US Accidents' Severity Prediction

Samarth G Vasist
Department of Computer Science and Engineering
PES University, Bangalore, India
samarthgvashist2000@gmail.com

Sreejesh Saya
Department of Computer Science and Engineering
PES University, Bangalore, India
saya.sreejesh@gmail.com

Vishesh P
Department of Computer Science and Engineering
PES University, Bangalore, India
visheshp172000@gmail.com

Abstract-- In this day and age where traffic accidents are increasing in number and severity, determining what environments they occur in and how that can be used to prevent future accidents is of the utmost importance. We have observed the dataset named "US-Accidents, A Countrywide Traffic Accident Dataset" and visualised the different aspects of the accidents such as Location and Severity of accidents occuring in each state among others. After performing the appropriate preprocessing steps on the dataset, we implemented several classification algorithms such as Logistic Regression, KNN, Decision Trees, Random Forest and Naive Bayes. A comparative study was performed to determine the best classification method to accurately classify the severity of the accidents observed given the details of the accidents. Further, we have also determined the conditions under which accidents of higher severity generally occur more frequently in the US.

I. Introduction

In this day and age of traffic movement, public safety has become more important than ever, but it's prevention has been neglected all over the world. Hence there is a compelling need to find out the causes and risks behind the unfortunate accidents occurring across the world and present it in such a way that it is easy for everyone to understand. We will be predicting the severity of each accident based on the conditions under which the accidents have occurred. This is done by developing several classification models, testing all of them, comparing the results of the models with each other and finally picking out the most accurate model. We restrict our research and analysis to the accidents that have occurred recently in the United States of America, using the Dataset "US-Accidents : A Countrywide Traffic Accident Dataset" which has been created using collected in real-time, using multiple Traffic APIs, ranging from accidents that have occurred all across the country between February 2016 and June 2020.

II. Related work

Sobhan Moosavi et al. [1] talks about the need to have a publicly-available dataset that has a larger coverage in terms of more attributes describing the accidents such as environmental stimuli, textual information describing each accident and also have a dataset that is up to date containing details of accidents that have occurred recently and not outdated data which has little to no relevance with current Data Analytics. They further discuss the creation of a dataset that matches their requirements and naming it "US-Accidents". The paper presents a collection of 2.25 million accident records which are further augmented with the help of map-matching along with context details like weather, hour of day along with the points-of-interest. The process they carried out for the dataset creation mainly consisted of 3 main steps, namely real-time traffic data gathering followed by integration and augmentation of the collected data. "MapQuest Traffic" and "Microsoft Bing Map Traffic" API's helped in gathering the real time traffic data through the help of law enforcement agencies, cameras and sensors used in the road-networks. In the integration step, duplicate records were removed while building a unified dataset. Finally coming to the data Augmentation step, a process of data augmentation is carried out by augmenting GPS data with reverse Geo-coding, weather data with Weather Underground API and points of interest from OpenStreetMap (OSM). Accident types such as junction and intersection for a particular accident were assigned based on the annotations and the data was prepared for prediction.

Sobhan Moosavi et al. [2] uses a large publicly available dataset known as US-Accidents to train their deep-neural network model named Deep Accident Prevention (DAP) to predict an accident in a given region and in a 15 minute time interval. The dataset used contained information such as weather, point of interest, traffic events, the latter two, first time to be included in any dataset. The DAP model once trained with the dataset, was put to test in 5 different US states each having different attribute values of important factors such as weather, point-of-interest against baseline models - Logistic Regression, Gradient Boosting Classifier and a four-layer deep neural network. The DAP model proved to be superior in five out of the six cities over the baseline models in Accident situations (the baseline model - logistic regression had better accuracy in non-accident situations but that is ignored due to negative sampling and the fact that we are more focussed on accident related cases).

Zhuoning Yuan et al.[3] used a big dataset including all the motor vehicle crashes ranging from 2006 to 2013. They incorporated the spatial structure (one of the first) of the road network into the predictive models by leveraging new features generated through eigen-analysis of the road network to address the spatial heterogeneity challenge. Experiments on four classification models, i.e., SVM, Decision Tree (CART) , Random Forest and Deep Neural Network (DNN) were performed and their results were compared and discussed and showed that Decision trees, random forests and DNNs outperform the SVMs which verifies that linear models are not effective for predicting traffic accidents. They chose to fill in missing values using interpolation instead of removing the records.

III. Dataset

The dataset we have used contains over 2.25 million records where each record represents an accident occurring somewhere in the US. The dataset contains records of accidents ranging from February 2016 to March 2020. Each record has several features describing the accident such as the location, textual information about the accident and the severity of each accident among others. It also includes a key attribute the points-of-interest (eg. traffic signal, stop sign). We will be focusing only on 5 states in the US viz. Kentucky, Louisiana, Massachuesetts, California and Maryland. The total number of accidents in these states is roughly 900 thousand in the given timeframe out of which 800 thousand occurred in the state of California.

There are a total of 49 attributes in the given dataset, 20 of which are categorical, 16 numerical and the remaining binary. The numerical attribute 'Severity' is ordinal in nature, ranging from 1 to 4 indicating the severity of the accident, 1 being least severe and 4 being the most severe.

IV. Methodology

A. Exploratory Analysis

After loading the dataset of the accidents that have occurred in the 5 states in the USA chosen accounting upto about 1 million entries, we plot the accidents geographically using their latitude, longitude. We observe that the accidents form 5 clusters representing the 5 states respectively. On analyzing the number of accidents state-wise we found that the number of accidents was highest in the state of California followed by Lousiania and Maryland while Kentuckey recorded the least number of accidents.

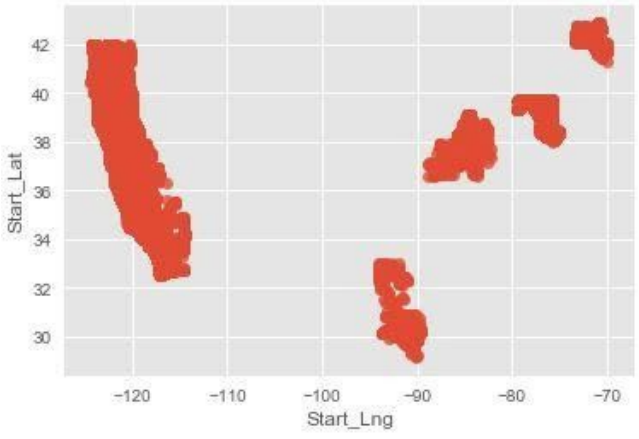

Fig.1 Clusters of 5 states based on location

Severity being the target variable which is ordinal in nature, we construct a bar plot to see the distribution of the accidents based on severity, we see that virtually all accidents reported belong to either severity 2 or severity 3. When plotted state-wise, the graphs grouped by states show that the above results are consistent among the states.

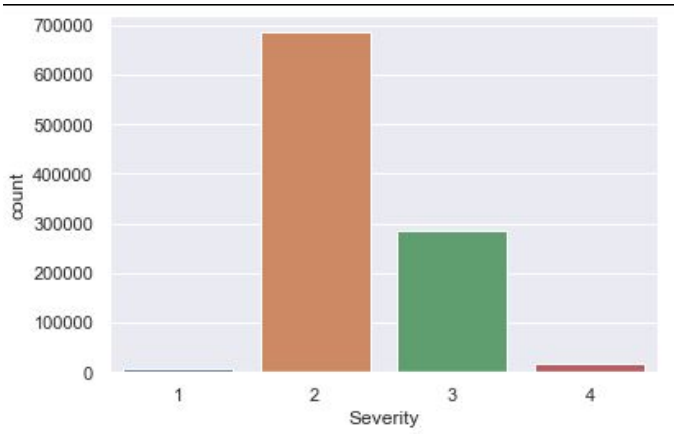

Fig.2 Distribution of severity

On creating additional attributes such as Start_year, Start_Month and Start_day using the timestamp attributes we analysed how the weekdays and weekends influence on the number of accidents. In total there are 993530 accidents reported from the 5 states , with 1586 unique days, roughly 626 accidents per day. We performed a time-series analysis (resampled by month) for the accidents per state for every month over the timeline. We see that the number of accidents in California over time is nearly constant with not much of variation whereas the rest of the states have a spike in the number of accidents over time.

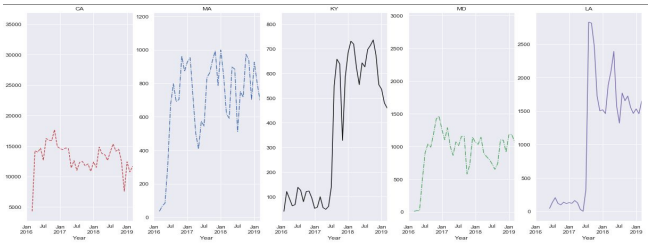

Fig.3 Time Series analysis resampled by months for all five states

Analysis through plots for the number of accidents based on the day of the week reveal that the number of accidents on weekends are fairly less compared to the count on weekdays for all the states. The results from the pie charts are also consistent with the bar plots with weekends constituting only 10-12% of the total accidents approximately for each state.

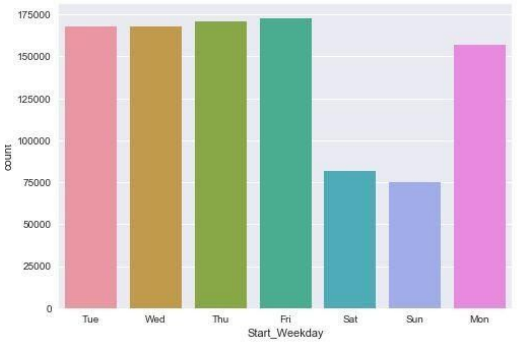

Fig.4 Count of accidents for every day of the week

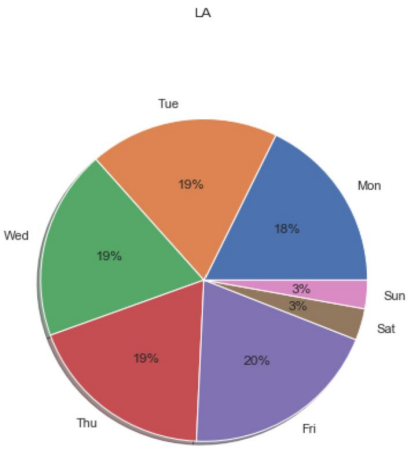

Fig.5 Percentage of accidents occurring on each day of the week in Louisiana

A time series analysis for the accidents for each hour (hourly distribution) for each state indicates that the number of accidents in California is very high compared to the rest of the states. The time series data is plotted for all days and separately for weekdays and weekends. Results being consistent with all days and weekdays plots, a slight difference was noted in the time-series plots on weekends. Generally a higher number of accidents were reported during the 8th and 17th hour for all

the states on weekdays whereas it was the 12th and 14th hour on weekends.

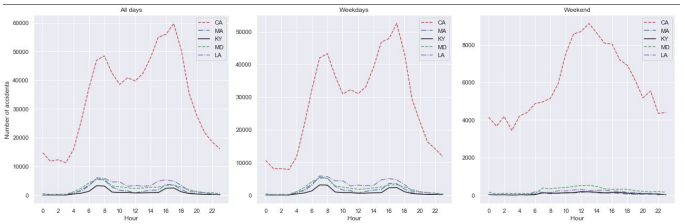

Fig.6 Time-series analysis resampled by Hour for all five states

We then plot pie charts to see the individual city contributions for accidents in each state and find that Los Angeles, Boston, Louisville, Baltimore and Rouge are the cities having the highest number of accidents for the states of California, Massachusetts, Ken Tuckey, Maryland and Lousinia respectively.

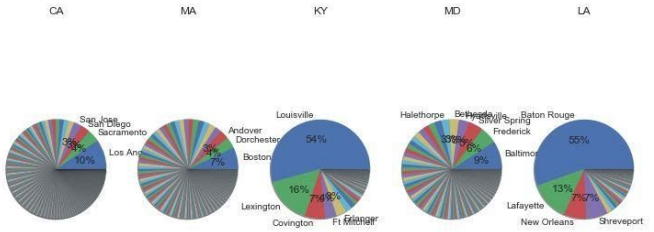

Fig.7 Percentage of accidents occuring in each city for all five states

Pie charts for the accident locations show us that junctions constitute a large percentage of accidents. Pie charts for the weather conditions show us that most accidents took place in a clear sky.

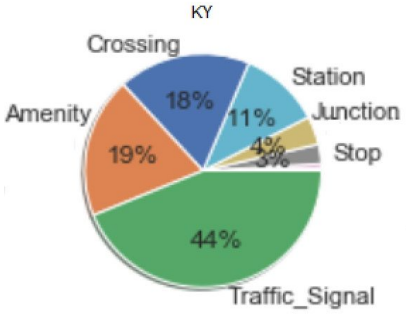

Fig.8 Percentage of accidents occurring at different locations for Kentucky

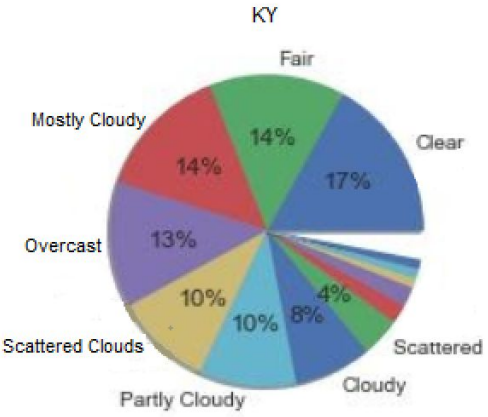

Fig.9 Percentage of accidents occurring in different weather conditions in Kentucky

A box plot was plotted to check whether temperature influenced the number of accidents in the 5 states and concluded that the temperature in Louisiana is slightly higher than that for other states.

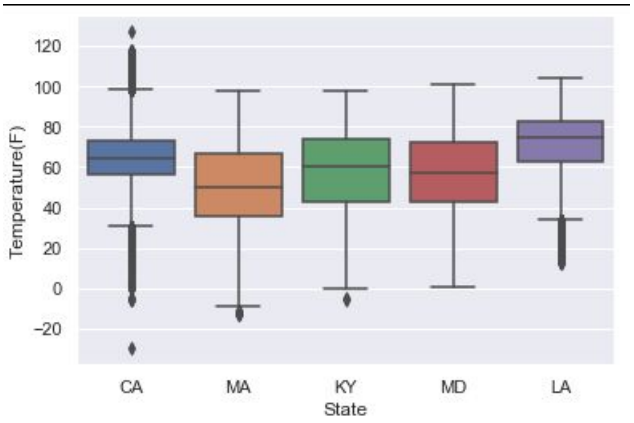

Fig.10 Boxplot showing the distribution of temperature at the time the accidents were recorded for each state

B. Data Cleaning

Since the above dataset has many outliers and missing values it is necessary to clean the data to ensure that the former values do not significantly affect our data. We analyse the accidents that have occurred in each state individually in order to make sure that there is no relation between the accidents taking place in one state with the other. We analyzed the accidents relating to the state of Kentucky, performed data cleaning and

incorporated the same method to the remaining 4 states. Since large missing values significantly influence our data we have dropped columns having missing values that contribute upto 50% of the total number of entries belonging to a particular state. We then drop columns such as 'Turning_Loop', 'Civil_Twilight', 'Nautical_Twilight', 'Astronomical_Twilight', 'Weather_Timestamp', 'TMC', 'Start_Time', 'End_Time', 'Description' , 'Street', 'Unnamed: 0' and 'ID' as they don't contribute in influencing the response variable. We also dropped the columns 'State' and 'Country' as those values are the same for the respective states.

We then analyzed each of the columns having missing values and treated the outliers and NULL values accordingly. The temperature column is analyzed first through a box plot alongside calculating the number of outliers and since it followed an approximate normal distribution for few of the states under study which was verified by plotting histograms for the same we filled the missing values with the median. The same measures were taken for filling the humidity attribute where we filled the missing values with median humidity for a particular state.

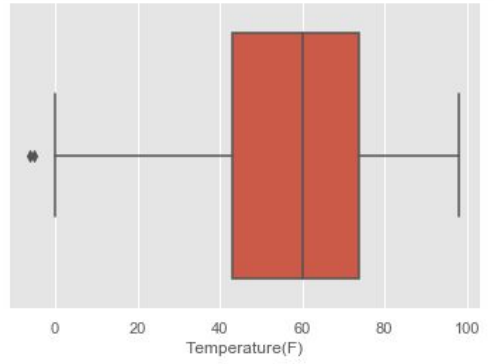

Fig.11 Boxplot showing the distribution of Temperature at the the time the accidents were recorded for the state of Kentucky

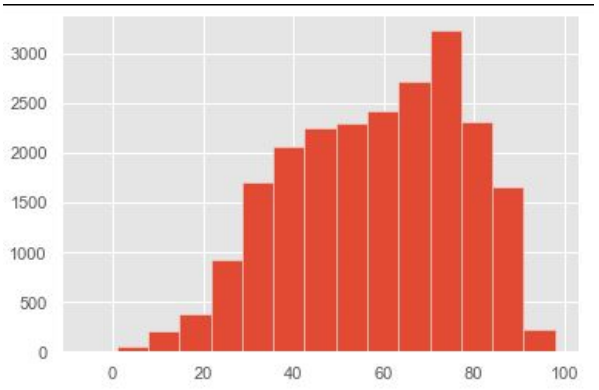

Fig.12 Distribution of Temperature for the state of Kentucky

The pressure attribute had a significantly higher number of outliers reported for a few states and hence we used the method of flooring and capping in order to treat the outliers where all the values less than the 10th percentile and values greater than the 90th percentile are replaced by the respective percentile values. The missing values were then filled by the median of the column. The same method of flooring and capping was used to fill the missing values in the columns of Visibility and Wind Speed since the outliers followed the same pattern and median was used to fill the missing values.

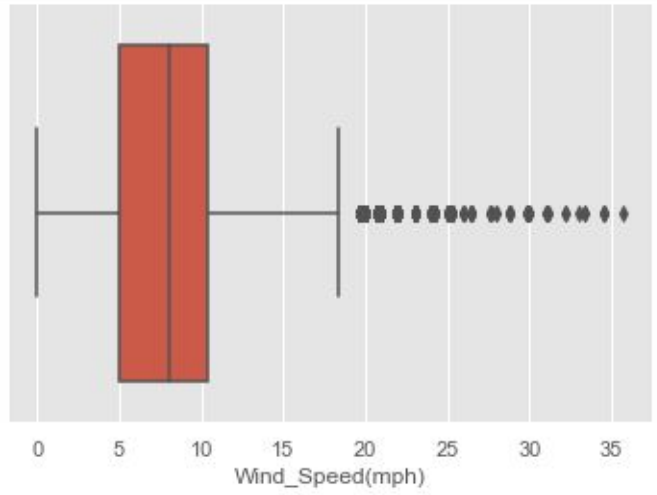

Fig.13 Boxplot showing the distribution of Wind Speed at the time the accidents were recorded for the state of Kentucky

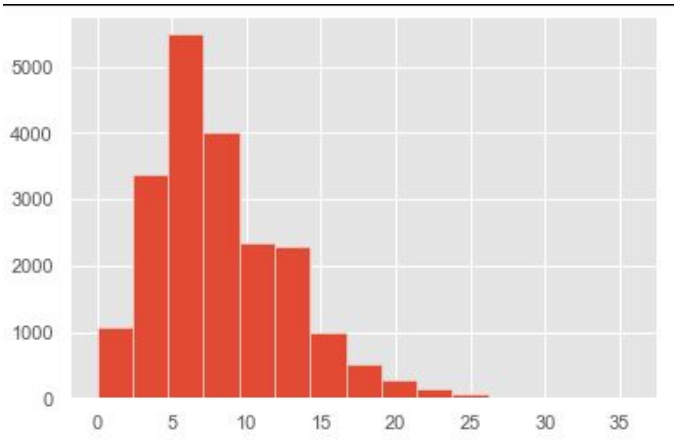

Fig.14 Distribution of Wind Speed after performing Flooring and Capping on the data for the state of Kentucky

Wind_direction, Weather_Condition and City being categorical variables, the missing values are filled by the most frequently occurring value (i.e mode) of the column. We plot correlation plots between the numerical columns and infer the Pearson correlation coefficient using a heatmap and infer that each state has its own different correlations established between the variables with the negative correlation between humidity and temperature being the common one reported for all the states.

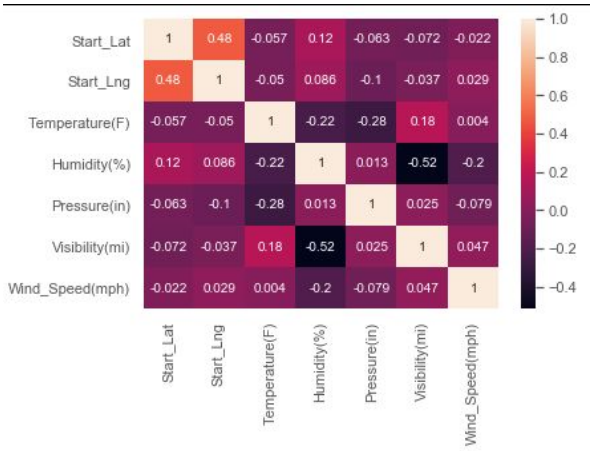

Fig.15 Heatmap showing the correlation between every pair of numerical attributes for the state of Kentucky

C. Data Preprocessing

Since the numerical values have different scales, there is a need for feature scaling in order to have a common scale. We use the StandardScaler package of python in order to standardize the numerical attributes which produce new distributions for each numerical attribute, all of which will have a mean of zero and unit variance. As many of our columns are categorical, there is a need for categorical encoding and hence we have implemented both one hot encoding and label encoding for a particular set of attributes. Attributes such as 'Side', 'Crossing', 'Bump', 'Give_Way', 'No_Exit', 'Junction', 'Roundabout', 'Railway', 'Stop', 'Station', 'Traffic_Signal', 'Traffic_Calming', 'Sunrise_Sunset', 'Year', 'Month', 'Day', 'Hour', 'Weekday' are one hot encoded using the get_dummies function of pandas. The remaining categorical features such as 'Source', 'City', 'County', 'Wind_Direction', 'Weather_Condition' are encoded using label encoding where each value is replaced with a unique integer based on alphabetical ordering.

D. Model Definition

1)Logistic Regression

Initially the dataset containing the accident records of any one state(performed for all five states separately) was split into training and testing sets with a testing size of 20%. The same was done to test every model. Multi-class logistic regression was implemented for predicting the severity of accidents. On filling the Logistic regression model onto the training data and then predicting the test results we obtained an accuracy of 77.499%. We also implemented dimensionality reduction using Principal Component Analysis (PCA) where we chose 2 components such that 95% of the variance is retained. After fitting the PCA with the training data it is used to transform the training and testing data. On performing a logistics regression on this data we obtain an accuracy score of 54.954% which indicates the fact that PCA does not perform well for our dataset.

When we use only label encoding for all the categorical features and implement logistic regression we get a score of 75.903% and using PCA for the same we obtain a score of 54.932%

and hence is consistent with the conclusion that PCA is not suited for the dataset.

2)K-Nearest Neighbours

For the K-nearest neighbours, we used elbow method to determine the best K value for the multi-classification. A K-value of 7 was considered as the best and an accuracy of 72.34% was achieved.

3)Decision Trees

Coming to decision trees, we tried experimenting on the model parameters such as max_depth and criterion such as entropy and gini-index inorder to achieve the best results. A max_depth of 10 with the criterion as gini gave us the best results with an accuracy of 81.42%.

4)Random Forest

We then implemented Random Forest classification, maintaining 10 trees and setting the criterion to be *entropy*. The results proved to be satisfactory as the accuracy score of the model on the testing set for the state of California turned out to be 88.7%.

5)Naive Bayes Classifier

We further implemented the Naive Bayes classifier, particularly Gaussian Naive Bayes on the dataset. However, the results proved to be unsatisfactory as the accuracy scores never crossed 60% for any state.

V. Results

Model	Accuracy(%
Random Forest	88.7
Decision Tree	81.42
Logistic Regression	77.49
K-Nearest Neighbours	72.34
Naive Bayes	57.12

Table 1. Model vs Accuracy score for the state of California

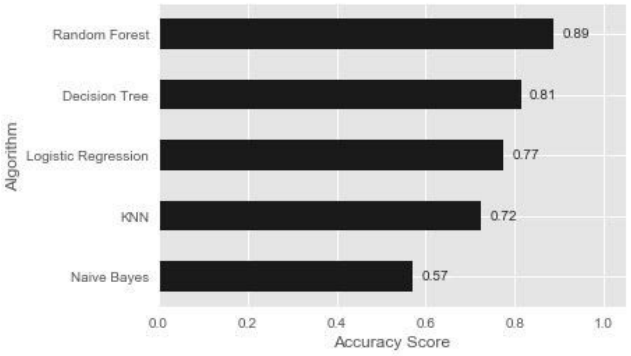

Fig 16. Bar plot showing the accuracies of various classification methods implemented for the state of California

From Table 1, it is evident that Random Forest is the best method for predicting severity of accidents for the state of California with a testing accuracy of 88.7%. Similar results were obtained while checking the accuracy of Random Forests created for the other four states(Kentucky - 81.9%, Louisiana - 87.9%, Massachusetts- 81.5%, Maryland - 75.4%). Results obtained from Support Vector Machines built for the five states were very unsatisfactory, hence they haven't been tabulated above or noted down for any other state.

VI. Conclusion

Inferring from Table 1 and the results specified above, we have concluded that Random Forest combined with the updated traffic accident data is the best performing classification method to classify severity of accidents in the US.

As expected, California, having the highest population among the five states of our choice, has the highest number of accidents. The accidents were also recorded at times of generally lower temperature. Although the number of accidents that occured in different weathers were similar, they were slightly more in number on clear days. The accidents were mostly extremely severe in every state, which seems like the most obvious inference. Rush hour was the time at which more accidents occurred throughout the country, owing to the need to make less use of our own vehicles and make use

of Public Transport at these times. Finally, Traffic Signals are considered to be the center for the majority of accidents, as shown in the dataset. We feel any combination of the above mentioned conditions are favourable for accidents of higher severity.

We further plan on performing Time-series analysis on the data. Using the description attribute of the dataset which contains textual data, we plan on implementing NLP techniques to help improve our classification process and also help the process of data cleaning as there may be more information there. We also have thought about refining the model parameters - running cross validation tests in order to achieve better results.

References

[1] Zhuoning Yuan, Xun Zhou et al “Predicting Traffic Accidents Through Heterogeneous Urban Data: A Case Study ” 2017

[2] Sobhan Moosavi, et al. "*A Countrywide Traffic Accident Dataset*" 2018.

[3] Sobhan Moosavi, et al. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights" 2018.

[4] Link for the dataset: https://www.kaggle.com/sobhanmoosavi/us-accidents