UE18CS312
Data Analytics

Analysis on the Impact of Weather and Road Conditions on traffic accidents in USA

Pavan A	PES1201800157	pavanappanna7@gmail.com
Harsha Kulkarni	PES1201802000	kulkarniharsha14@gmail.com
Rahul Kata	PES1201802018	katarahul@gmail.com

➢ Data Description:

The countrywide car accident dataset collected from February 2016 to June 2020, which covers 49 states of the USA. The road and weather conditions during the time of the accident and also the location of the accident are recorded. The severity of the accident is also recorded on a scale of 1 to 4.

- **Shape** = 3513617 x 49
- Total unique records(**rows**) = 3513616
- Total Attributes(**columns**) = 49
- Target column is **Severity**, which is of *ordinal data* type.

```
'ID', 'Source', 'TMC', 'Severity', 'Start_Time', 'End_Time',
'Start_Lat', 'Start_Lng', 'End_Lat', 'End_Lng', 'Distance(mi)',
'Description', 'Number', 'Street', 'Side', 'City', 'County', 'State',
'Zipcode', 'Country', 'Timezone', 'Airport_Code', 'Weather_Timestamp',
'Temperature(F)', 'Wind_Chill(F)', 'Humidity(%)', 'Pressure(in)',
'Visibility(mi)', 'Wind_Direction', 'Wind_Speed(mph)',
'Precipitation(in)', 'Weather_Condition', 'Amenity', 'Bump', 'Crossing',
'Give_Way', 'Junction', 'No_Exit', 'Railway', 'Roundabout', 'Station',
'Stop', 'Traffic_Calming', 'Traffic_Signal', 'Turning_Loop',
'Sunrise_Sunset', 'Civil_Twilight', 'Nautical_Twilight',
'Astronomical_Twilight'],
```

➢ Missing values:

23 columns out of 49 columns have **missing values** in them, as shown below. (which is a part of the code). The importance of the columns with missing values are gauged with respect to the goals we have put forth. Based on this importance we have dropped the following columns:

- > End_Lat, End_Lng (Start_Lat, Start_lang are more than enough to visualize the location of the accident and usually the accident).
- > Number(street number).
- > TMC - Traffic Message Channel (TMC) codes are difficult to map and interpret.
- > Civil_Twilight, Nautical_Twilight,Astronomical_Twilight - Sunrise_Sunset is enough to know whether it's a day/night.
- > Timezone, Airport_Code .
- > Precipitation(in) - >50% missing values.
- > Wind_Chill(F) - >50% missing values.

```
The dataframe has 49 columns.
There are 23 columns that have missing values.
```

	Missing Values	% of Total Values
End_Lat	2478818	70.5
End_Lng	2478818	70.5
Number	2262864	64.4
Precipitation(in)	2025874	57.7
Wind_Chill(F)	1868249	53.2
TMC	1034799	29.5
Wind_Speed(mph)	454609	12.9
Weather_Condition	76138	2.2
Visibility(mi)	75856	2.2
Humidity(%)	69687	2.0
Temperature(F)	65732	1.9
Wind_Direction	58874	1.7
Pressure(in)	55882	1.6
Weather_Timestamp	43323	1.2
Airport_Code	6758	0.2
Timezone	3880	0.1
Zipcode	1069	0.0
Sunrise_Sunset	115	0.0
Civil_Twilight	115	0.0
Nautical_Twilight	115	0.0
Astronomical_Twilight	115	0.0
City	112	0.0
Description	1	0.0

Mean Imputation is done for *temperature(f), pressure(in) and humidity(%)* because of low standard deviation and less number of outliers.

Median imputation is done for *visibility(mi)* due to a large number of outliers.

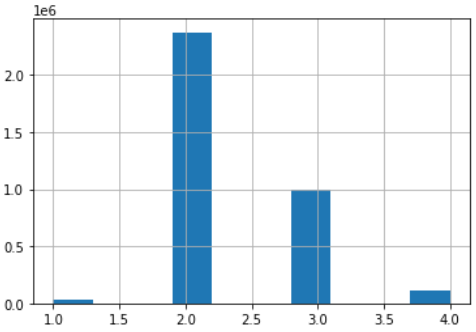
After dropping the above specified columns and a few other non relevant columns, the dataset has the following features:

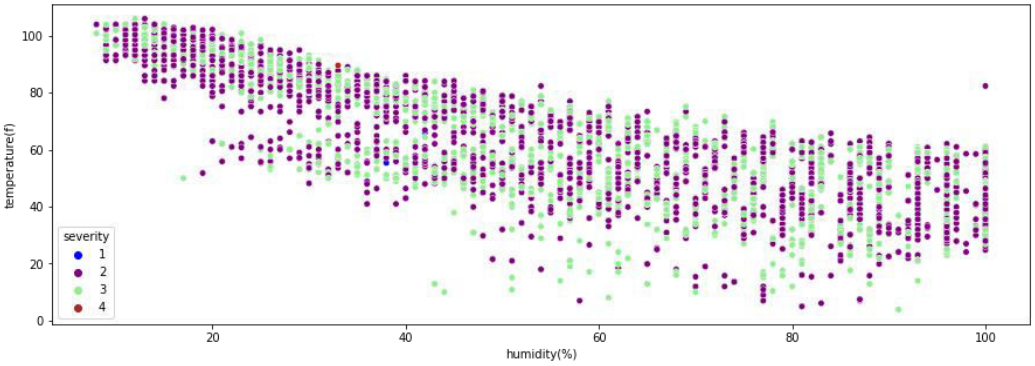
```
'id', 'source', 'severity', 'start_time', 'end_time', 'start_lat',
'start_lng', 'description', 'side', 'city', 'state', 'temperature(f)'
'humidity(%)', 'pressure(in)', 'visibility(mi)', 'wind_direction',
'wind_speed(mph)', 'weather_condition', 'amenity', 'bump', 'crossing'
'give_way', 'junction', 'no_exit', 'railway', 'roundabout', 'station'
'stop', 'traffic_calming', 'traffic_signal', 'sunrise_sunset'],
```

➢ Outliers:

There are only 7 numerical type columns after dropping the irrelevant columns. Visually determining the outliers using a boxplot is impossible due to the sheer volume of data. Hence we have used *(Q3 + 1.5IQR)* and *(Q1 - 1.5IQR)* as upper and lower bound to filter out the values that lie out of that range. The table clearly shows how many outliers are present and the outlier percentage. It is observed that the number is very high and hence no immediate action is taken to eliminate the outlier. We intend to create a baseline model and then eliminate the outliers and validate if there's a significant improvement in the model.

	#_Outliers	Outliers %
visibility(mi)	701179	19.956704
pressure(in)	395638	11.260506
wind_speed(mph)	48281	1.374156
temperature(f)	26075	0.742137
humidity(%)	0	0.000000
start_lng	0	0.000000
start_lat	0	0.000000





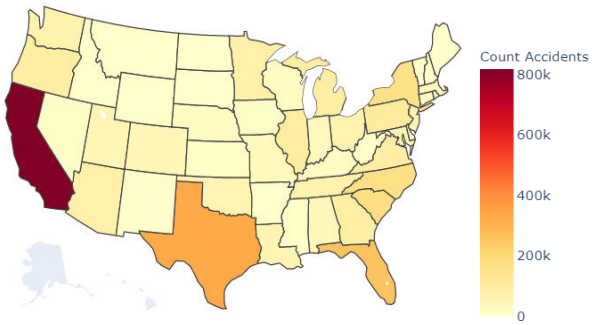
➢ Initial insights:

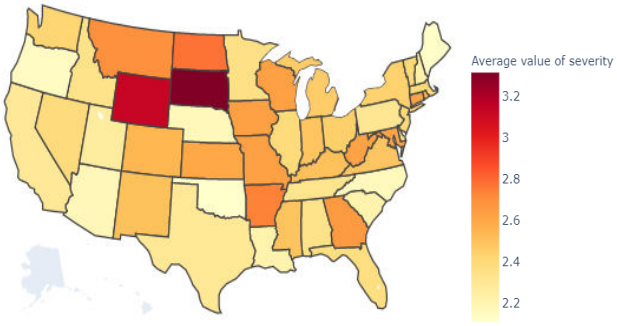
Fig: 1 [severity]

Fig: 2 [temperature - humidity]

Fig: 1 shows that there are more data points with a severity rating of 2 when compared to other ratings. Thus this class imbalance has to be dealt with in the future while building the model by resampling the dataset.

Fig 2 shows the scatter plot of *temperature* vs *humidity* with severity index. The two are negatively correlated to each other and show no formation of clusters as such. Further sections such as *Correlation* and *PCA* provide further insights on correlation type and cluster absence.





From the above two plots it is evident that although the number of accidents are more in the extreme west (California) and extreme south(Texas), the severity of accidents in the northern region(Wyoming, Dakota, Montana) is higher.

➢ Handling Categorical Data:

The following columns are binary in nature with true/false. We have encoded the records in these columns with **true = 1** and **false = 0**.

- > Amenity
- > Bump
- > Crossing
- > Give_way
- > Junction
- > No_exit
- > Railway
- > Roundabout
- > Station
- > Stop
- > traffic_calming
- > traffic_signal

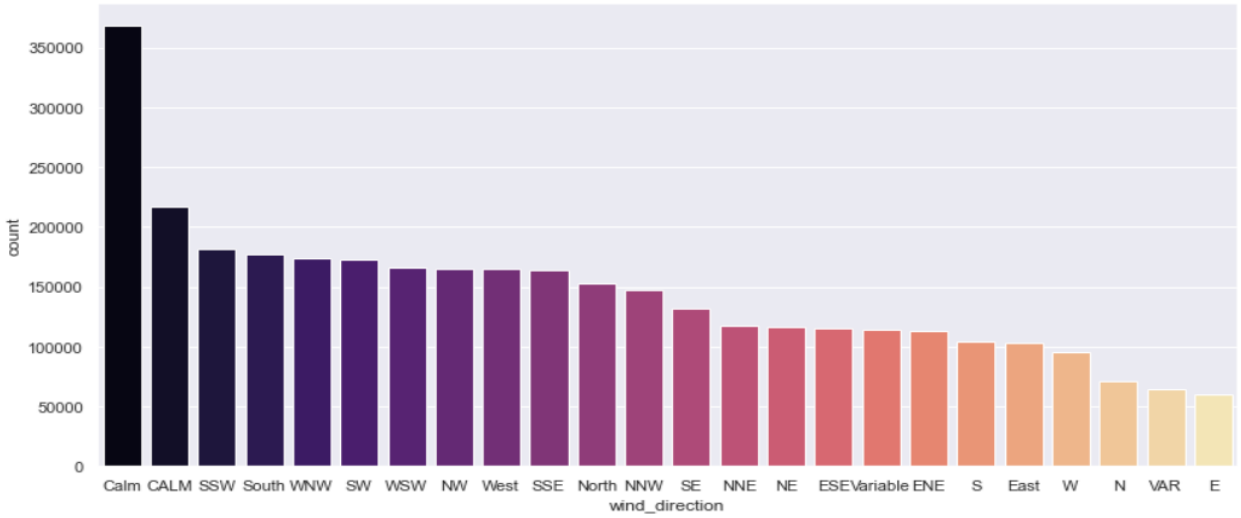
	amenity	bump	crossing	give_way	junction	no_exit	railway	roundabout	station	stop	traffic_calming	traffic_signal
0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	1
3	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	1
5	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	1	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	1
14	0	0	0	0	0	0	0	0	0	0	0	1

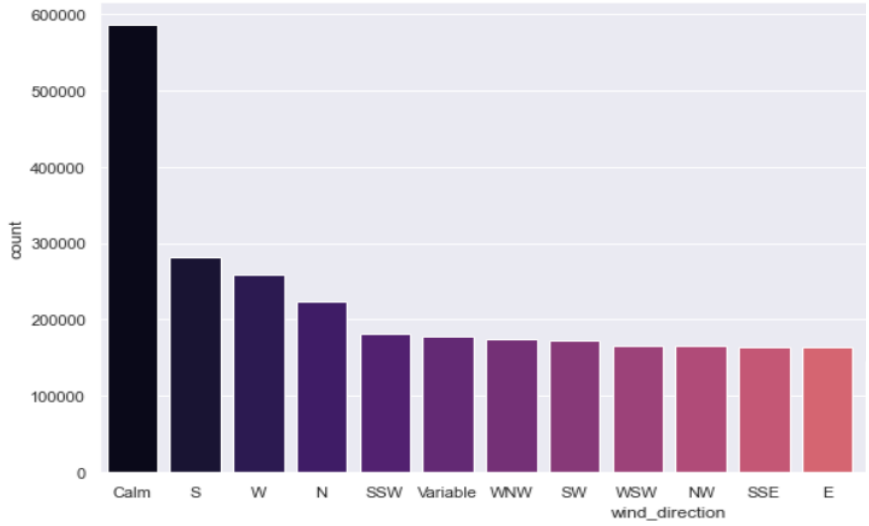
‘ side ’ column is binary in nature with R/L. we have encoded it as **R = 1** and **L = 0.**

‘ sunrise_sunset ‘ is also binary in nature.It is encoded as **Day = 1** and **Night = 0**.

‘ wind direction ‘ has 24 unique values with some incorrect data, such as:

- Calm and CALM mean the same.
- Four cardinal directions are in abbreviated form and in full form also.
- Variable and VAR mean the same.





After the replacements, there are 18 unique values. These values are label encoded using sklearn's inbuilt *LabelEncoder()* from 0 to 17. OneHotEncoder() would increase the size of the data so we intend to experiment with it after developing a baseline model. ' *weather_condition* ' has 127 unique values which are label encoded in the same way.

➤ PCA

We have applied Principal Component Analysis using the sklearn's inbuilt PCA() function. Initially the number of principal

	Explained_variance	Cumulative Variance
0	5.780268e-01	0.578027
1	2.698647e-01	0.847891
2	1.232571e-01	0.971149
3	1.494943e-02	0.986098
4	1.005259e-02	0.996151
5	3.185812e-03	0.999336
6	3.497028e-04	0.999686
7	8.907769e-05	0.999775
8	7.648342e-05	0.999852
9	6.379498e-05	0.999916
10	3.321806e-05	0.999949
11	2.424261e-05	0.999973
12	9.358936e-06	0.999982
13	6.719942e-06	0.999989
14	5.112044e-06	0.999994
15	3.719040e-06	0.999998
16	1.238598e-06	0.999999
17	5.956549e-07	1.000000
18	2.300228e-07	1.000000
19	3.662449e-08	1.000000
20	2.459796e-08	1.000000

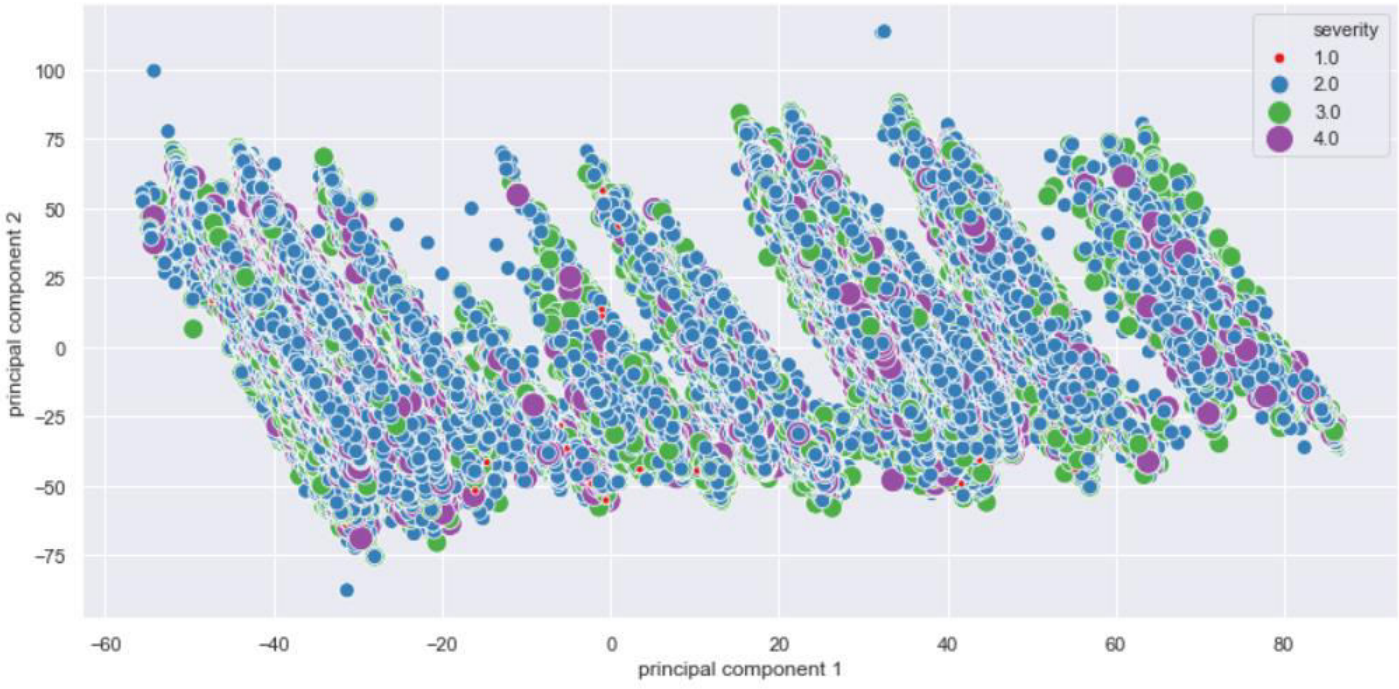
components is set to default (equal to the number of features). The features which are used are shown below.

We intend to use these features to predict the target column *severity* as well. When we examined the explained_variance and cumulative_variance table we noticed that only the first 6 principal components are enough to explain 99.9% variance of the features shown above. For visualization purposes we have selected the first 2 principal components which explain 84.7% variance. When these 2 components are plotted using a 2-D scatter plot, there are no differentiating cluster for each severity level.

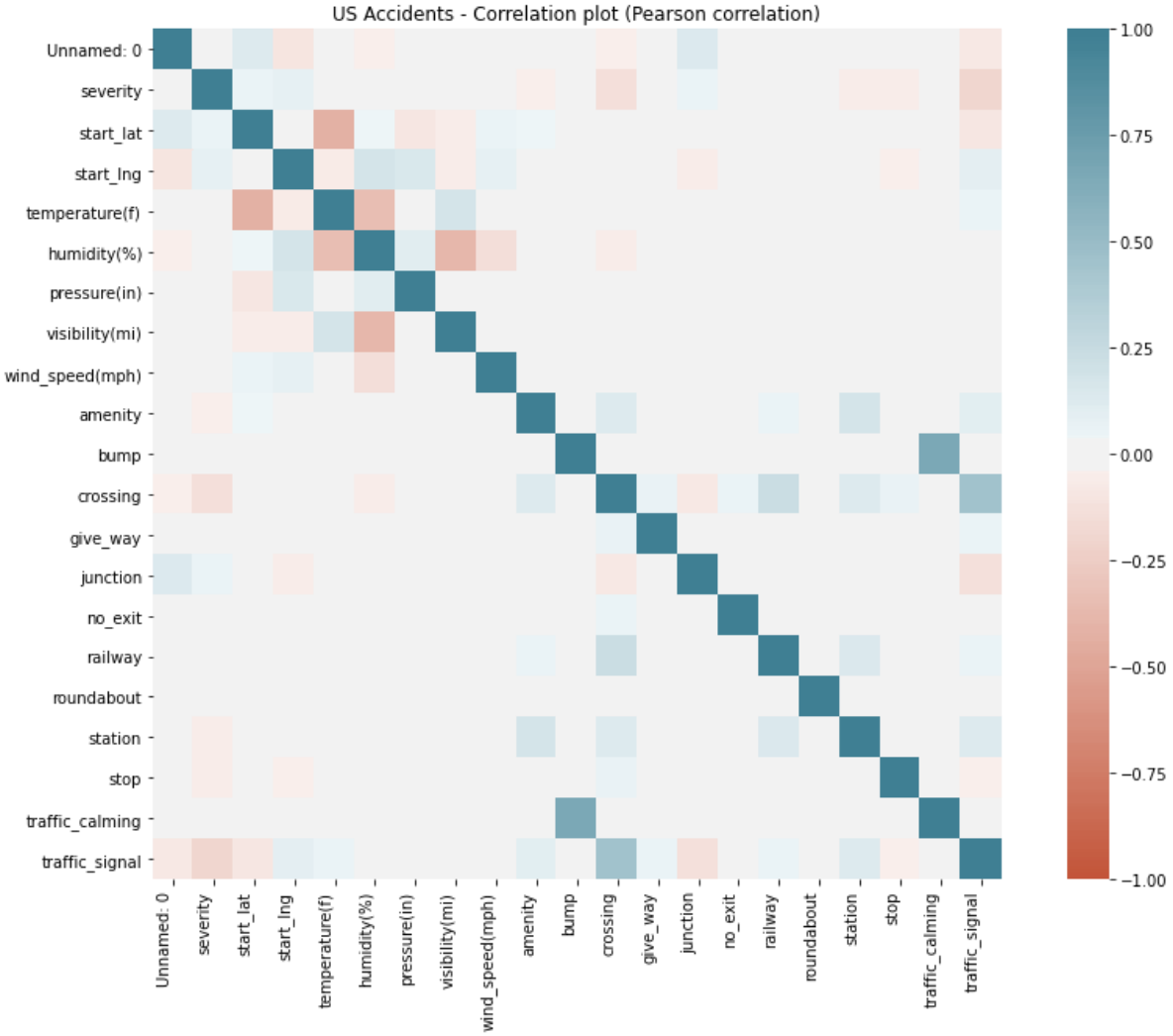
```
df[['side', 'temperature(f)','humidity(%)', 'pressure(in)', 'visibility(mi)', 'wind_direction',
   'wind_speed(mph)', 'weather_condition', 'amenity', 'bump', 'crossing',
   'give_way', 'junction', 'no_exit', 'railway', 'roundabout', 'station',
   'stop', 'traffic_calming', 'traffic_signal', 'sunrise_sunset']]
```

➢ Correlation:

A Pearson correlation has been performed to understand the linear relationship among different features. From the correlation it can be observed that *traffic_signal* and *crossing* have a positive correlation with each other. *Traffic_calming* and *bump* too have a significantly higher



positive correlation with each other. *Temperature* and *Humidity* show a negative correlation between them. However though *visibility* shows negative correlation with *humidity*, due to the volume of data, scatter plot didn't reveal a significant negative slope. The correlations possibly indicate important features that influence the occurrence of an accident, in general.



➤ Literature Review:

Road Accident Analysis and Prediction of Accident Severity by Using Machine Learning in Bangladesh

Authors: Md. Farhan Labib, Ahmed Sady Rifat, Md. Mosabbir Hossain, Amit Kumar Das, Faria Nawrine.

The paper uses supervised machine learning techniques to classify the severity of accidents that take place in Bangladesh. The paper also uses different visualizations to interpret the impact of a given feature on the accidents being caused. The authors have collected a dataset from the ARI of BUET that consists of a total 43,089 traffic accidents[2001-2015] in Bangladesh. The dataset has 34 features with 8.7% missing values in total. The paper uses three feature selection algorithms to reduce the number of features: Univariate Feature Selection, Recursive Feature Elimination, and Feature Importance. Mean imputation is used to deal with 1.65% missing values in the selected features. Decision Tree, K-Nearest Neighbors (KNN), Naïve Bayes and AdaBoost are applied to classify the severity of the accidents. The target column is found to be imbalanced, hence Feature Creation is used to balance the target column. Model performance is found to improve slightly when the target column is balanced. Adaboost outperforms the other three algorithms as it is a combination of multiple weak classifiers.

Accidents are relatively unforeseen and spontaneous, so direct observation is quite difficult. For that reason, getting 100% accurate data is quite impossible.

The paper claims that training time and the risk of overfitting can be reduced with a suitable feature selection algorithm.Missing values also affect the performance of the model, thus an appropriate data imputation method should be used.

The key takeaways from this paper is that data preprocessing techniques like data cleaning and feature selection are very integral before passing the data to a machine learning model. Imbalanced data can lead to a biased/skewed model.

Summary of Analysis of road traffic fatal accidents using data mining techniques.

Authors: Liling Li, Sharad Shrestha, Gongzhu Hu
IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA), pp. 363-370. IEEE, 2017.

This paper is about road traffic accidents analysis which shows the relationship of various factors affecting the same. They used FARS (Fatal Accidents Reporting System) dataset for their study, containing 37,248 records and 55 attributes. Since our paper is about effects of weather on road accidents, this paper is most relevant as it explains all features affecting the accident. Careful analysis of road traffic data is critical to find out the variables that are closely related to fatal accidents. Association rules were discovered by Apriori algorithm, classification model was built by Naïve Bayes classifier, and clusters were formed by simple K-means clustering algorithm with Euclidean distance as dissimilarity measure.

Data cleaning is important so all the tuples with missing values are removed and numerical values are converted into nominal values in accordance with the requirement.

The claim of this paper is that if more data, like non-fatal accident data, weather data, mileage data, and so on, are available, more tests could be performed thus more suggestions could be made from the data.

The takeaway from this paper is that selection of appropriate features is crucial as they will help in more accurate analysis of data.

A Statistical Analysis of Recent Traffic Crashes in Massachusetts

Authors: Zhang, Aaron, Evan W. Patton, Justin M. Swaney, and Tingying Helen Zeng
arXiv preprint arXiv:1911.02647 (2019)

The main purpose of the study was to determine the greatest risk factors for crashes that result in injury such that a greater emphasis can be placed upon eliminating those risks. Represented the categorical variables numerically using one-hot encoding. Then the authors use Logistic Regression with and without Recursive Feature Elimination (RFE) to perform feature selection and identify whether a certain set of conditions would increase the likelihood of an injury-inducing accident. The dataset for this paper was a randomly selected sample of 46,077 crash entries for the last five years, retrieved from MassDOT's crash portal, based in Massachusetts.

They assumed that crash data is better interpreted by linear modelling. They removed columns such as co-ordinates which had too many distinct possibilities to group together.

The paper claims that prediction accuracy can be increased by using suitable feature selection algorithms. They also suggested using SVM's in the future to better model the non-linearity of certain attributes contributing to the accident. According to the authors features like tough weather and road conditions, senior/teen drivers were the major accident-causing factors.

The key takeaways from this paper is that data preprocessing techniques like data cleaning and feature selection are very integral before passing the data to a machine learning model. By encoding the categorical values, the machine learning algorithms can better understand data. The attributes of the accident dataset might not always be linearly related and imbalanced data can lead to a biased/skewed model.