Final Report

Name: **Accidents Severity Analysis and Prediction**

Project Type (Analysis) DSC 478
Team Members: Di Han, Wanshu Wang

1. Executive Summary

Nowadays the road accident has become a severe problem and is marked as the ninth prominent cause of death in the world. Traffic accident analysis is in high demand. This is a countrywide car accident dataset, which covers 49 states of the USA. The accident data are collected from February 2016 to Dec 2020. In this project, we analyze this dataset to predict the severity of an accident (Severity). This dataset is found at Kaggle.com. https://www.kaggle.com/sobhanmoosavi/us-accidents. The main variables are the severity of the accident, latitude, and longitude of the start point and endpoint, the address which includes a number of streets, the name of the street, city, county, and zip code. It also has information on weather, like temperature, wind chill, humidity, pressure, visibility, etc. The traffic situation is described as some variables, such as crossing, bump, railway, stop sign, traffic signal, station, etc. Which States and Cities have the most traffic accidents, what the most common weather conditions on the days of the accidents are, how many accidents have a severity level of 1, 2, 3, and 4, and what the most important features could affect the severity level, especially related to the level 3 and 4. We applied KNN and Decision Tree for the classification and evaluated the accuracy and recall of the results. We found that the Pruned Decision Tree performed the best accuracy. It could group Level 1 best, but for Level 3, and level 4 it is still not with good precision and recall. From the results of the pruned tree, we concluded the most important feature is the distance. Second important features are time-related features, like a year, day of the week. Following that are the weather variables like pressure, and etc., which also play important roles on the severity level. Visibility, which we may usually think of as an important point to the accident, is actually not very crucial. Pressure may affect people's mental situation, which affects their behaviors. For this point, we need to do further research in the future.

2. Analysis

2.1. Data Pre-processing

This dataset has 1,516,064 rows and 47 variables. There are 33 Discrete columns and 14 continuous columns. The size of the file is 907.1 Mb. It has numeric variables, including Distance, Visibility miles, Wind speed (mph), and Pressure. Most of the variables are nominal data, which are Severity, Bump, Amenity, No-Exit, Railway and etc. Some of them are binary data, showing if the location of the accident was with some features. The following graph is showing the percentage of each level of severity of accidents. We can see that most accidents are in 2 or 3 levels. And there are 3.07% in the 4th level (Fig1). Although the severest accidents are not a majority, they are important for us to evaluate them in advance and find out what features would affect these severe accidents most and to avoid that finally. We dropped some features with a high percentage of missing data and filled some data for some other features. Also, we transformed time features into a date, month, year and etc. According to the correlation matrix and the domain language, to avoid redundancy, we dropped some features about road condition, twilight information, and address information. We applied Synthetic Minority Oversampling Technique to balance the data, and eventually, the number of each severity level has been equaled. That would be beneficial for the clustering and classification, and we wouldn't have very high accuracy but overlooking the severity level 4 or 3. At last, before the classification and clustering method was applied, we normalized the numeric data and attained the reduced-feature dataset by PCA.

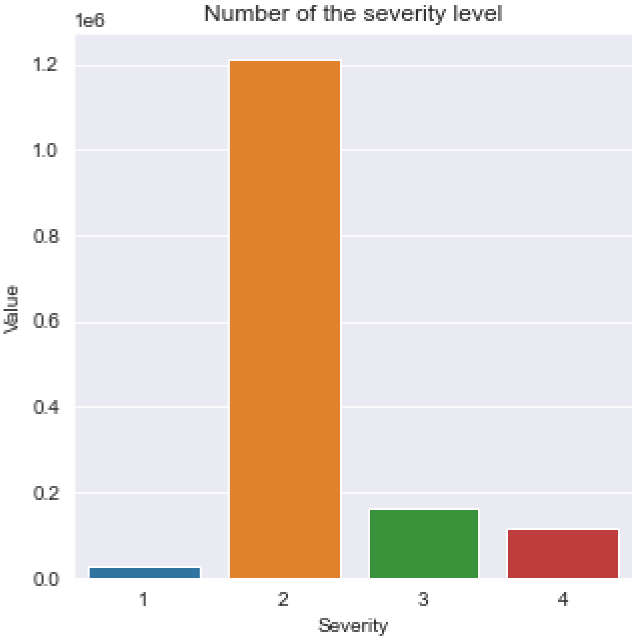


Fig 1 Number of the severity level

2.1.1. Missing Data

There are three features having over 30% missing data. We will drop or fill the missing data by researching those features. (Fig 2) We deleted Number because it has address information, we have many features including address information, like zip code, street, etc. We will drop more this kind of feature to avoid redundancy. For Precipitation, we don't want to simply fill in with the mean/median value because it is related to the humidity, visibility and other weather condition, pressure and many other weather-related features. Considering this point, we dropped this feature, because we have many other weather-related features which will provide this information. Similarly, we deleted the Wind_Chill(F). We fill Zipcode with the value which has the same pair of City and State pair and fill City with the value which has the same pair of Zipcode and State. For those who have no Zipcode and City value, we fill it with the most frequent in the state. For these features, Wind_Speed(mph), Humidity(%), Visibility(mi) Temperature(F), Wind_Direction, Pressure(in, we fill them with the mean value of the month in a certain city or state.

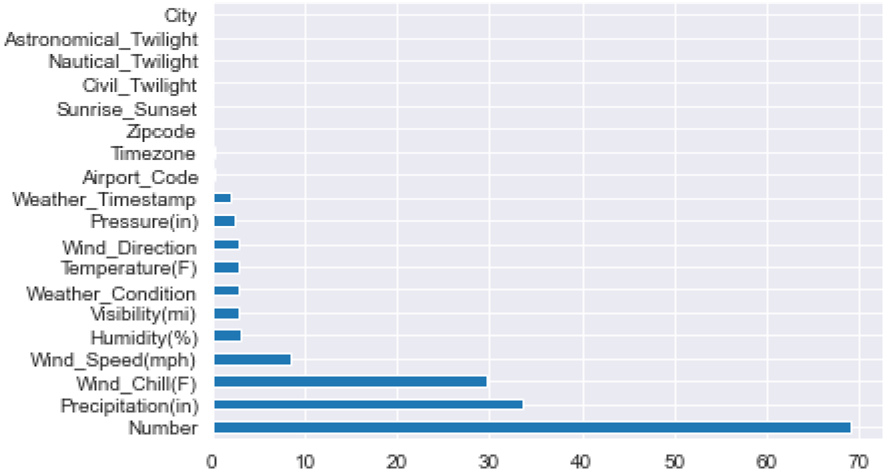


Fig 2 Percentage of Missing Data by Features

2.1.2. Data transformation/normalization

In this section, we extracted month, day of the week, date, and year information from Start_time. For the binary features like Amenity, bump and etc., we transformed their values from True or False into 1 or 0. Moreover, when we analyzed the weather conditions, we found there are lots of them, so it's better to reduce the number of unique conditions. replace them with a more generic description. After transformation, we have 11 categories, and we have Rain', 'Cloudy', 'Snow', 'Clear', 'Fog', 'Thunderstorm', 'Smoke', 'Windy', 'Hail', 'Sand', 'Tornado'. Similarly, we reduced the categories to wind direction. In wind direction, it has “CALM”, and “Calm”. They are supposed to be in the same category. So we transformed those data like this to have more accurate data. After transforming, we have 18 categories instead of 24. Next, we did dummy variables to Wind Direction, Weather Condition, and Sunrise_Sunset.

2.1.3. Feature Reduction

Besides, we dropped features with a high percentage of missing data, and some redundant features, for the preparation of classification and clustering we removed location information, like Start_Lng, Start_Lat, End_Lat, End_Lng, and Zipcode, the original Start_Time, End_Time. According to the correlation matrix (Fig), we can see that the start and end coordinates of the accidents are highly correlated. The end of the accident is usually close to the start, so we can consider just one of them for the machine learning models. Moreover, the wind chill (temperature) is directly proportional to the temperature, so we can also drop one of them. We can also see that the presence of a traffic signal is slightly correlated to the severity of an accident meaning that maybe traffic lights can help the traffic flow when an accident occurs. From the matrix, we can also note that we couldn't compute the covariance with Turning_Loop, and that's because it's always False. Finally, we have 53 features for the dataset. For the exploratory analysis, we still kept those features to analyze the relationship between the severity level, the number of accidents, and the state, city, weather condition, road condition.

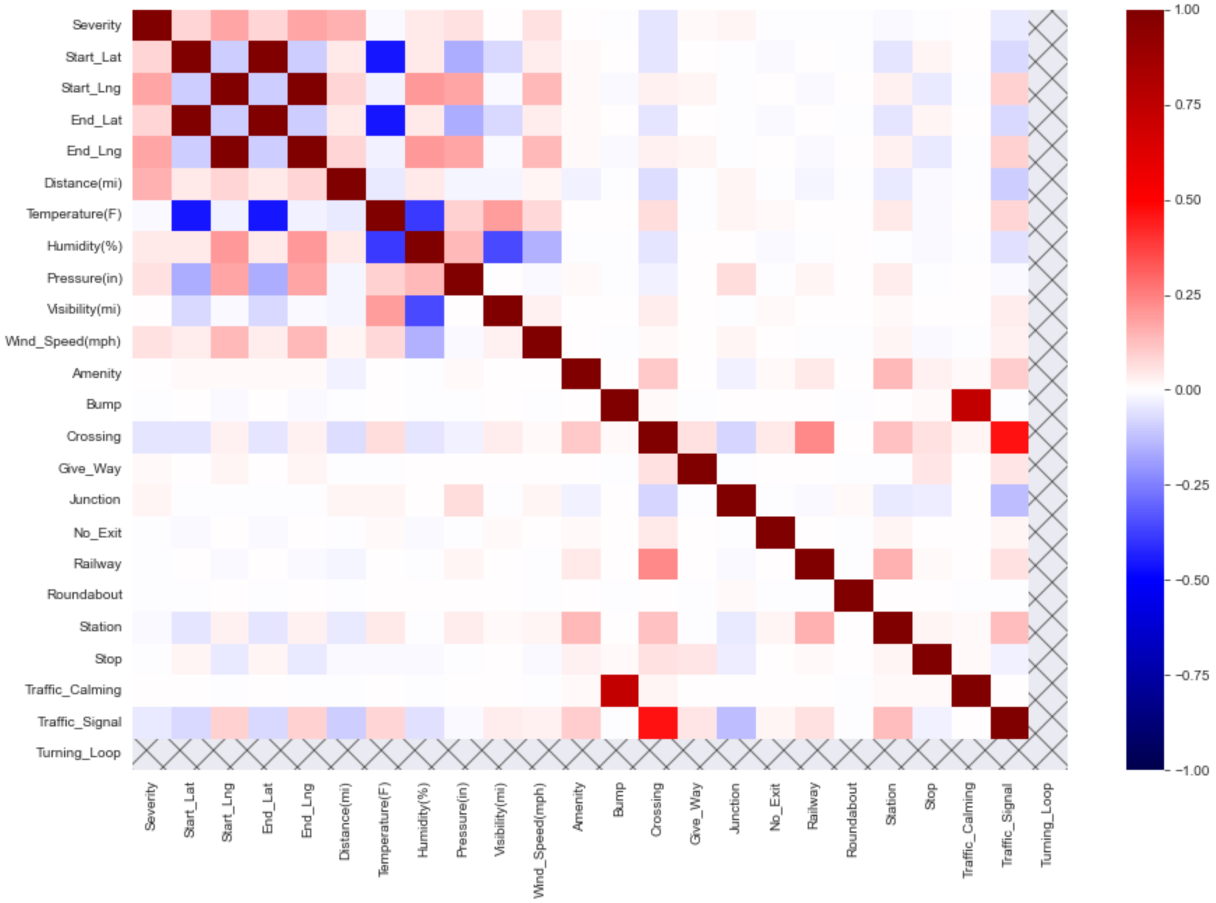


Fig 3 Correlation Matrix

2.1.4. Synthetic Minority Oversampling Technique

Due to the imbalance of this dataset mentioned in section 2.1, we used Synthetic Minority Oversampling Technique to balance the dataset. We used *imblearn* package and SMOTE method. The following is balanced for the four levels. There are 4849528 observations after SMOTE, compared to 1516064 as before.

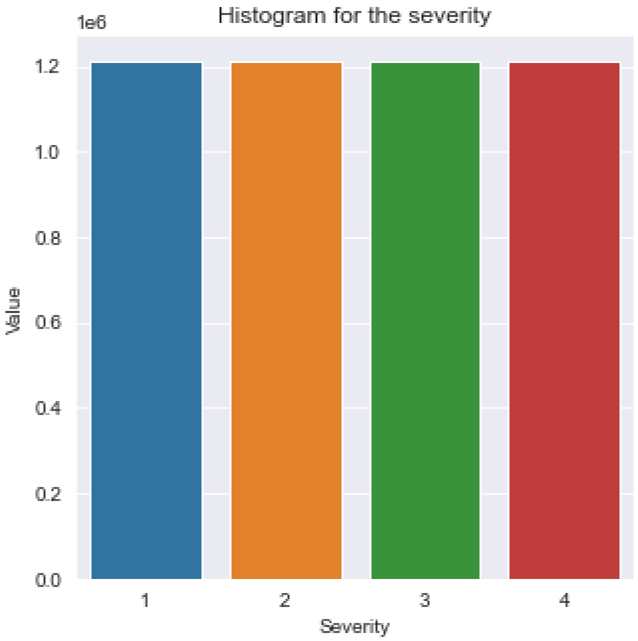


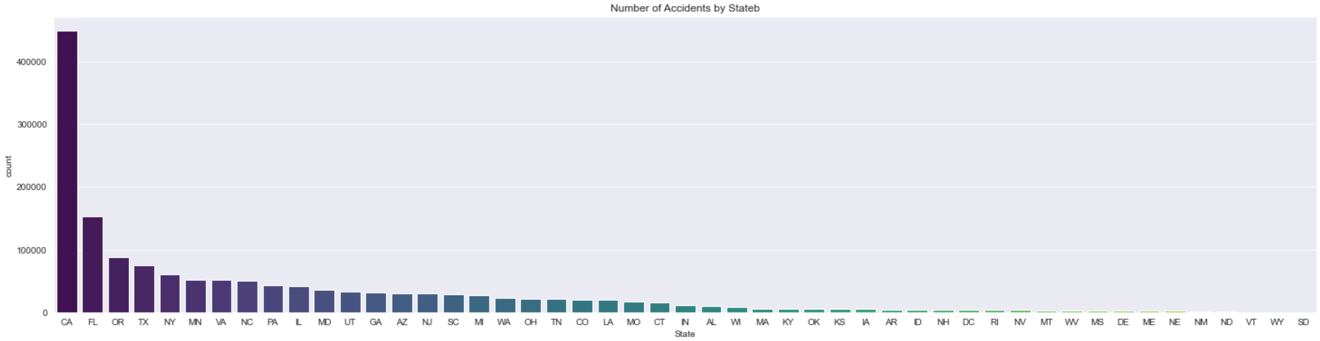
Fig 4 Number for Severity Levels

2.2. Data Exploratory Analysis

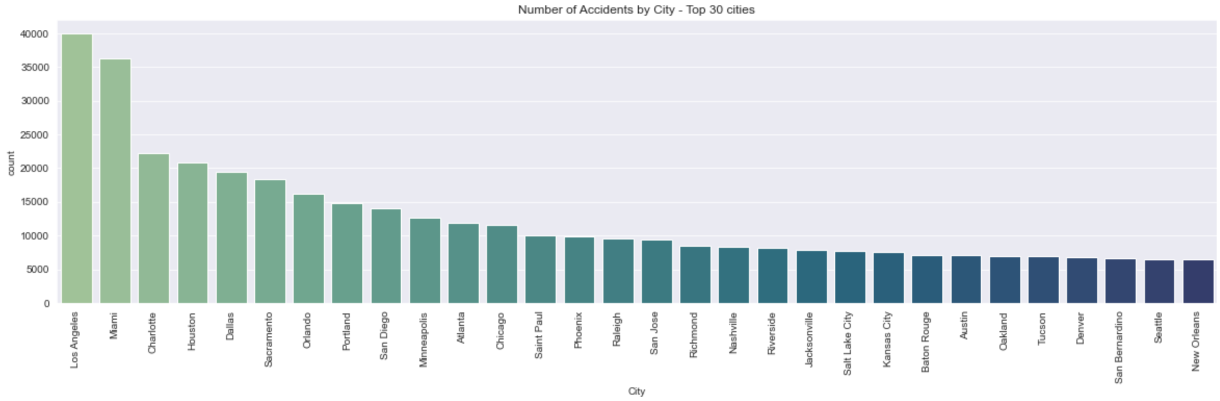
2.2.1. Analysis by State and City

a) Number of Accidents by State

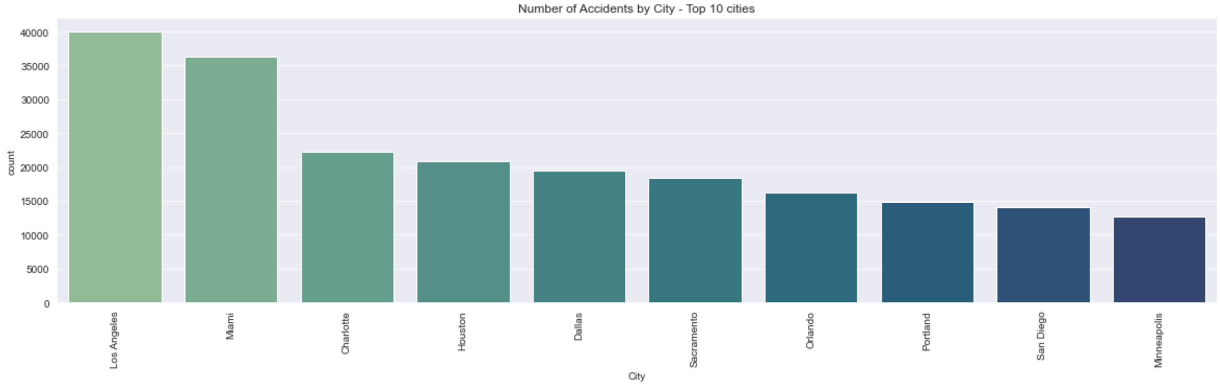
California(CA) is the most populated state, followed by Texas(TX) and Florida(FL), Oregon (OR) is the 3rd state with the most number of accidents. CA is almost two more times than the second state FL.



b) Number of Accidents by City

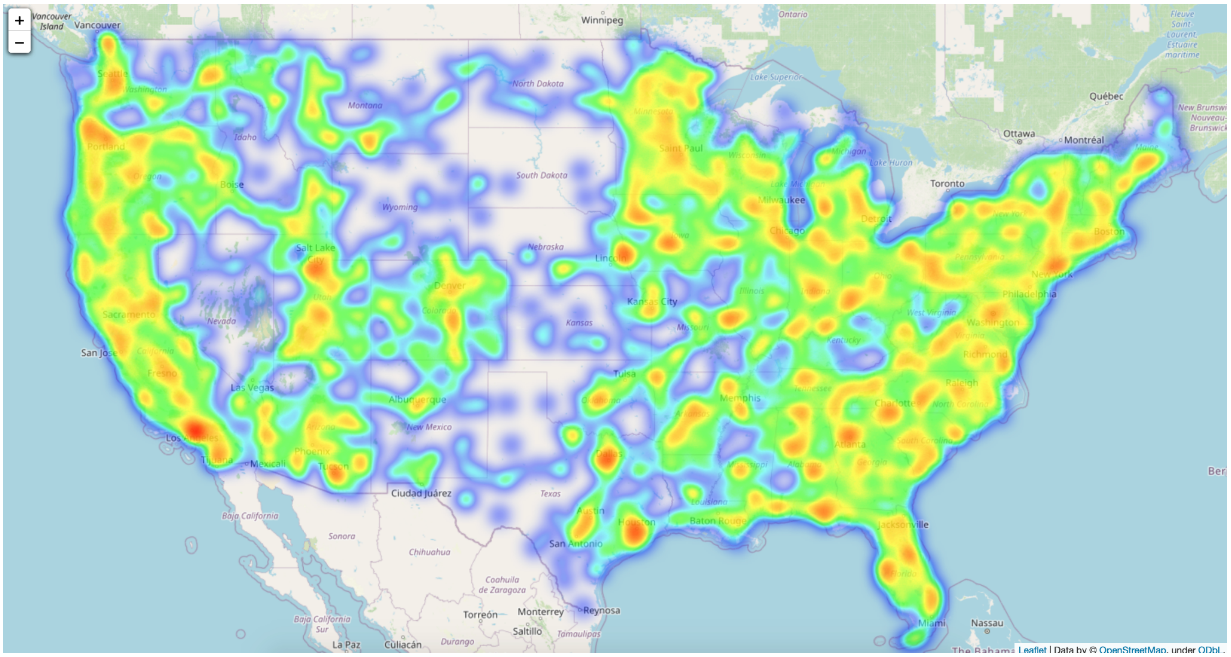


There are over 30 cities having over 5000 accidents. The top 5 cities are Los Angeles, Miami, Charlotte, Houston, and Dallas. The top 10 cities all have a number of accidents over 10000.



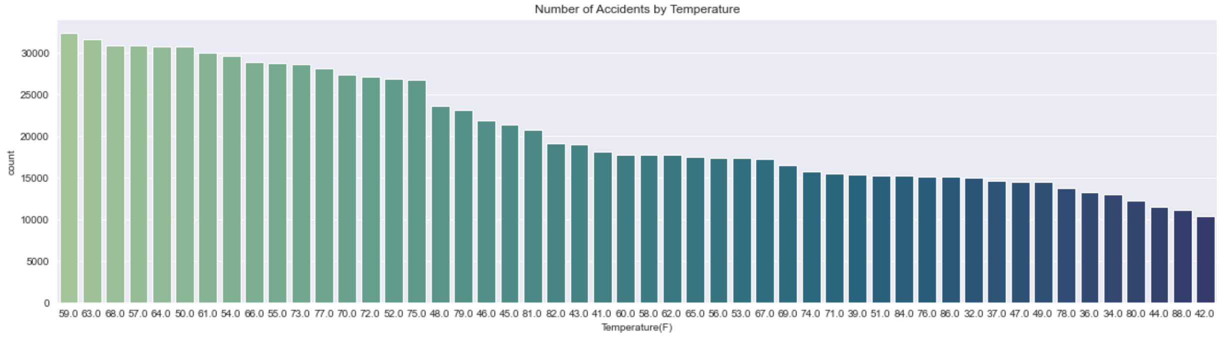
c) Number of the accidents distributed on a map

From the following heat map graph, we can see the higher number of accidents (red) that happened around coasts. In the central area, the number is very low.

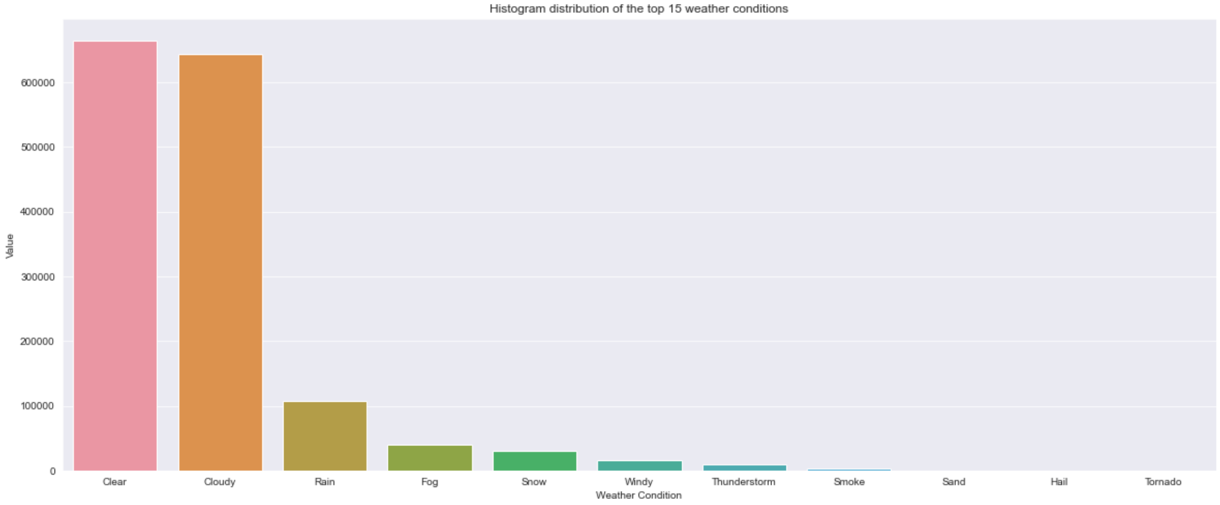


2.2.2. Analysis in Weather Conditions

Most accidents happen on the days with a Fairweather, follow by days Mostly Cloudy. Most accidents happen on days with temperatures between 50°F and 75°F (10°C and 23°C).



We can see the top 2 weather conditions are kindly good. They are Clear and Cloudy. None of them are extreme weather conditions, like rain, snow, or fog.



2.2.3. Clustering

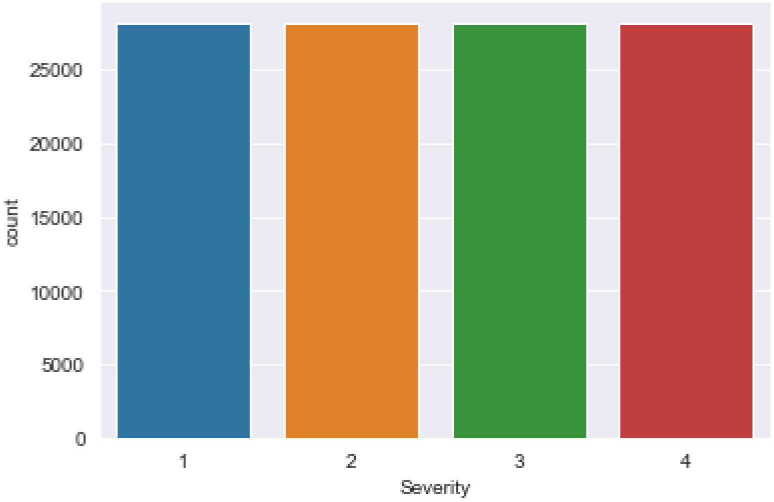
We used K-Means with four clusters for clustering to group the balanced dataset and compared the groups with the original groups with different severity levels. We used PCA for reducing features. The following are the results for clustering before PCA and after PCA. The completeness score after PCA is 0.014798401142677924, compare to the previous completeness score: 0.01474845550687712, which has improved. The homogeneity score is 0.01352069029332516, compare to the previous homogeneity score: 0.013482016769888854, it has improved. The clustering didn't perform very well. It is with bad completeness and homogeneity. After PCA, it had improved very slightly.

From our previous study, we know there is an important feature: source, which means who is responsible to record this accident record. The record source means that severity reported by different sources may differ in their underlying impact on traffic. This updated dataset in this project has been edited by removing many features including source because of some requests of the departments. Considering all these above, we may research further for the appropriate type for the accidents.

2.3. Supervised Learning

To predict the severity level, the methods for classification we applied are KNN, Decision Tree, and Naïve Bayes, Linear Discriminant Analysis, cross-validation for these methods and then compare their accuracy and recall.

At first, we tried to use *oversampling* balanced data to do the classification. However, due to the extremely slow speed, we failed. We afterward used an undersampling balanced data. The data size is 112,712 rows x 25 columns.



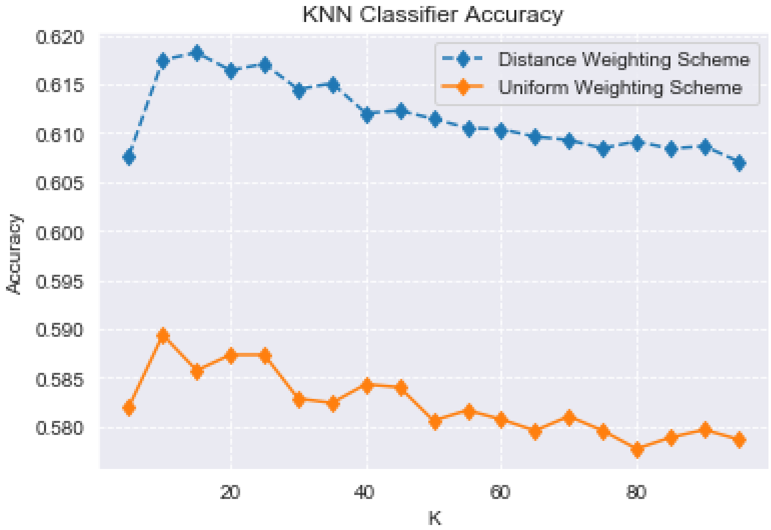
KNN method

We first normalize the data so that all attributes are on the same scale. Then we run the KNN classifier with K of 20.

The confusion matrix shows that class 1 has the best result. The precision is 0.71 and recall is 0.96. Class 2 and 4 have a precision value of 0.69 and 0.53 and recall values of 0.56 and 0.52, respectively. Level 3 has the worst result. Its precision is 0.49 and recall is only 0.42.

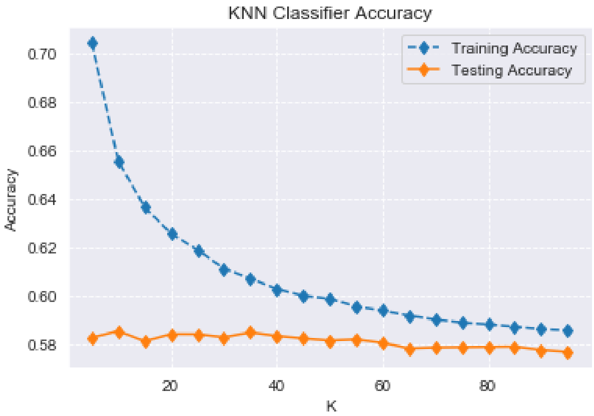
```
              precision    recall  f1-score   support

           1       0.71      0.96      0.82      5739
           2       0.69      0.56      0.62      5608
           3       0.49      0.42      0.45      5656
           4       0.53      0.52      0.52      5540

    accuracy                           0.62     22543
   macro avg       0.61      0.61      0.60     22543
weighted avg       0.61      0.62      0.60     22543
```

The knnclf score for the train set is 0.998 and for the test set is 0.616.
We next experimented with different values of K and the weight parameters. The result shows that the “best” K for this model is 10.

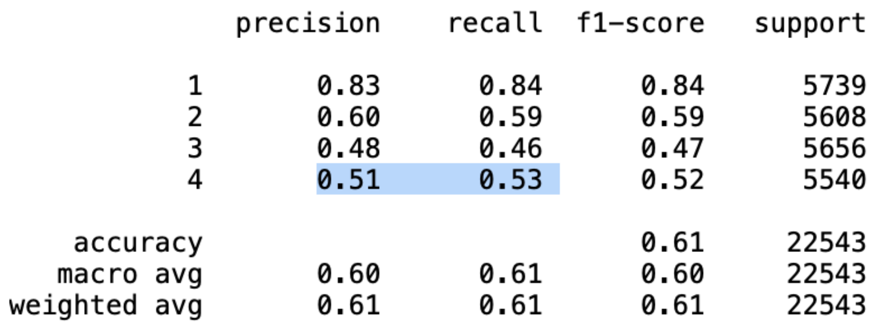


The accuracy for training is 0.656, and for testing is 0.585.

Next, we used only “uniform” weights, compare the accuracy of the KNN classifier across the different values of K on the training and the test data. We plotted training accuracy and testing accuracy in the following graph. When K is more than 50, the overfitting would disappear.



Then we used the non-normalized training and testing data to perform the decision tree classifier. And the confusion matrix is as follows:



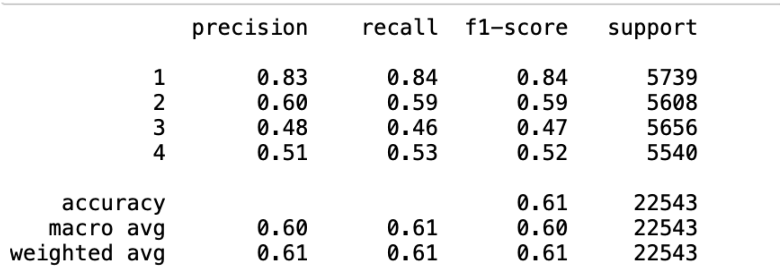
```
              precision    recall  f1-score   support

           1       0.83      0.84      0.84      5739
           2       0.60      0.59      0.59      5608
           3       0.48      0.46      0.47      5656
           4       0.51      0.53      0.52      5540

    accuracy                           0.61     22543
   macro avg       0.60      0.61      0.60     22543
weighted avg       0.61      0.61      0.61     22543
```

Compared to the KNN result, the precision of Level 1 has been improved, the recall for Level 4 and Level 3 were both better than KNN. The accuracy for training is 0.997 but the accuracy for testing is 0.607, which means there is overfitting for this model.

Then we applied Naïve Bayes and Linear Discriminant Analysis. For Naïve Bayes we have 0.529 for training and 0.533 for testing. For LDA, we have 0.545 for training and 0.541 for testing. For each of these methods, we performed 10-fold cross-validation on the 80% training data and reported the overall average accuracy as follows.

Method	Accuracy
Naïve Bayes	0.533
LDA	0.541
Decision Tree	0.607

According to this form, we can conclude that the Decision tree method has the highest accuracy. But after we tested the Decision tree model using testing data, we found that the evaluation results indicate overfitting. Pruning the tree may help in reducing overfitting. The accuracy of training is 0.99. The accuracy of testing is 0.59. Therefore, we would prune the tree to avoid it. We have the result for the pruned tree as follows. The precision and recall and accuracy are similar to the above, but the accuracy for testing is 0.611, compared to the accuracy of 0.61 on training, we concluded that there is no overfitting issue for this model.



```
              precision    recall  f1-score   support

           1       0.83      0.84      0.84      5739
           2       0.60      0.59      0.59      5608
           3       0.48      0.46      0.47      5656
           4       0.51      0.53      0.52      5540

    accuracy                           0.61     22543
   macro avg       0.60      0.61      0.60     22543
weighted avg       0.61      0.61      0.61     22543
```

We found that the Pruned Decision Tree performed the best accuracy. It could group Level 1 best, but for Level 3, and level 4 it is still not with good precision and recall. From the results of the pruned tree, we concluded the most important feature is the distance. Second important features are time-related features, like a year, day of the week. Following that are the weather variables like pressure, and etc., which also play important roles on the severity level. Visibility, which we may usually think of as an important point to the accident, is actually not very crucial. Pressure may affect people's mental situation, which affects their behaviors. For this point, we need to do further research in the future.

3. Work Distribution

We split pre-processing tasks. Wanshu handled the missing data and feature reduction. Di did data transformation and normalization. Wanshu also analyzed for Data Exploratory and clustering. Di applied Classification for severity prediction and evaluated the results. They finalized the final deliverables together.

References

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. "A Countrywide Traffic Accident Dataset.", 2019.

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.