**CSCI-4930/5930: Machine Learning**

**Final Project Report**

1. **Title of the project**: : A Model To Predict the Severity of the Traffic accidents and Impact of Environment on Accidents
2. **Team**:

Druva Sai Kumar Bobbilla

Prem Rishi Kranth Gaddam

Sampath Mandava

1. **Introduction**

**Definition of the task(s)/problem**: In this project we have analyzed the data of US - Accidents and predicted the Severity of the accidents based on weather, time and place. We performed a neural network on text classification . By using the data we identified certain patterns which are disclosed in our results section and we have discovered a few more insights where we can discover facts which are

crucial in our project. There are various many other factor such as the effect of the Temperature(F), Wind\_chill(F), Humidity(%), Pressure(in), Wind \_direction, Wind\_Speed(mph), Precipitation(in) and other environmental effects on the occurrence of accident , we used these as the significant features in our model . We visualized the facts identified during the analysis which makes it easier to detect patterns in the data.

**Purpose of the task(s)**:

We utilized the neural networks which gives more precision to our model for forecasting the severity of the traffic accidents. We Analyzed the patterns to extract insights from those which helped us in discovering the facts which are really being used throughout the project.

1. **Information on the Dataset**:

We have analyzed the dataset completely and the features. There are 49 attributes in the dataset .We have considered the relevant attributes for our model and Each of those attributes are listed below and explained as follows

 1)Severity-Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay as a result of the accident) and 4 indicates a significant impact on traffic (i.e., long delay).

 2)Start\_Time-it Shows start time of the accident in the local time zone.

 3)End\_Time-Shows end time of the accident in the local time zone.

4)Start\_Lat-Shows latitude in GPS coordinate of the start point.

5)Start\_Lng-Shows longitude in GPS coordinate of the start point.

6)End\_Lat-Shows latitude in GPS coordinate of the end point.

7)End\_Lng-Shows longitude in GPS coordinate of the end point.

8)Distance(mi)-The length of the road extent affected by the accident.

9)Description- Shows natural language description of the accident.

10)Street-Shows the street name in the address field.

11)City-Shows the city in the address field.

12)County-Shows the county in the address field.

13)State-Shows the state in the address field.

14)Zip Code-Shows the zip code in the address field.

15)Timezone-Shows timezone based on the location of the accident (eastern, central, etc.).

16)Weather\_Timestamp-Shows the time-stamp of the weather observation record (in local time).

17)Temperature(F)-Shows the temperature (in Fahrenheit).

18)Wind\_Chill(F)-Shows the wind chill (in Fahrenheit).

19)Humidity(%)-Shows the humidity (in percentage).

20)Pressure(in)-Shows the air pressure (in inches).

21)Visibility(mi)-Shows visibility (in miles).

22)Wind\_Direction-Shows wind direction.

23)Wind\_Speed(mph)-Shows wind speed (in miles per hour).

24)Precipitation(in)-Shows precipitation amount in inches, if there is any.

25)Weather\_Condition-Shows the weather condition (rain, snow, thunderstorm, fog, etc.).

26)Bump-A POI annotation which indicates presence of speed bump or hump in a nearby location.

27)Crossing-A POI annotation which indicates presence of crossing in a nearby location.

28)Give\_Way-A POI annotation which indicates presence of give\_way sign in a nearby location.

29)Junction-A POI annotation which indicates presence of junction in a nearby location.

30)No\_Exit-A POI annotation which indicates presence of no\_exit sign in a nearby location.

31)Roundabout-A POI annotation which indicates presence of roundabout in a nearby location.

32)Stop-A POI annotation which indicates presence of stop sign in a nearby location.

33)Traffic\_Calming-A POI annotation which indicates presence of traffic\_calming means in a nearby location.

34)Traffic\_Signal-A POI annotation which represents the presence of traffic\_signal in a nearby location.

35)Turning\_Loop-A POI annotation which indicates presence of turning\_loop in a nearby location.

36)Sunrise\_Sunset-Shows the period of day (i.e. day or night) based on sunrise/sunset.

**Proposed method:**

We Analyzed the dataset thoroughly by applying various algorithms such as decision trees , naïve bayes, K-nearest neighbor ,Random forest, Support Vector Machine and the Logistic Regression we gained important insights .Apart from the previously mentioned algorithms we have also used neural network on text classification in our project We compared and analyzed various results from which we obtained the best ones. We are successful in our analysis and the results are described as follows.

**Results**: You can present here some preliminary results. Please maintain a private Github repository, and send me an invite for verification. My github handle is ashiskb

By using the data we have identified certain patterns, we discovered some interesting facts based on these such as there are a greater number of accidents in California which is twice the number when compared to the next highest state Texas. Most of the accidents occurred during clear weather rather than the rainy and snowy days. We have completed extracting the results and insights. We have used different kinds of visualization techniques such as the Folium, Factor plot and more.

The following are the insights obtained in our Analysis

**Insights:** This plot describes the percentage of the accidents in top15 states.

A picture containing accessory, umbrella

Description automatically generated

**Insights:** This plot describes the distribution of the severity of all the accidents

A picture containing text

Description automatically generated

**Insights:** This plot describes the severity of accidents on each state

A close up of a colorful background

Description automatically generated

**Insights:** This plot describes the proximity to traffic object . In this plot we can observe that there are high number of accidents occurred near the traffic signals and then the junctions

A picture containing clock

Description automatically generated

**Insights:**

By observing this California has highest number of accidents , from this plot we can conclude that the central and the eastern time zones are having more high severity cases

A close up of a logo

Description automatically generated

**Insights:** From this plot we can say that the more impacted sides of the vehicles is the right side . we can also infer that the driver side is less impacted during most of the accidents .

A picture containing device

Description automatically generated

**Insights:**This plot depicts the state wide severity rate .From this plot we can say that the there are more number of high severity cases in South Dakota

A picture containing object, antenna

Description automatically generated

**Insights:** Here comes the correlation plot .we can see that there few parameters which are closely related to others .So over here wind chills and temperature are closely correlated

A close up of a screen

Description automatically generated

**Keyword Detection**:

This is the word cloud based on the accident description

A close up of a black background

Description automatically generated

This is the word cloud obtained considering the weather conditions

A picture containing table

Description automatically generated

**Insights:** This plot is based on the weather conditions

A screenshot of a cell phone

Description automatically generated

**Insight:** This plot is based on the severity with respect to temperature

A picture containing building, door

Description automatically generated

**Insight:** This plot represents the severity with respect to different weather conditions.

A picture containing stationary, implement, pencil

Description automatically generated

**Insight:** This plot represents the different time zones in the USA.

A picture containing text

Description automatically generated

**Insights:** The plots below the day, week and the year wise accidents count.

A close up of a screen

Description automatically generated

**Insights:** This plot gives the information about time to clear the accident area and max area are cleared with-in 30 minutes and max time is around 5 hours.

A screenshot of a cell phone

Description automatically generated

**Insights:** This show the accidents severity from 2016-2019

A close up of a map

Description automatically generated

**Insights:** This plot shows the accidents rate hourly and we can see that 7am records the highest accident rate.

A picture containing drawing

Description automatically generated

**Insights:** From the monthly analysis we can say that the 12th day of the month has the highest number of accidents.

A picture containing pencil

Description automatically generated

**Insight:**  This plot represents the severity with respect to day and night.

A screenshot of a cell phone

Description automatically generated

**Insight:** Accidents from 2016 to 2019 in denver

A close up of a map

Description automatically generated

**Insight:** Plot represents the accidents severity with respect to the temperature in Colorado

A picture containing sitting, green, bright, side

Description automatically generated

**Insight:**

A screenshot of a cell phone

Description automatically generated

A picture containing screenshot

Description automatically generated

A screenshot of a cell phone

Description automatically generated

**Insight:** There is a steep increase in accidents from July 2017.

A close up of a map

Description automatically generated

**Insights:** These are the days with most accidents recorded in Colorado

A picture containing drawing

Description automatically generated

**Insights on Feature Importance:**

The first plot is plotted before the data cleaning and the second one is plotted after the data preprocessing is done. From this we can say that the location and the Distance attribute plays a major role.

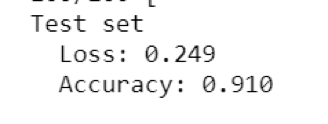
A screenshot of a cell phone

Description automatically generatedA picture containing drawing, fence

Description automatically generated

**Text Classification:**

We performed the text classification using the method neural network by using the accident description along with the severity. We got the accuracy of 91% by using the neural network. The following is the loss graph.



A close up of a map

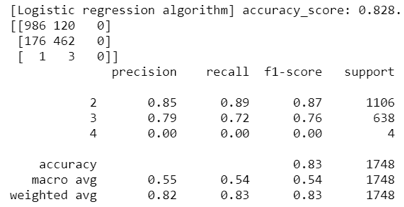
Description automatically generated

**Algorithms with accuracy details:**

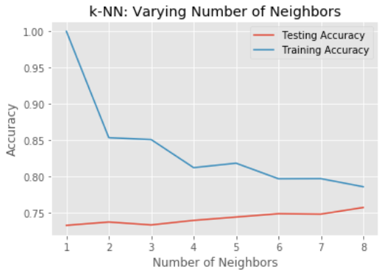
|  |  |  |
| --- | --- | --- |
| **Algorithms** | **With Data Pre-Processing** | **Without Data Preprocessing** |
| **Logistic Regression** | **83%** | **78%** |
| **K-nearest neighbours** | **75%** | **74.4%** |
| **Decision Tree** | **83.5%** | **81.2%** |
| **Random Forest** | **89.2%** | **91.6%** |
| **SVM** | **83%** | **71%** |
| **Naïve Bayes** | **65.5%** | **19%** |

**Results Evaluation:**

To evaluate the results we are using the confusion matrix for each and every algorithm we applied during the project.



The interesting fact we found out in the KNN is the test accuracy is increasing with increase in number of neighbors.



Github link:

<https://github.com/premrishikranth-gaddam/Model-to-predict-the-severity-of-the-Trafic-Accidents>

1. **Summary of contributions**: List all members of the team, and who contributed in what part.

|  |  |  |
| --- | --- | --- |
| Member name | Task | Comment to explain the contribution a little bit better |
| Druva Sai Kumar Bobbilla | EDA and Data Preprocessing, Analysis, Extracting Insights  Documentation | Performed the exploratory data analysis , Preprocessing tasks such as dealing with noise in the data. Successfully driving the completion of each area. Attention to the details related to each area throughout the project . Preparing the action plan on progressing through the completion of project and Documentation . |
| Prem Rishi Kranth Gaddam | Pattern Extraction and Multi Class classification with prediction of severity | Worked on pattern extraction and focused on the implementation of various algorithms along with the text classification by neural network. |
| Sampath Mandava | Data Visualization | Visualization of the insights and facts which better conveys the knowledge from the data |

Apart from the above-mentioned tasks and contribution we have collectively worked on the project on each of the aspects and made the project successful.

**Conclusion:**

* We are able to derive a variety of insights with respect to the location, time, weather, and points-of-intersection of an accident
* We have predicted the severity based on other variables related to weather , traffic ,time .we are successful in doing that and very much satidfied by the results
* Our model can be integrated with the many government agencies , road traffic departments and also this can be useful for the normal people who are using the roads .

**Applications:**

The Analysis and knowledge extracted from our project can be utilized in many applications .Mainly in studying the car accidents and the hotspot locations. Our project also used in analyzing the severity the accidents and also in extracting the cause and effect relationship to predict car accidents . It can also be very much useful in studying the impact of weather conditions and also other environmental stimuli on accident occurrence

1. Future work

We are planning to enhance our model with one of the important parameters which is the driver behavior . As it is not available in the dataset , if it is included that will enhance our model .

The Implementation of our project will definitely help the public along with the road traffic departments in reducing the number of accidents by alerting the users beforehand.

1. **List of references:** Enumerated list of references you cite in the project report. If you add a figure in the report that you copied from some website, put a number, say at the caption of the figure and add a reference entry here in this section containing the web address.

[1]<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6448880/pdf/pone.0214966.pdf> [2]<https://towardsdatascience.com/live-prediction-of-traffic-accident-risks-using-machine-learning-and-google-maps-d2eeffb9389e>

[3] <https://smoosavi.org/datasets/us_accidents>

[4] <https://arxiv.org/pdf/1906.05409.pdf>

[5]<https://arxiv.org/pdf/1906.05409.pdf>

[6] <https://arxiv.org/abs/1909.09638>