FINAL REPORT

Introduction

The US Department of Transportation has announced a sum of 6,734,000 vehicle crashes happening the nation over, in 2018, bringing about roughly 1,900,000 wounds and 34,000 fatalities. These mishaps, and ensuing wounds, can be ascribed to a few elements, including occupied driving, the number of vehicles, speeding, and the number of individuals engaged with the mishap. Stakeholders of this project are Public Development Authority of Seattle and car drivers. By using recorded accident information, dissecting each revealed occurrence, and the properties related to each crash, a model can be made that can help caution drivers of expected dangers and may enable neighborhood governments to allot extra assets to forestall future mishaps.

Data

The dataset used for this project is based on car accidents which have taken place within Seattle from the year 2014-2020. This data is regarding the severity of each car accident along with the time and conditions under which each accident occurred. While there are 37 different factors associated with accidents, the project will use only the factors that has the most impact on accidents.

Methodology

The dataset contains 194,673 recorded mishaps, with 136,485 mishaps recognized as Severity Code 1, and 58,188 mishaps distinguished as Severity Code 2. Given the distinction in the number of cases in every classification, it is accepted that quite a few different elements, additionally revealed in this dataset, can be utilized to decide future mishap's degree of seriousness.

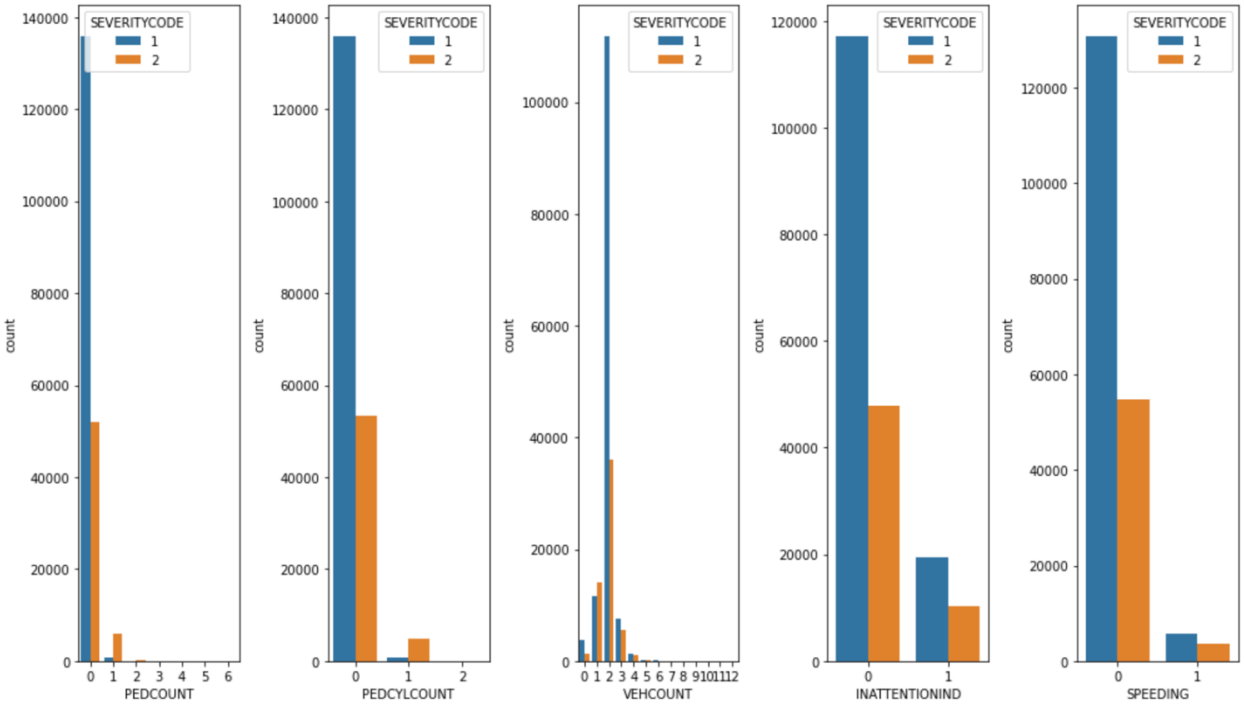
The initial step is investigating this dataset is to choose the accepted variables that drive the seriousness level. For this undertaking, this scientist accepted the accompanying elements assumed a significant job: pedestrian count, vehicle count, person count, pedestrian/cycle coun, inattention indicators, and speed. It was important to pare down the original dataset and eliminate the implicit highlights, and the outcome is the accompanying table:

	SEVERITYCODE	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	INATTENTIONIND	SPEEDING
0	2	2	0	0	2	NaN	NaN
1	1	2	0	0	2	NaN	NaN
2	1	4	0	0	3	NaN	NaN
3	1	3	0	0	3	NaN	NaN
4	2	2	0	0	2	NaN	NaN

Data cleaning was important, in any case as the characteristics of fragments Inattention Indicators, and Speeding was missing for certain records or was appeared as one of the various characteristics. Cleaning the characteristics brought about the accompanying table:

	SEVERITYCODE	PERSONCOUNT	PEDCOUNT	PEDCYLCOUNT	VEHCOUNT	INATTENTIONIND	SPEEDING
0	2	2	0	0	2	0	0
1	1	2	0	0	2	0	0
2	1	4	0	0	3	0	0
3	1	3	0	0	3	0	0
4	2	2	0	0	2	0	0

Because the data is now clean, we can start the analysis.



For each contributing element, the tallies of Severity Codes were determined during the exploratory investigation to recognize any disturbing elements. Moreover, AI models were made utilizing K-Nearest Neighbors, Decision Tree, and Logistic Regression, once the data was split into training and testing sets.

```
#Desision Tree
accidentTree=DecisionTreeClassifier(criterion='entropy',max_depth=4)
accidentTree.fit(X_trainset,y_trainset)
predTree=accidentTree.predict(X_testset)
Treef1=f1_score(y_testset,predTree,average='weighted')
Treeacc=accuracy_score(y_testset,predTree)
```

```
#K-Nearest Neighbours
KNN=KNeighborsClassifier(n_neighbors=4).fit(X_trainset,y_trainset)
predKNN=KNN.predict(X_testset)
KNNf1=f1_score(y_testset,predKNN,average='weighted')
KNNacc=accuracy_score(y_testset,predKNN)
```

```
#Logistic Regression
LR=LogisticRegression(C=0.01,solver='liblinear').fit(X_trainset,y_trainset)
predLR=LR.predict(X_testset)
LRf1=f1_score(y_testset,predLR,average='weighted')
LRacc=accuracy_score(y_testset,predLR)
```

```
#F1 and Accuracy Scores
score={
    'Model':['Decision Tree','KNN','Logistic Regression'],
    'F1 Score': [Treef1,KNNf1,LRf1],
    'Accuracy': [Treeacc,KNNacc,LRacc]
}

score=pd.DataFrame(score)
score
```

Results

The outcomes didn't affirm the speculation of this examination. As can be found in the abovementioned, there was no huge contrast in the frequency of increased severity among the expected elements. There were slight contrasts in factors including walkers notwithstanding, these outcomes were not huge.

Discussion

The Logistic Regression model was generally precise with an exactness of 74.82%, and an F1 score of 0.692, most likely because of the feeble relationship between's highlights of the dataset and accident severity. Extra models may not give any noteworthy enhancements.

	Model	F1 Score	Accuracy
0	Decision Tree	0.683253	0.747785
1	KNN	0.681430	0.732657
2	Logistic Regression	0.692335	0.748170

Conclusion

This report addresses the issue of foreseeing vehicle accident seriousness with an end goal to furnish the driver with advance information on conceivable mishap during their drive. Tragically, the connection couldn't be recognized.