Analysis and Severity Prediction of Traffic Accidents

Aditi G S
Department of Computer Science and Engineering
PES University
Bengaluru, India
aditi07gs@gmail.com

Aditi Soori
Department of Computer Science and Engineering
PES University
Bengaluru, India
aditisoori@gmail.com

Akash Agarwal
Department of Computer Science and Engineering
PES University
Bengaluru, India
akash1729agarwal@gmail.com

***Abstract*— Not only to the growing industry of automobiles but also in the public interest, reducing traffic accidents has been a classic problem. Extensive research has been done on it over the past few decades. Using an exclusive dataset which recognizes a variety of attributes such as traffic-events, weather data, points-of-interest and time put together by paper[2][5]. The aim of this project is to perform detailed analysis of this dataset to investigate the various factors that have a potential impact on traffic accidents of varying levels of severity. We have employed the appropriate data cleaning and pre-processing methods and further performed a thorough exploratory data analysis (EDA) to gain insights on the significant road and weather conditions along with geographical data that help us classify traffic accidents on the basis of level severity. The primary goal of this paper is to use multiple machine learning algorithms to predict the severity of traffic accidents and compare the results obtained to evaluate which gives a better performance.**

Keywords - Accident severity prediction, Spatio-temporal data, exploratory data analysis, machine learning algorithms, Logistic regression, Random forest, Naive Bayes, Decision tree

I. Introduction

Across the world, almost 1.35 million people lose their lives every year due to road accidents. Road traffic injuries are a major cause of death and disability globally, with a disproportionate number occurring in developing countries. Road traffic injuries are currently ranked ninth globally among the leading causes of disability adjusted life years lost. Moreover, about 90% of the disability adjusted life years lost worldwide due to road traffic injuries occur in developing countries. The problem is increasing at a fast rate in developing countries due to rapid motorization and other factors[1].

According to numerous studies, environmental factors including the weather, the state of the roads, and the amount of light may have an effect on the likelihood of a traffic collision. Predicting traffic accidents is now more feasible owing to the recent rapid advancement of data collection techniques and the accessibility of large datasets. This is because abundant environmental data, public records transportation, and reports of vehicle crash reports can all be gathered and combined.

However, traffic accident prediction is a very challenging problem. First of all, the causes of traffic accidents are complex. Besides the common factors listed above, random factors such as vehicle mechanical problems and driver carelessness may also cause traffic accidents. Second, traffic accidents are rare events. Precisely predicting individual accidents is challenging due to lack of enough samples. Finally, the factors that may cause traffic accidents vary from place to place. For example, the main factors that lead to traffic accidents in an urban region with busy local roads might be very different from on a rural expressway. Handling the spatial heterogeneity in the data is challenging.

This paper explores the different machine learning approaches to predict severity of traffic accidents. We would like to study and comprehend the impact of various attributes that reflect the geographical data, road conditions and weather conditions on accidents. We seek to understand and perceive the correlations that exist between the attributes and the target variable - Severity.

II. Review of Literature

A. Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights

The authors of paper [2] have identified the very common and important shortcomings of not having a dataset good enough to train a model with which could be useful to make real-time predictions. These include small-scale datasets with limited coverage, being dependent on extensive sets of data and not being applicable for real-time purposes. To overcome this problem, they have created a large-scale, publicly available database of accident information named US-Accidents. They did so through a comprehensive process of data collection, integration and augmentation.

In the paper, they have employed a special deep neural network model named DAP which utilizes a variety of data attributes such as traffic events, weather data, points-of-interest and time. All these attributes are crucial to perform accident predictions with high accuracy with real time series data.

The DAP model is broken into four components:

1. Recurrent Component: this component uses Recurrent Neural Networks, specifically the LSTM model to process the set of 8 vectors, each of size 24, due to their temporal order. The output of this component is a vector of size 128.

2. Embedding Component: Given the index of a grid-cell, this component provides a distributed representation of that cell which encodes essential information in terms of spatial heterogeneity, traffic characteristics, and impact of other environmental stimuli on accident occurrence. This representation is then fed to the feed forward layer of size 128 that uses the sigmoid function as their activation function.

3. Description-to-Vector Component: this component utilizes the English description of historical traffic events in a grid-cell. This too is fed into a feed-forward layer of size 128 with sigmoid function as its activation function.

4. Points-of-Interest Component: This component utilizes points-of-interest data (a vector of size 13), which is a representation of spatial characteristics. The POI vector is fed to a feed-forward layer of size 128 which also uses the sigmoid activation function.

The last component Fully-Connected Component utilizes the outputs of all the above four components and makes the prediction.

B. Attention based Stack ResNet for Citywide Traffic Accident Prediction

The aim of this research paper[3], was to tackle the challenge of aggregating the various types of cross-domain data, which contain spatial and temporal dependencies, to predict accidents using deep learning models. Existing traffic accident predictions models (as of 2019) have used classic machine learning approaches, using historic accident records, that fail to capture the periodical patterns and trends due to the inter-correlation between temporal dimensions. In order to overcome these limitations, the author has proposed an Attentive based Stack ResNet model. The dataset used is traffic-related, cross-domain data in New York City (except Staten Island) from 2017. The dataset, which includes diverse attributes for Road Network Structure, Meteorological Data, Social Data, Human Mobility Data and Calendar Data, has been categorized into three types:

Type I: Variables spatially varied but temporally static

Type II: Variables both spatially and temporally varied

Type III: Variables only temporally varied but spatially static

The model comprises of three components, (i) CNN feature extractor: the 2017 NYC datasets are split into one-hour interval subsets with each feature assigned to its corresponding region and then encoded into vectors, (ii) Citywide Speed Inference Model: the author makes the assumption that, road average speed depend on the following region wise casualties: (1) geographically adjacent road segments tend to share similar traffic speed patterns and (2) those road segments which are geographically distant but have the same road type and functionality share similar average speed, to fill the missing speed values. Thus, the inference model can be formulated as a weighted-regression problem, (iii) ASRAP: the proposed model consists of three ResNet and three stacked CNNs structures to model the properties of Type I data in the form of multi-channel frame and Type II data in the form of feature maps. Type III data is encoded by two fully-connected layers. This attention mechanism reweights the different temporal dependencies autonomously.

The study conclusively showed that, ResNet and attention mechanisms used for accident prediction tasks, outperforms other deep learning methods in terms of MSE and accuracy rate, and reach 0.16 and 88.89% respectively. In conclusion, this paper gave us an impeccable insight into the approach that can be employed to find a solution to our problem statement, as the type of data used here and our problem statement seem to intersect.

C. Spatio-Temporal Transformer for Accident Prediction

Paper[4] seeks to forecast traffic accidents with increased accuracy based on a spatio-temporal Transformer. Compared with traditional data mining techniques, deep learning possesses the ability of using distributed and hierarchical feature representation to model complex linear phenomenon. Transformer, a deep learning method, takes a sequence as the input, scans through each element in the sequence and learns their dependencies. This feature makes the transformer intrinsically good at capturing global information in sequential data. Based on this, the enhanced spatio-temporal Transformer is able to depict both the time and spatial dependence of the traffic flow sequence.

The paper uses two sets of real large-scale high-speed traffic flow data sets, and data fusion with the real high-speed traffic accident data sets, to obtain a traffic data set with accident labels.

The proposed framework for accident prediction, ST-TAP is shown to predict accidents accurately and give corresponding risk warnings. The model is composed of four main parts, which are the input layer, the spatial pre-order codec Transformer, the temporal Transformer and the prediction layer. The input layer consists of parallel convolution neural networks for feature extraction of the time-sorted traffic speed matrix and traffic flow matrix. In the spatial-temporal Transformer module, the model captures the temporal and spatial features of the data.

It is found that the spatial features of traffic flow can be divided into static spatial-temporal dependence and dynamic spatial-temporal dependence. In order to capture the static and dynamic spatial dependence of traffic flow, the spatial Transformer is used in the spatial-temporal Transformer block first. Compared to the temporal Transformer, spatial Transformer uses static image convolution and dynamic image convolution to capture spatial dependence. So as to detect the spatial dependence, it is essential to ensure the simultaneous capture of the temporal and spatial aspects of the traffic flow at various periods since the input of this paper is a time-labeled road upstream and downstream traffic speed matrix and traffic flow matrix with distinct temporal and spatial steps. To solve this problem, the paper implements spatiotemporal position embedding in the spatiotemporal Transformer results.

The paper defines road topology as a graph constructed on the basis of physical connectivity and distance between the sensors. The static spatial dependence determined by road topology is captured using a static graph convolution network. The dynamic graph convolution captures implicit spatial dependencies that change over time like the hot spot location of traffic flow, by training and modeling high-dimensional latent spaces. To capture long-distance time dependence, the paper uses a self-attention mechanism combined with the sliding window.

Since the spatial dependence of static graph convolution and dynamic graph convolution learning cannot be directly fused, it is necessary to use a gating mechanism for feature fusion. After spatial-temporal Transformers extract the spatial-temporal features of the traffic flow, the output traffic flow sequence is used as the mapping between the input training of the convolutional neural network and the accident. The traffic mapping probability and the velocity mapping probability obtained are passed through the fully

connected layer to obtain the final output, indicating whether there is an accident.

Finally, to verify the advance of the proposed framework, the research paper compares it with a variety of existing methods. It concludes that the presented model has a shorter training time and the model optimization is realized faster. The model also obtains a higher accuracy rate because the use of dynamic graph convolution to capture the hidden space dependence of road changes over time. However, it is found that the recall rate of the model is less effective than ST-GCN i.e. Spatial-Temporal Graph Convolutional Networks.

III. Problem Statement

In real-world circumstances, traffic accidents frequently result in serious human casualties and significant economic losses. A timely, accurate traffic accident forecast has a significant deal of potential to safeguard public safety and minimize financial damages. Due to the complicated causality of traffic accidents, which involves numerous aspects such as spatial correlations, temporal dynamic interactions, and external impacts in traffic-relevant heterogeneous data, it is difficult to anticipate traffic accidents.

Classical methods of accident prediction aim at fitting regression models or other models to predict the number of traffic accidents on specific roads or certain regions. Some other work at identifying correlation between attributes (e.g., weather, road topology etc.) and the accident risk. The major drawback of these data mining techniques is that they apply to small scale traffic accident data with limited features. They also have low accuracy since they don't address unique data properties like spatial and temporal characteristics.

Recent research tries to tackle the traffic accident prediction problem using deep learning models. A major criticism that has been raised about these models is that although conventional neural network models can fit the training data with high precision, when it comes to prediction, they may produce predicted values with unacceptable variances.

The main reason for this is overfitting and neural network models that suffer from the overfitting problem generally have poor generalization ability, which limits their applicability for accident predictions.

A. Dateset

We have used the US-Accidents dataset for the course of the work[2][5]. The dataset contains more than 2.8 million cases of traffic accidents which covers 49 States of the USA from February 2016 to Dec 2021. The data attributes are mainly belong to the following categories:

- Traffic: consists of traffic events(i.e., accident, broken-vehicle, congestion, construction, event, lane-blocked, and flow-incident).
- Time: includes weekday, hour-of-day and daylight.
- Weather: comprises 10 weather attributes including temperature, pressure, humidity, visibility, wind-speed, precipitation amount and special events like rain, fog, and hail.
- POI: encompasses amenity, speed bump, crossing, give-way sign, junction, no exit sign, railway, roundabout, station, stop sign, traffic calming, traffic signal, and turning loop.
- Location: Location features include the geographic coordinates of the accident site, as well as other information such as the Street, State, City, County, and Zip Code
- Description: natural language description of the accident.
- Road Infrastructure: These features specify the road infrastructure present at or near the accident site, like Bump, Crossing, Roundabout, Traffic Signal, and others.
- Severity: Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic.

B. Data Preprocessing

The US-Accidents data considered was obtained as a single compiled file of Comma-Separated Values (CSV), of all accidents over the span of 5 years (Feb 2016 – Dec 2021). The latest updated dataset is a collection of about 2.8 million records on traffic accidents, gathered over 47 different attributes.

Due to the large size of the dataset, preprocessing the data was a significant part of the project, in order to make the data more suitable for exploratory data analysis.
The data was first checked for columns with more than 60% null instances or missing data and they were dropped (refer to *Fig.1* Appendix). Redundant columns i.e., attributes that were reflected by another attribute were eliminated. Data cleaning methods like mean replacement of NaN values in numerical columns with notable missing data were used. We dropped a few attributes that were deemed irrelevant to our analysis based on the information they provided or their format.

After removing irrelevant features, we had a mix of both categorical and numerical variables in our dataset. We also performed feature extraction and created features like Duration, Month, DayOfWeek, Hour using start_time and end_time. We performed label encoding to create binary encoding for all the unordered categorical variables.

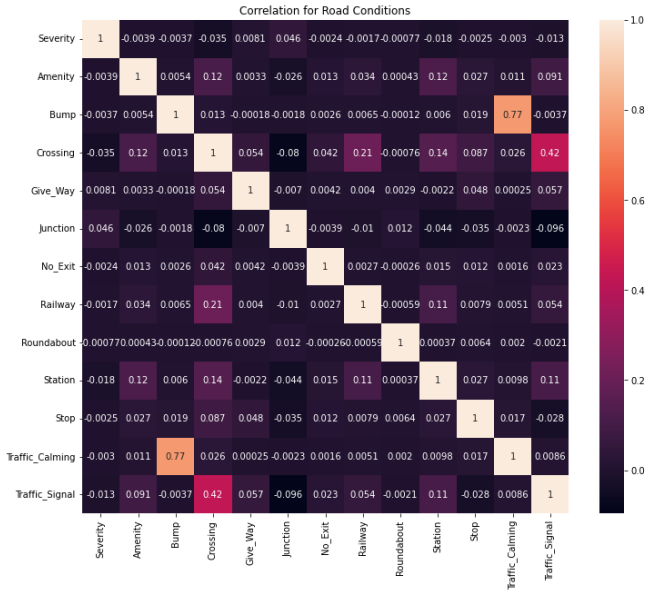
After encoding and feature extraction, we did a stratified random split based on Severity and created training (80%) and test (20%) datasets. The data split was done to handle data standardization and dimensionality reduction to avoid leaking any test and validation data information in training our models.

C. Exploratory Data Analysis

We began our analysis of data by exploring the relationships shared among different attributes and between severity of the accidents. Initially, we approached our analysis in two directions:

- analyze the potential impact of weather conditions on accidents
- analyze the potential impact of road conditions on accidents

We employed heatmaps to view the correlations between the attributes reflecting road conditions and between the roadway features and accident severity.

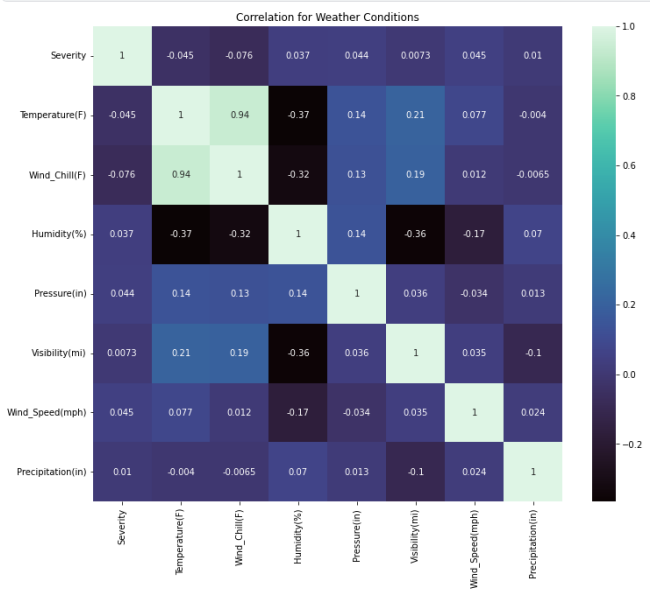


Certain relationships between the attributes stand out:

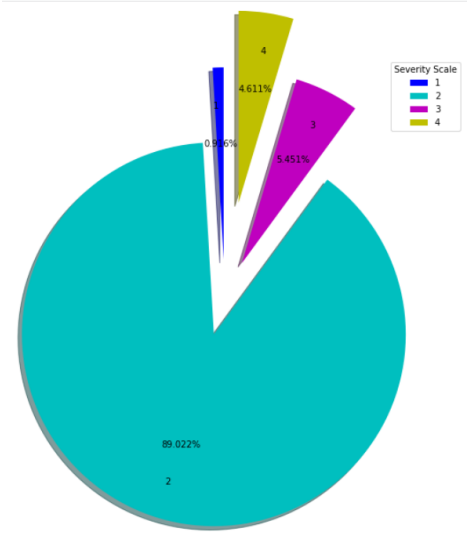
- Traffic_Calming and Bump
- Traffic_Signal and Crossing
- Railway and Crossing

have positive correlations with fair strength. These correlations are to be expected as these road features are often found together. However, our main focus is the relationships the different attributes share with the accident severity. It was observed that the majority of the road attributes have a negative, weak correlation with severity. Junction is the only feature to exhibit a positive correlation with severity. We agreed to drop the columns displaying extremely weak relations with severity as they won't contribute much information to our analysis.

We later plotted the frequency of accidents against the weather conditions for every level of severity and we noticed most of the accidents have occurred when the weather condition was either Fair or Mostly Cloudy (refer *Fig.2* Appendix). Temperature(F) and Wind Chill(F) exhibit a strong positive correlation. Humidity(%) has a negative correlation of considerable strength with Temperature(F), Wind Chill(F) and Visibility(mi).

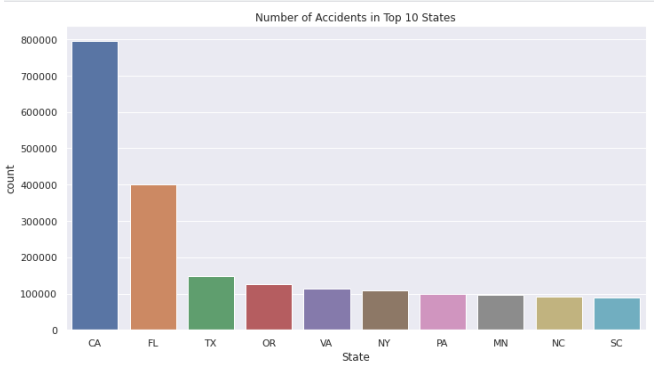


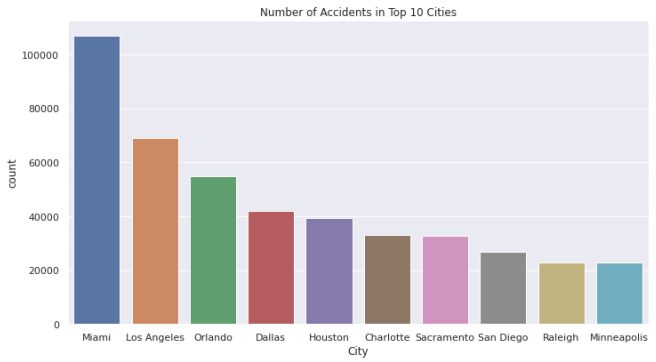
Since our analysis and prediction revolves around the attribute Severity of accidents, we viewed the distribution of accident severity.

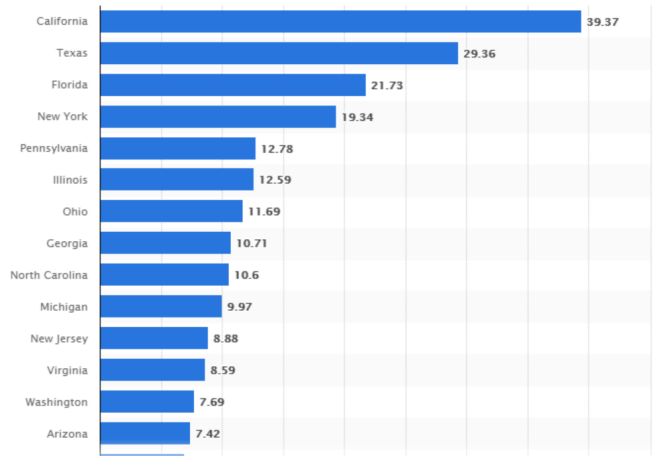


An overwhelming number of instances are classified as Severity Level 2 and scarcely as Severity Level 1. This might pose a problem when we train our models. To overcome this issue, we used a combination of undersampling and oversampling on our training sets to create a more even distribution among our target variable.

We further plotted the top 10 states and cities that have the maximum occurrence of accidents. Since we didn't have demographic data, we had to refer to an external source[6] to investigate whether population impacts accidents frequency. The comparison of our results with the demographics aligned with our assumption that population can be one of the factors of accidents.







D. Fitting the models and Evaluation

The goal of this project is to develop a model which predicts accident severity with acceptable accuracy.
Our approach differs from current work on the topic where we plan to make use of other spatial and temporal features to predict accidents.
In this section, evaluation of the performance and reliability of the model, and comparing the proposed approach with the conventional models will be discussed briefly.

For the prediction task , we evaluated six models :

1. Logistic Regression: Using gradient descent, logistic regression's fundamental form generates weights for its covariates. A probability is initially calculated by feeding the value into the logit function rather than utilizing a linear term. The values that gradient descent seeks to maximize are these. The regression equation can then be multiplied with our results to determine if it is 0 or 1, based on a threshold that we choose, after weights and a bias term have been chosen.
On prior analysis,it was found that a certain number of features resulted in high correlation. This observation motivated dimensionality reduction using Principal Component Analysis (PCA). Logistic regression model was trained with PCA as well as without PCA.

2. Naive Bayes: It is a basic method for building classifiers, consisting of models that give class labels to problem occurrences, represented as vectors of feature values, where the class labels are chosen at random from a finite collection. Naive Bayes is using Bayes Theorem which states that the posterior probability is equal to the conditional PDF, times the prior probability, divided by the normalizing PDF or the aggregate of the PDFs for both classes. The model chooses the larger value of posterior probability given the data for a cell of that row and column. In our approach we used multinomial naive bayes and transformed each feature individually by scaling it such that it is in the range of the training set.

3. Decision Tree: selecting the most biased feature and comprehensible nature. It is also easy to classify and interpret easily. Also used for both continuous and discrete data sets. After instantiating a DecisionTreeClassifier with a maximum depth of 8, Gini Impurity and Entropy methods were used to build an appropriate decision tree for selecting the best splitter.

4. Random Forest: The main concept is that bagging, a method of combining and averaging the output of those trees, is used after random forest produces multiple decision trees. Then it continues to prefer some tree groups while removing others. RandomForestClassifier was used to fit 100 decision trees and train the model using the training sets.Further a new random forest classifier was also created for the most important features and the accuracy of the full and the limited feature models were compared.

5. ANN: They are multi-layer fully-connected neural nets consisting of an input layer, multiple hidden layers, and an output layer. Every node in one layer is connected to every other node in the next layer. We make the network deeper by increasing the number of hidden layers. In our approach we have implemented 3 hidden layers with the ReLu activation function, 1 input layer and an output layer with the softmax activation function.The batch size is 40 and the epochs have been set to 100 and an adam optimizer is used for the best accuracy.

6. KNN: K Nearest Neighbor is one of the fundamental classification machine learning models which classifies new data points based on the similarity measure (Euclidean Distance) of the known data points. With respect to our dataset, KNN works the best under the assumptions that we have properly labeled and small dataset. Since different regions have different environments, we assumed subsetting by region might produce more accurate results as environment specific features will be similar if region specific and hence strengthening the impact of predictors by controlling the variations.We created a smaller dataset by

subsetting by specific region, specifically, city with maximum accidents for our KNN model. After evaluating our model with the dataset, along with parameter tuning for varying K values, we observed that the model yielded better results when k_nearest_neighbors = 2 with about 9.8% better accuracy. This performance is not robust, however, it was able to classify more than half of the instances accurately.

IV. Experimental Results

Fig. 1. PRACTICAL RESULTS

MODEL	Logistic Regression (with PCA)	Logistic Regression (without PCA)
ACCURACY	0.944	0.9439
TIME TAKEN(in seconds)	111.732	1.410

Fig. 2. PRACTICAL RESULTS

MODEL	Naive Bayes	ANN
ACCURACY	0.926	0.9433
TIME TAKEN(in seconds)	0.230	137.813

Fig. 3. PRACTICAL RESULTS

MODEL	Decision tree with entropy as information criterion	Decision tree with gini as information criterion
ACCURACY	0.959	0.959
TIME TAKEN(in seconds)	26.358	23.811

Fig. 4. PRACTICAL RESULTS

MODEL	Random Forest	Random Forest (Only important features)
ACCURACY	0.966	0.965
TIME TAKEN(in seconds)	579.0821	209.0594

Fig. 5. PRACTICAL RESULTS

MODEL	K Nearest Neighbor
ACCURACY	0.638
TIME TAKEN(in seconds)	229.7641

V. Conclusion

This paper primarily focuses on severity prediction for traffic accidents which is a key step in traffic accident management.
Results show that the Random Forest Model performed better with higher accuracy rate, while the Naive Bayes model performed better in terms of time used in the building of the model.Although the Random Forest Model performed better than other models, it performs poorly in terms of time taken on this dataset.Compared to RF, Logistic Regression and Decision tree classifier perform admirably.However they are less accurate.

In this proposed study, we have implement some basic machine learning algorithms to classify the severity of Traffic accidents and presented valuable insights on
The implementation of machine learning is a functional and a great approach to take an accurate decision to tackle the issue of road accidents and the findings of the analysis part (Section C. Exploratory Data Analysis) can be suggested to traffic authorities for reducing the number of accidents.

In future work, the approach presented in this study for predicting and analyzing road traffic accident data, can provide potential usage on road traffic accident dataset for India and other countries globally.

Acknowledgment

We wish to extend our special thanks to Prof. Anand MS and the team of Data Analytics Course, for providing us guidance, suggestions and resources over the course of our project. We would also like to acknowledge the Department of Computer Science and Engineering at PES University, for encouraging us to gain more practical knowledge and for providing us the exposure and opportunities in the industrial and research fields.

References

[1] Nantulya VM, Reich MR. “The neglected epidemic: road traffic injuries in developing countries.” BMJ. 2002 May 11;324(7346):1139-41. doi: 10.1136/bmj.324.7346.1139. PMID: 12003888; PMCID: PMC1123095.

[2] Sobhan Moosavi, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, 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 (2019),* URL: https://doi.org/10.48550/arXiv.1909.09638.

[3] Z. Zhou, "Attention Based Stack ResNet for Citywide Traffic Accident Prediction," *2019 20th IEEE International Conference on Mobile Data Management (MDM)*, 2019, pp. 369-370, doi: 10.1109/MDM.2019.00-27.

[4] W. Liu et al., "ST-TAP: A Traffic Accident Prediction Framework Based on Spatio-Temporal Transformer," 2021 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), 2021, pp. 360-365, doi:10.1109/DASC-PICom-CBDCom-CyberSciTech52372. 2021.00068.

[5] Moosavi, , Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, arXiv preprint arXiv:1906.05409 (2019).

[6] URL: https://www.statista.com/statistics/183497/population-in-the-federal-states-of-the-us/

Appendix

A. Contributions

Aditi G S (PES120CS015) : Literature survey, exploratory data analysis, data pre-processing and cleaning, undersampling and oversampling the training dataset, fitting and validating KNN model, analyzing the results.

Aditi Soori (PES1UG20CS017) : Literature survey, data pre-processing and cleaning, fitting and validating Logistic regression, Decision tree, Random Forest, ANN and Naive Bayes models, analyzing the results.

Akash Agarwal (PES1UG20CS026) : Literature survey

B. Data Visualizations

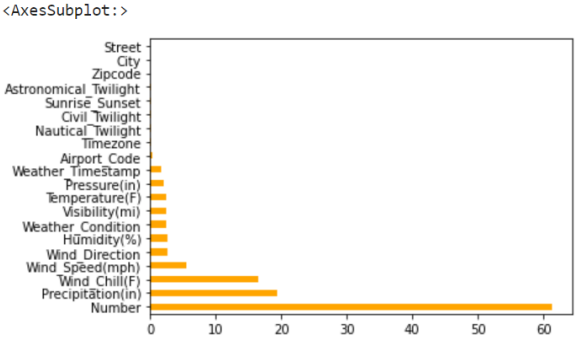


Fig. 1. Percentage Distribution of Missing Data

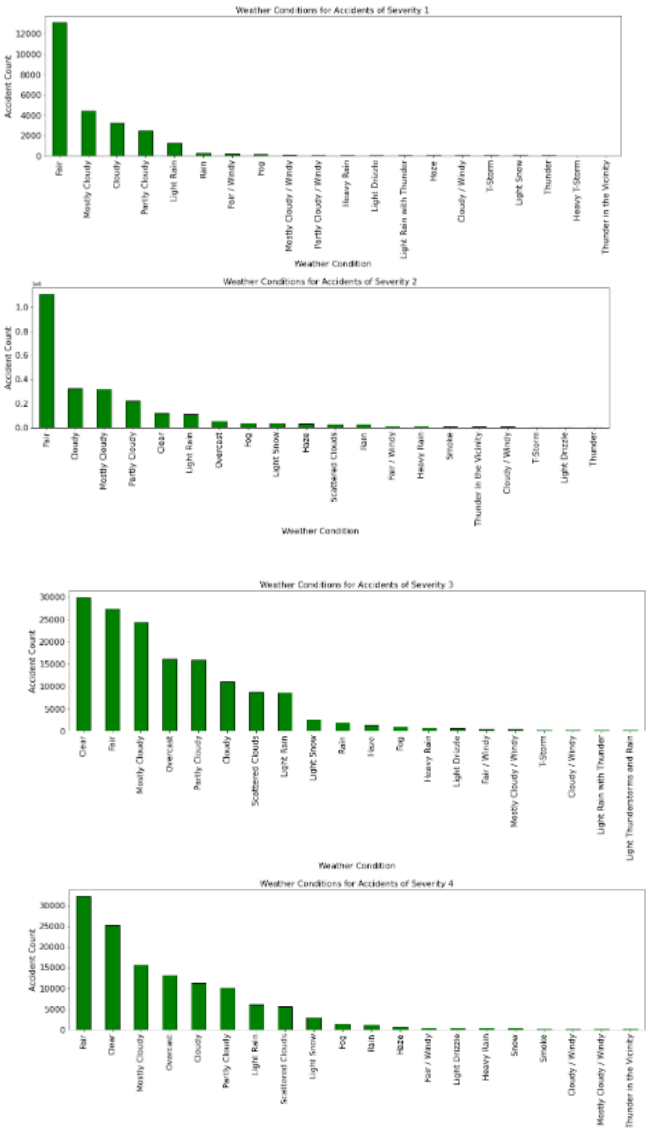


Fig. 2. Accidents frequency under different Weather conditions for all levels of Severity

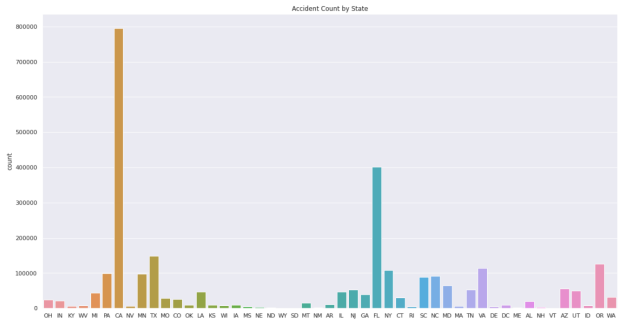


Fig. 2. State-wise Accident Frequency