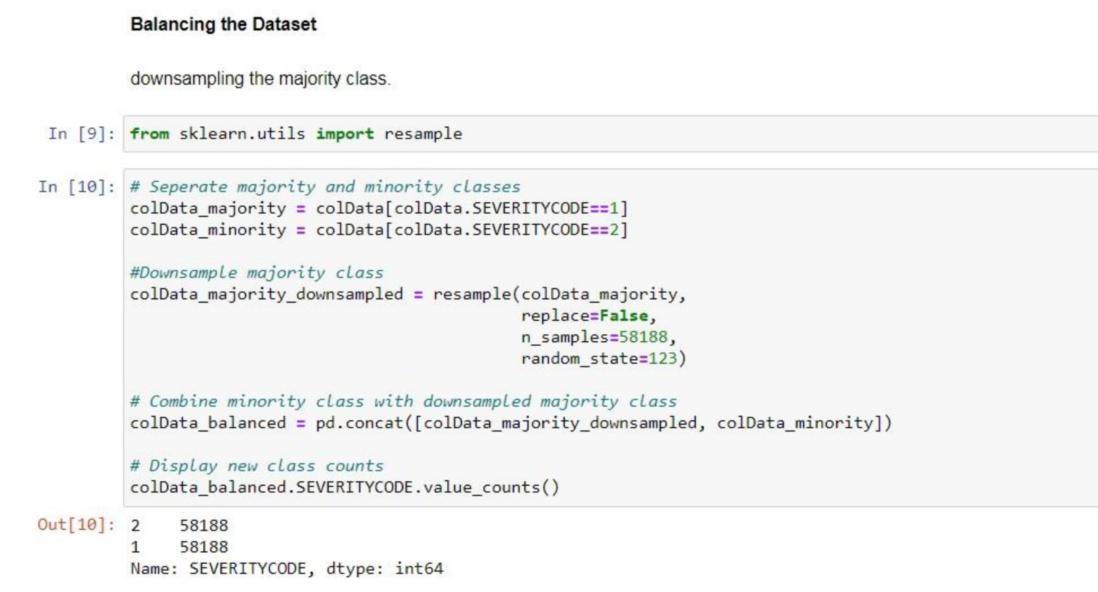
Convert data



SEVERITYCODE	WEATHER	ROADCOND	LIGHTCOND	WEATHER_CAT	ROADCOND_CAT	LIGHTCOND_CAT
2	Overcast	Wet	Daylight	4	8	5
1	Raining	Wet	Dark - Street Lights On	6	8	2
1	Overcast	Dry	Daylight	4	0	5
1	Clear	Dry	Daylight	1	0	5
2	Raining	Wet	Daylight	6	8	5

METHODOLOGY

Balancing Data



Balancing the Dataset

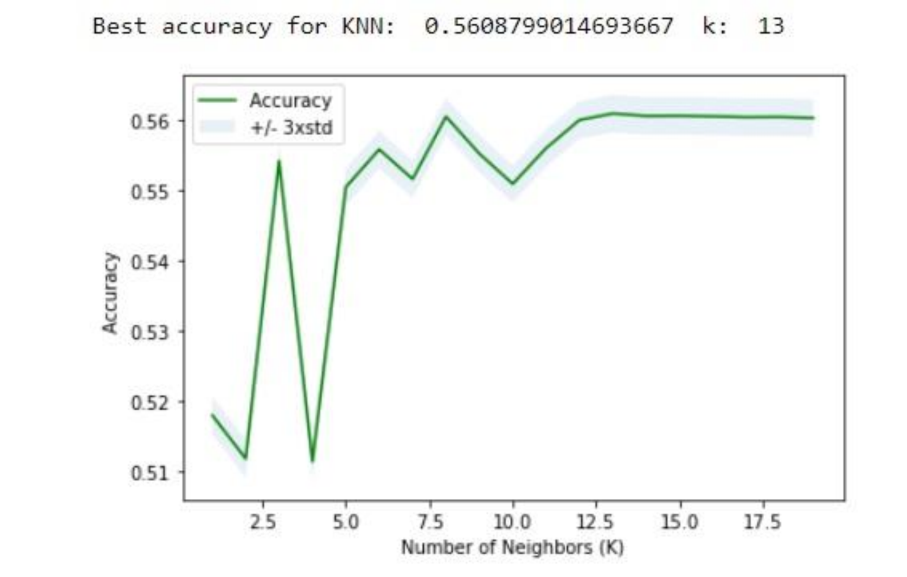
downsampling the majority class.

```
In [9]: from sklearn.utils import resample
```

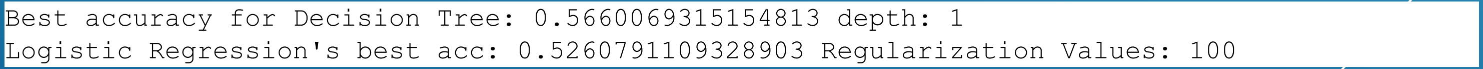
```
In [10]: # Seperate majority and minority classes
         colData_majority = colData[colData.SEVERITYCODE==1]
         colData_minority = colData[colData.SEVERITYCODE==2]

         #Downsample majority class
         colData_majority_downsampled = resample(colData_majority,
                                                 replace=False,
                                                 n_samples=58188,
                                                 random_state=123)

         # Combine minority class with downsampled majority class
         colData_balanced = pd.concat([colData_majority_downsampled, colData_minority])

         # Display new class counts
         colData_balanced.SEVERITYCODE.value_counts()

Out[10]: 2    58188
         1    58188
         Name: SEVERITYCODE, dtype: int64
```

BUILDING MODELS

All models are previously searched in the space of [1,25] for k, [1,10] for depth and [0.001,0.01,0.1,1,10,100] for regression in logistic regression.



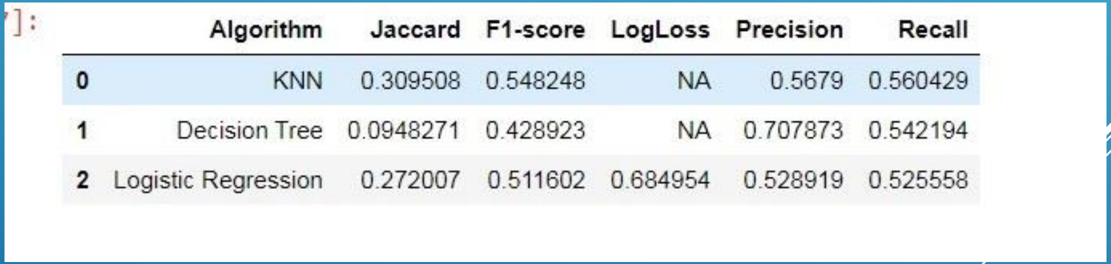
```
Best accuracy for KNN:  0.5608799014693667  k:  13
```



```
Best accuracy for Decision Tree: 0.5660069315154813 depth: 1
Logistic Regression's best acc: 0.5260791109328903 Regularization Values: 100
```

RESULT

In the results, it shows that among three machine learning methods, KNN excels other methods with only a small difference in precision. Although KNN provides the best performance evaluation, parameter tuning in KNN is computationally exhaustive.



]:

	Algorithm	Jaccard	F1-score	LogLoss	Precision	Recall
0	KNN	0.309508	0.548248	NA	0.5679	0.560429
1	Decision Tree	0.0948271	0.428923	NA	0.707873	0.542194
2	Logistic Regression	0.272007	0.511602	0.684954	0.528919	0.525558

CONCLUSION

- The f1-score is highest for k-Nearest Neighbor at 0.75. However, later when we compare the precision and recall for each of the model, we can see that the k-Nearest Neighbor model performs poorly in the precision of 1 at 0.08.
- The Decision Tree has a more balanced precision for 0 and 1
- Among environment conditions, weather is imperative to contribute in car collisions, while location of the collision, such as junctions, are exceptionally insignificant.

RECOMMENDATION

- The developmental body for Seattle city can assess how much of these accidents have occurred in a place where road or light conditions were not ideal for that specific area
- The car drivers could also use this data to assess when to take extra precautions on the road under the given circumstances of light condition, road condition and weather, in order to avoid a severe accident, if any.