Introduction

In this project, we will use the data to get information about the possibility of getting into a car accident and how severe it would be in the city of Seattle, US.

The collision of the vehicle's data will be analyzed in this project. Therefore, we will examine the correlation of accident severity with 3 factors which are weather, road, and light conditions. The prediction model can be used to inform people around Seattle, US. Besides, the data is provided by SPD (Seattle Police Department) and recorded by Traffic Records, and the timeframe of the data is 2004 to present.

Data

In total, we have 37 attributes but not all attributes are useful, so we need to decide what to keep. We will research the impact of environmental factors on the accidents. Based on the problem definition, we will use the data which are:

- The weather conditions during the time of the collision.
- The condition of the road during the collision.
- The light conditions during the collision.

Our target variable is ‘SEVERITYCODE’. The ‘SEVERITYCODE’ includes two types of collisions and these are:

1. Property Damage
2. Injury

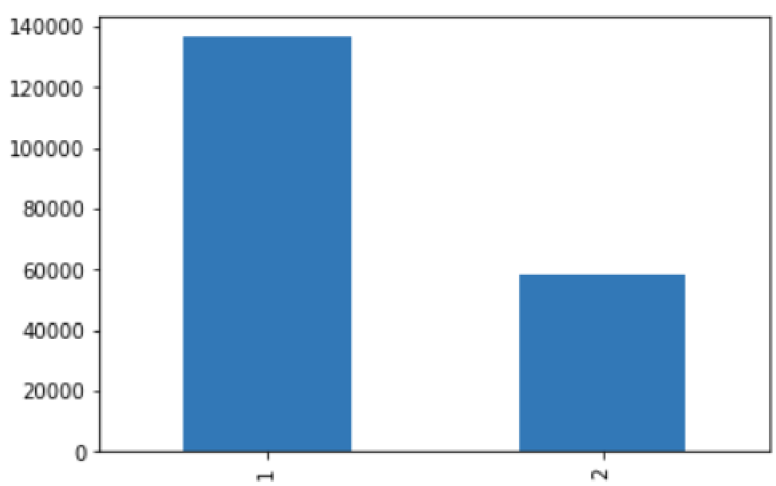

Figure 1: Severity of accidents before resampling

The figure above shows us the data is imbalanced, so we should balance the data to have accurate solutions. There are several techniques to balance data. In this project, the random under-sampling technique is used.

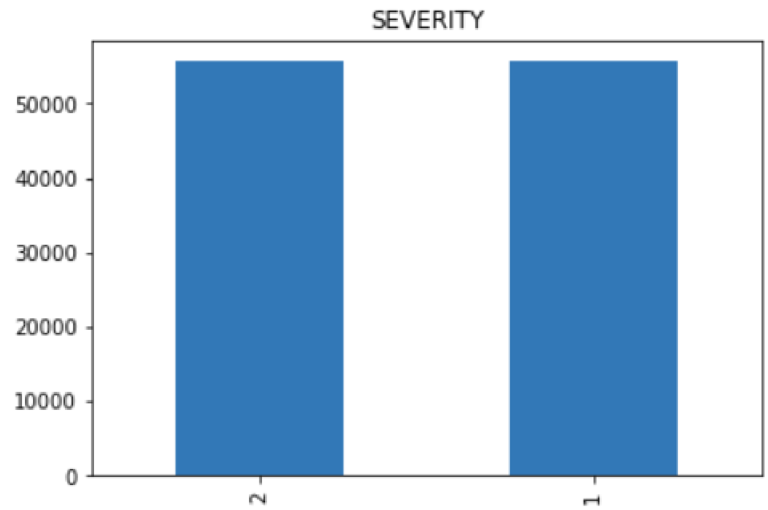

Figure 2: Severity of accidents after resampling

Methodology

The environmental conditions like ‘Weather Conditions’, ‘Road Conditions’, and ‘Light Conditions’ are the essential focus of this project. After importing the main libraries in python, the data is narrowed according to the environmental factors.

	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 3: Head of the data we worked on

Firstly, the correlation between the features is examined and the heatmap of this correlation is created with the help of the seaborn library (Figure 4).

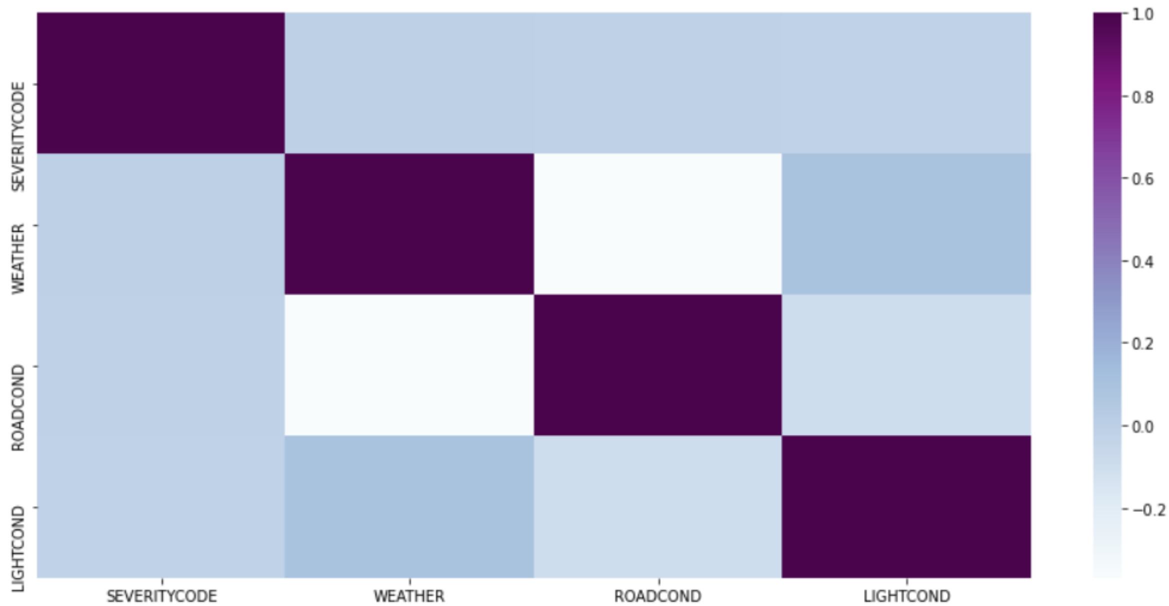

Figure 4: Correlation of selected features

Secondly, the dataset has lots of missing information, so the unknown information (NaN) is dropped from the data. The figure below shows that the number of missing values:

```
Number of NaN values for the column WEATHER: 5081
Number of NaN values for the column ROADCOND: 5012
Number of NaN values for the column LIGHTCOND: 5170
```

Figure 5: NaN values

Data Visualization

The number of accidents is plotted against each environmental feature (weather, road and light conditions) to see the effect of each factor clearly.

```
Clear                       108825
Raining                      32648
Overcast                     26923
Snowing                        825
Fog/Smog/Smoke                 553
Sleet/Hail/Freezing Rain       107
Blowing Sand/Dirt               46
Severe Crosswind                25
Partly Cloudy                    5
Name: WEATHER, dtype: int64
```

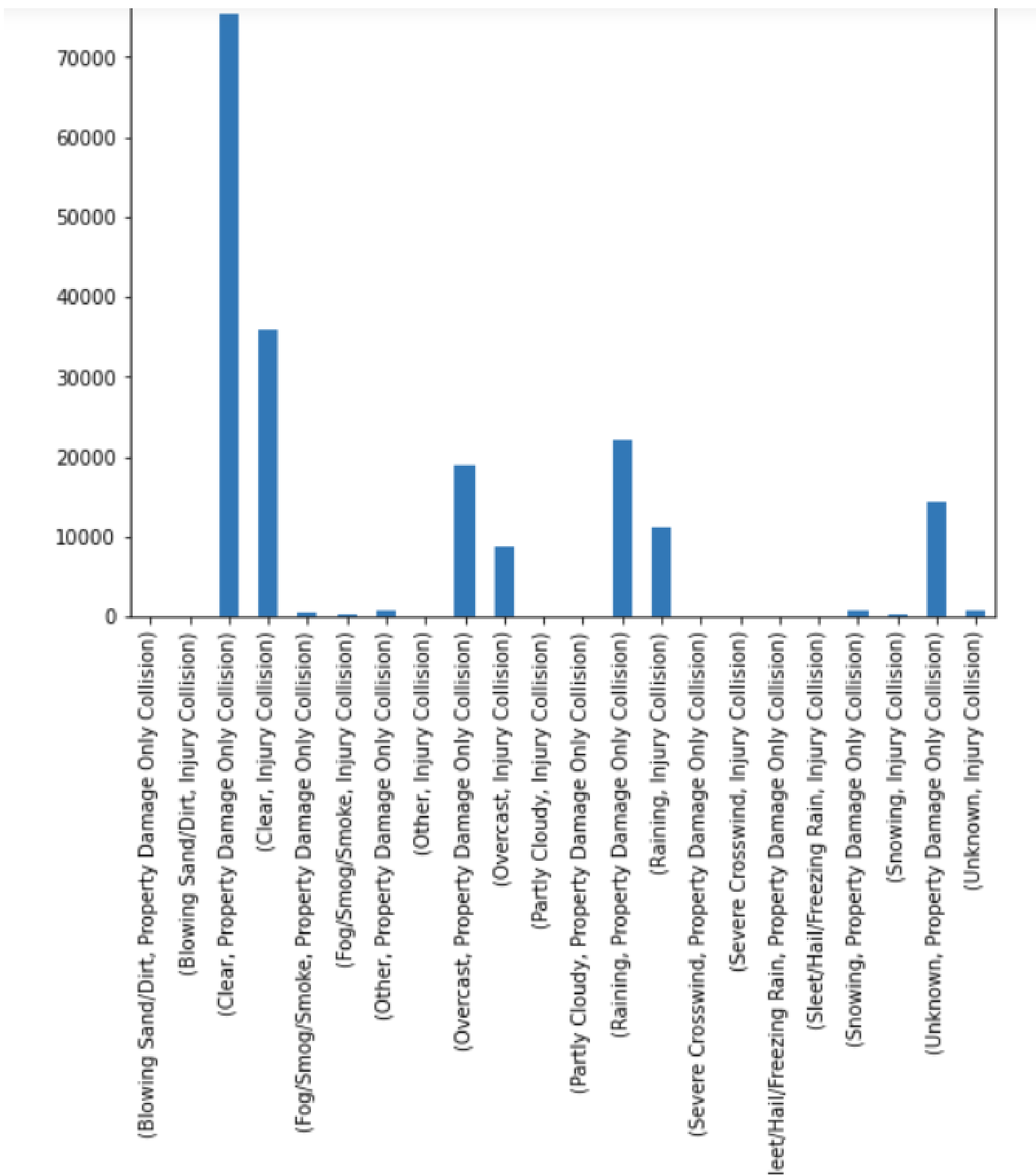

Figure 6: Weather conditions according to number of accidents

```
Dry                  121490
Wet                   46324
Ice                    1080
Snow/Slush              833
Standing Water          105
Sand/Mud/Dirt            65
Oil                      60
Name: ROADCOND, dtype: int64
```

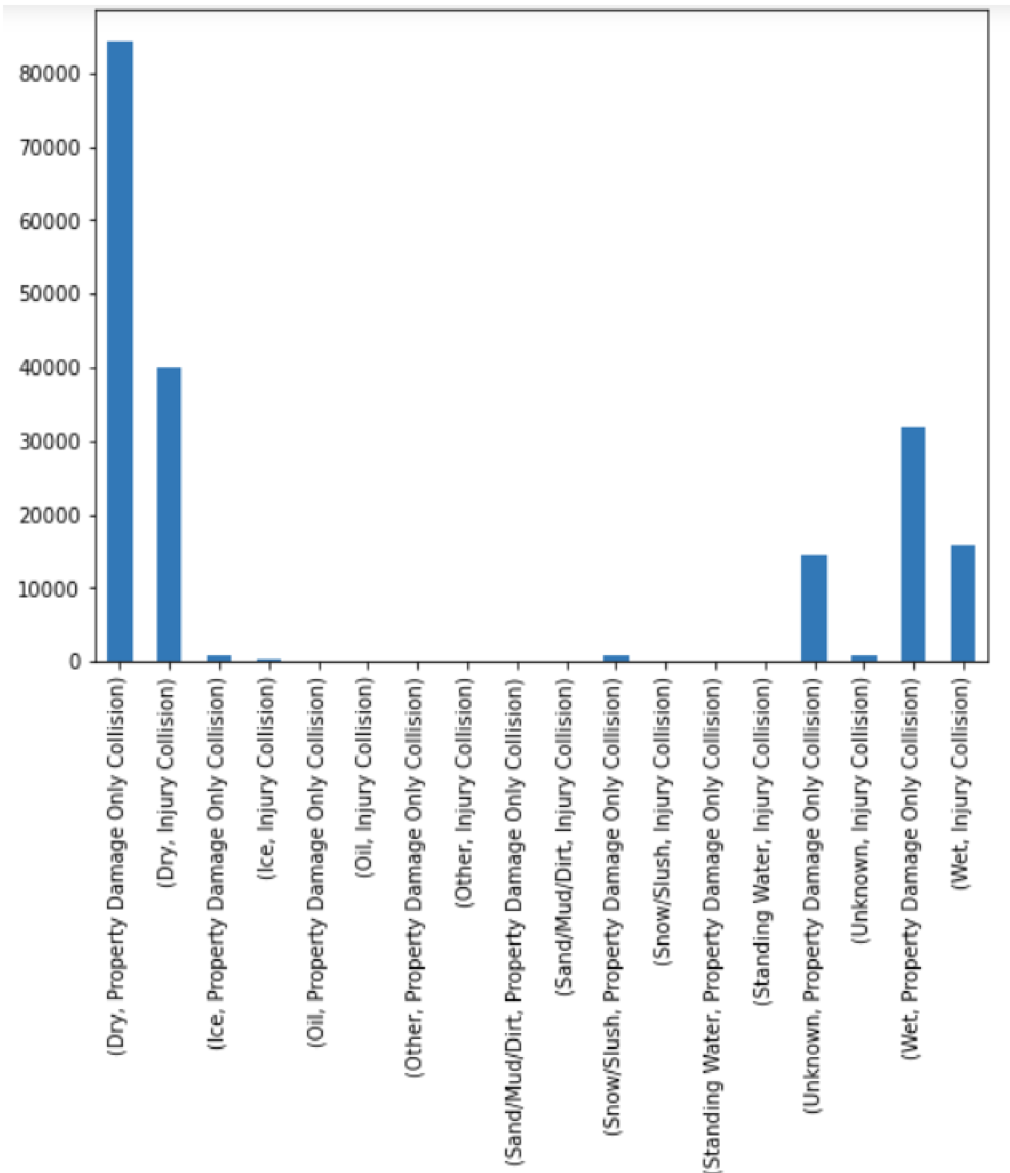

Figure 7: Road conditions according to number of accidents

```
Daylight                      112618
Dark - Street Lights On        46748
Dusk                            5648
Dawn                            2413
Dark - No Street Lights         1408
Dark - Street Lights Off        1114
Dark - Unknown Lighting            8
Name: LIGHTCOND, dtype: int64
```

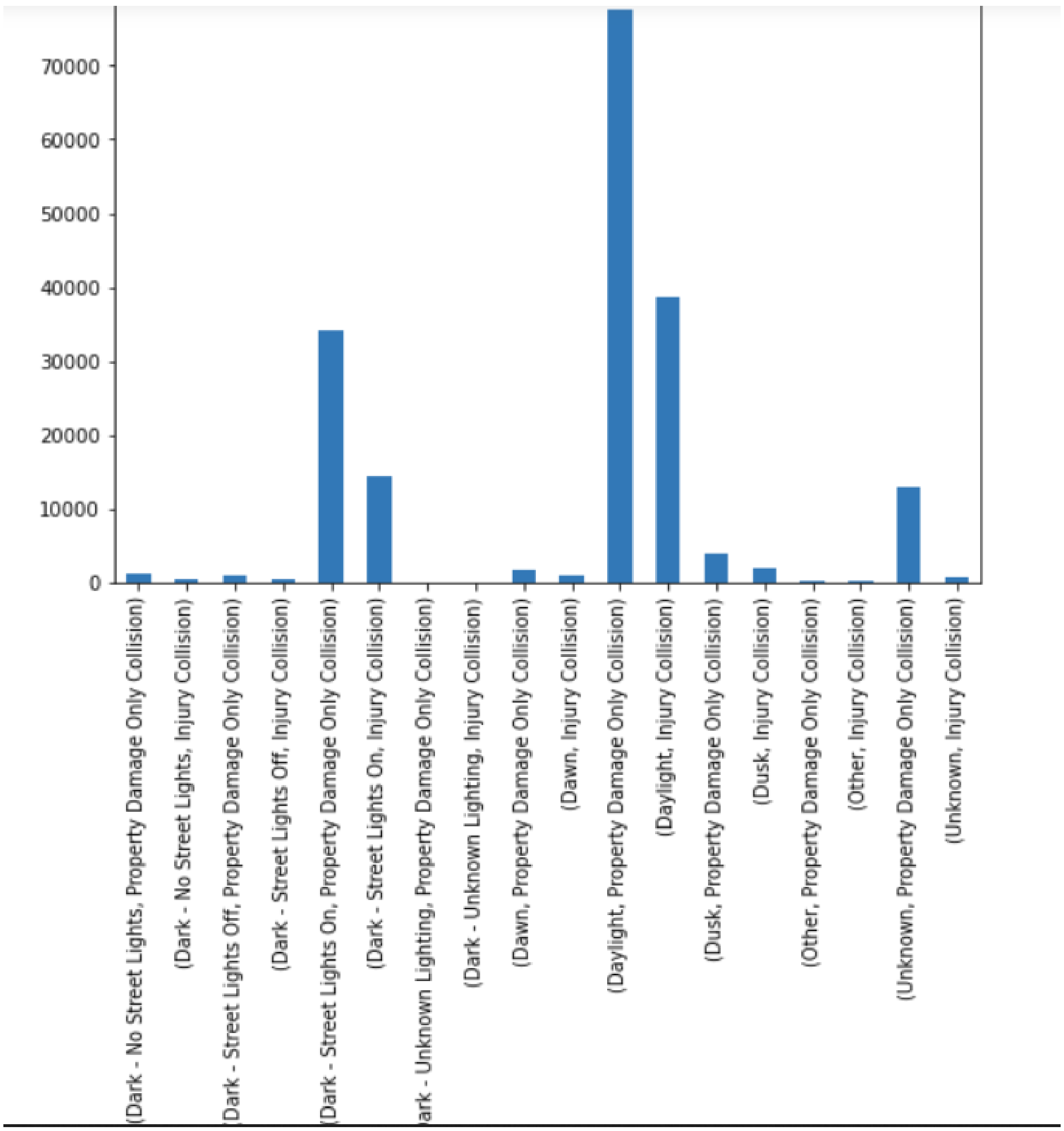

Figure 8: Light conditions according to number of accidents

As we can see in the figures above, contrary to expectations the accidents mostly happen in clear weather, dry roads, and daylight.

Results

The dataset is divided into training (70%) and testing (30%) samples after the cleaning and balancing of the data. The dataset is trained with different supervised machine learning methods which are K nearest neighbors (KNN), Decision Tree, Support Vector Machine (SVM), and Logistic Regression. The different classifier's results are indicated below:

K nearest neighbors (KNN)

	precision	recall	f1-score	support
1	0.53	0.31	0.39	16658
2	0.51	0.73	0.60	16752
accuracy			0.52	33410
macro avg	0.52	0.52	0.49	33410
weighted avg	0.52	0.52	0.49	33410

KNN's Accuracy: 0.5163723436096977

Decision Tree

	precision	recall	f1-score	support
1	0.53	0.34	0.41	16658
2	0.51	0.70	0.59	16752
accuracy			0.52	33410
macro avg	0.52	0.52	0.50	33410
weighted avg	0.52	0.52	0.50	33410

Decision Tree Acc. : 0.5186171804848848

Support Vector Machine (SVM)

	precision	recall	f1-score	support
1	0.53	0.27	0.36	16658
2	0.51	0.77	0.62	16752
accuracy			0.52	33410
macro avg	0.52	0.52	0.49	33410
weighted avg	0.52	0.52	0.49	33410

SVM's Accuracy: 0.5183178689015265

Logistic Regression

```
              precision    recall  f1-score   support

           1       0.52      0.34      0.41     16658
           2       0.51      0.70      0.59     16752

    accuracy                           0.52     33410
   macro avg       0.52      0.52      0.50     33410
weighted avg       0.52      0.52      0.50     33410

LogLoss: 0.6927051671267729
Logistic Regression's Accuracy: 0.5173900029931159
```

According to the tables above, all classifiers are performed almost the same, but the Decision Tree classifier has the highest accuracy compared to others (Figure 9).

	KNN	Decision Tree	SVM	Logistic Regression
Accuracy Score	0.511853	0.518617	0.518318	0.51739

Figure 9: Accuracy Score of Classifiers

Discussion

The dataset which has categorical values is converted by using label encoding to numerical values. After that, the data is balanced with a random under-sampling method to achieve more accurate results.

Therefore, the data is ready to utilize supervised machine learning techniques like KNN, Decision Tree, SVM, and Logistic Regression. The evaluation methods are shown in the result section.

Conclusion

The aim of the project is to examine the car accident data to evaluate the correlation between environmental factors and accident severity in Seattle, the US from 2004 to 2020. After several python operations, the data is cleaned and prepared to utilize ML techniques for the evaluation. The models are evaluated using different accuracy metrics. As we can see that most of the car accidents happened in clear weather, dry roads, and daylight.