US Accidents Control Data analytics project

Sahana G *PES1UG19CS412*
Sanjana G *PES1UG19CS430*
Sathvik K *PES1UG19CS435*
Sharan S *PES1UG19CS450*

***Abstract*—Road crashes are the single greatest annual cause of death of healthy U.S. citizens travelling abroad.Roadway traffic safety is of concern for transportation agencies as well as ordinary citizens in today's world. Hence in order to give safe driving suggestions, careful analysis of roadway traffic data is essential to find out parameters that are closely related to fatal accidents.Therefore, reducing traffic accidents is an essential public safety challenge today;leading to accident analysis being a subject of much research of present times.**

The prediction uses predictive modelling techniques and supervised learning to identify risk and key factors that contributes to accident severity.The prediction uses publicly available data from US department of transport.

***Index Terms*—safe driving suggestions, reducing traffic accidents, predictive modelling, supervised learning.**

I. Introduction

The aim of the prediction is to analyze the US accident data from various states to inform the US government agencies and the general public on trends and possible causes of traffic accidents and what could be done to reduce them.

The dataset includes various attributes which contribute towards the severity of the accident such as the weather conditions including wind speed, temperature, humidity, pressure, precipitation and other factors such as visibility, traffic signals, junctions etc.

It is disheartening and saddening to allow its citizen to die by a road accident. Consequently, to handle this critical situation, an apt analysis is required.This can be done by analyzing the models using machin elearning approach.

The basic idea is to train some models on the dataset and use those models to get the feature importance to figure out which factors contribute the most to an accident. These can be weather, points of interest at the region, time of day, month of the year, and the location of the driver as some areas are more prone than others.We are trying to figure out those significant parameters leading to an accident and lay some beneficent suggestions regarding the issue.

The final goal is to predict and analyze the most effective parameters that decreases the severity of accidents and present the most accurate prediction model for accidents.

It will be of great significance if the model predicts the outcomes rightly. This would help in preventing the number of accidents taking place.The prediction would help people to decide whether to drive or not based on various weather conditions such as precipitation, wind speed, pressure etc and other factors such as visibility, traffic.If the model predicts correctly that there are high chances of an accident occurring when any of the weather conditions are not favourable such as high wind speed, high humidity or high precipitation during a particular time period and more visibility, less traffic which are favourable then it would be of immense help for the citizens to skip an accident which would be likely to happen otherwise.

II. Literature Survey

Based on the works that are present, from [1] there are two different approaches namely using logit model and Artifical neural network approach (ANN).The use of factor analysis and logit model simultaneously could help in achieving the most comprehensive and efficient model specifying the major contributing parameters and their effects on the accidents and hence assisting the public to take effective measures and precautions to lessen accident impacts and improve road safety.

In ANN approach classification is based on considering Pattern recognition which works by classifying input data into objects or classes based on prominent features, using either supervised or unsupervised classification.The efficiency of model is checked by confusion matrices and ROC curves.

From [2] this work was done on Bangladesh road accident dataset, the authors have used the four most popular machine learning techniques for the road accident analysis including Decision Tree are used with the purpose of identifying the relationships between the variables of the dataset.This consists of various stages and can be called as an hierarchy with binary decisions, Kth nearest neighbour (KNN) - is a classification algorithm which is based on parameter similarity. It analyzes the data and measures the distance between the clusters and the similarities between the data and cluster them based on K values , Naïve Bayes - classification technique is based on the Bayes theorem. It predicts the probability of different class based on various attributes and provides a new class to the highest probability, AdaBoost - a boosting algorithm. Every point is weighted in the training dataset. To get the most prominent features, we experiment by applying three different algorithms of feature selection. They are Univariate Feature Selection - which explores each feature severally and only selects the best features based on univariate statistical tests like chi-squared test, Recursive Feature Elimination - which works removes the features recursively and uses the accuracy of the model to consider the features that are of high significance to predict the desired parameter, and Feature

Importance - which is a trained, supervised classifier to come up with the critical feature. It works as a classifier and evaluates each parameter to generate splits.The performance of each algorithm has been calculated for four accident severity classes 1 to 4 ie Fatal or Grievous or Simple Injury or Motor Collision).High accuracy of about 80 percent is obtained using Naive Bayes and Adaboost approaches.Compared to all the approaches Ada-Boost gives the best result beacause it uses iterative classification on Decision Tress.

From [3] this work was done on Sri Lanka road accident dataset , the authors have used Deep neural network (DNN) as classifier and decision tree . DNN classifier is predominantly used for binary classification and the multiple output classification. All the classification was made on different class. on that particular dataset decision tree performed well than DNN.

From [4] the authors have used decision tree approach , the most excessively adopted algorithms in Decision Trees is ID3 which mainly uses Entropy and Information Gain to determine node impurity, based on the Information gain values the decision tree is constructed .The performance evaluation of the model is done using the confusion matrix and values of accuracy, precison. Time series calendar model is used which is useful when the severity of the accident is more than normal time series data. Combination of Enhanced Decision Tree model the Time series calendar heatmap can be used to get efficient results.

A. Observations

As we see the performance evaluation is done by confusion matrices and values of accuary and precision , for any model designed for any test data how correctly the model predicts the output could be checked by confusion matrix . Like the work done in [2] we could use any supervised learning classifier model as our model considering the class label severity.

III. Initial Insights / Data Visualization

The data of Road Accidents in US was analysed and visualized. The dataset had records of various accident incidents covering almost 3 years.

There were totally 1516064 records spanning over 47 columns.

The data taken for consideration initially was not cleaned and had missing values.

It can be noticed that there are a large number of NAN values in **number** feild followed by wind chill and wind speed. There are few missing values in Humidity, Weather Condition, Wind direction, Pressure, and Weather Timestamp feilds as well which have to be either filled using some techniques or ignored.

The countries which are under attention due to large number of accidents are depicted in the bar graph below.

The graph shows the top 20 countries where the threat of an accident to occur is more.It can be concluded that Los Angeles has maximum number of accidents taking place followed by Miami.This means that only 5 percent of countries have less then 500 yearly accidents.

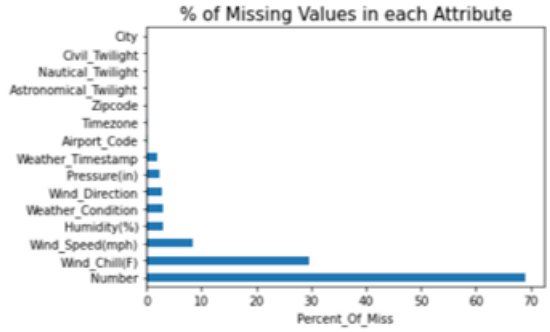

Fig. 1. percentage of missing values in each attribute.

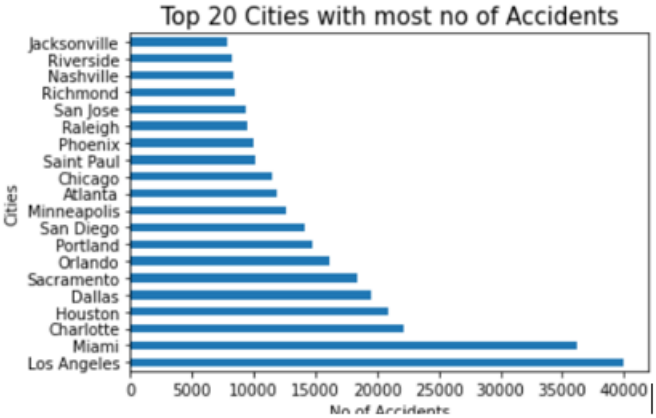

Fig. 2. Top 20 countries with most no. of accidents.

The relation between number of cities and accidents is depicted below.

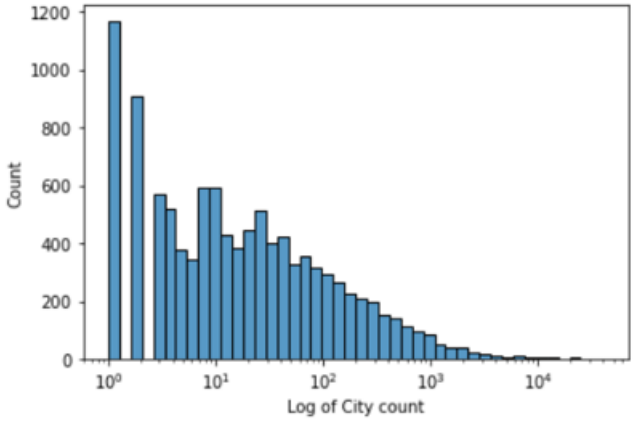

Fig. 3. Relation between number of cities and accidents.

It is observed that majority of the cities have only 1 accident and number of accidents per city decreases exponentially.

The number of accidents per hour and that per day are being analysed in the graph below.

From the graph below it can be noticed that large number of accidents are taking place during 9:00 and 10:00 AM in the

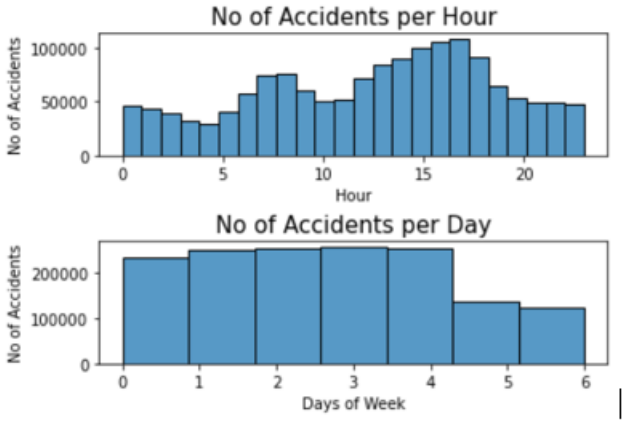

Fig. 4. Number of accidents per hour and per day.

morning which is the office reporting time and during 5:00 to 6:00 PM in the evening which is home returning time.

It is also observed that the number of accidents are more during weekdays ie from 0 to 4 when compared to weekends is 5 and 6.

The graph below shows the number of accidents on weekend and weekday.

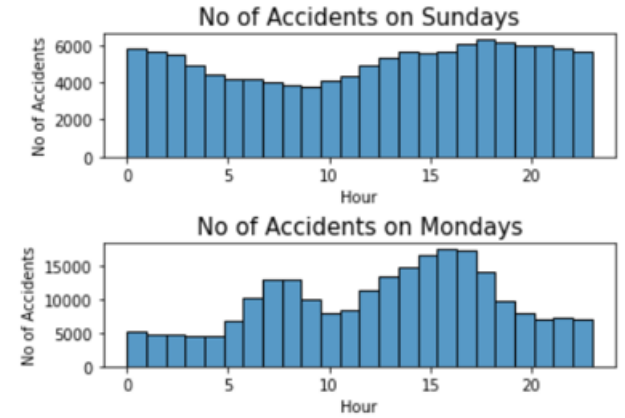

Fig. 5. Number of accidents on Sunday and Monday.

The number of accidents on weekends is constant between the span of a day whereas on weekday ie Monday number of accidents are more during the office reporting and home returning time as stated earlier.

The relation between Sunrise, Sunset and Severity of the accident is depicted in the graph below.

The values of Severity namely 1 and 2 are considered as 0 and that of 3 and 4 are considered as 1 in the graph above. It is noticed that the number of accidents are more in day time when compared to night which is contradicting.The severity of accidents is also more during day when compared to night.It implies that people are more cautious during night and are careless while driving in sunlight.

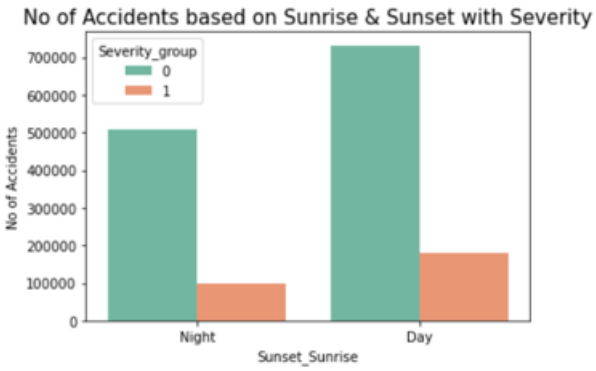

Fig. 6. Severity of an accident v/s Sunset and Sunrise.

The Heatmap below shows the most accident prone areas of US. The areas shown in red colour are the most accident

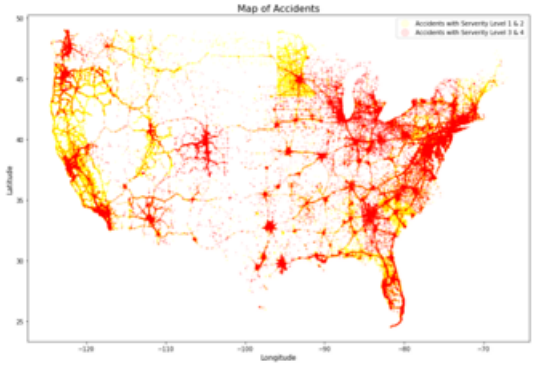

Fig. 7. Heatmap showing accident prone zones.

prone zones including New York, Vatican City of US and the with yellow colour in the middle is the barren land where probability of an accident is less.

The graph below shows the impact of weather conditions on the number of accidents. It is observed that the number of accidents are more when the weather is fair and less when it is cloudy which is again contradictory. It can be interpreted as people are more cautious when the weather is bad and are less prone to experience an accident and are carefree when the weather is fine on the other hand.

IV. Problem Statement

Predicting the least accident-prone environment and factors, least affected conditions(less severity).

Basically predicting the influence of all other factors on value of severity of accident this model will be able to answer any kind of questions of traveller.A traveller might always have a doubt that taking the particular route will he be able to reach destination possible , how safe is it to travel this way, what precautions should be taken while he travels through the route.

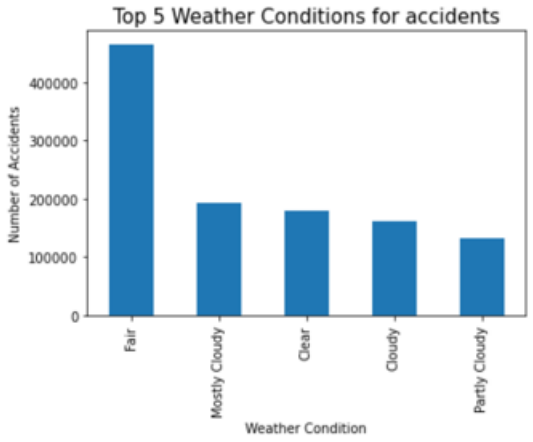

Fig. 8. Weather condition v/s Number of accidents.

Questions we are trying to attempt to answer are , is it safe for a driver to travel to the specific destination with specified condition? After meeting with an accident, how badly has the driver been injured ? will the basic first-aid suffice his injury or should he be taken to hospital? Which day and time are safe to travel? How can the State Government improve accident-prone infrastructure?

Dataset we used is from kaggle where it had columns like start and end time of accident, latitude and longitude which is required for positions where accident took place, distance, which side of the road did the accident happen and time , street , country , zipcode where accident took place , airport code nearest airport and all weather condition during accident like humidity, temperature, direction of wind, wind chill, pressure, wind speed, precipitation, period ie during sunset or sunrise etc and some features about the place where accident took place such as was a crossing, bump, amount of traffic, traffic signal, was it in junction , was there any station nearby and different measures to find the period such as natural twilight, astronomical twilight, civil twilight etc which really are a reason for an accident to occur.We are mainly interested in finding how severity is influenced by all these factors.

There are basically three types of algorithms in machine learning namely supervised, unsupervised learning, and re-inforcement learning. Among these three broad categories of algorithms of machine learning, the classification method approaches the supervised learning category.From the dataset we know that severity value can be between to 1 to 4(Motor Collision/Simple Injury/ Grievous/ Fatal), where 1 being least injured and 4 being highly injured.Any new test data given could be classified into any class of severity.The classification is done by model that uses supervised learning to learn from training data.

V. BUILDING MODEL

The data set we choose was split into two parts training and testing sets.Our problem statement was predicting the severity value. It comes under regression and supervised learning.We use the training set to fit the model.We have used a supervised regression model K-nearest neighbour.As our problem statement deals with regression so we consider only the non categorical columns as the basis for classifying instances. K nearest neighbour classifies based on instances that are near to them and Decision tree is like a flowchart to the terminal nodes presenting classification of outputs.

VI. CONCLUSION

After analysing the data set , We built a model using K-nearest-neighbour(KNN),where we predicted the severity value(real number) based on other attributes present in our data set.Using the train data the model created gives accuracy of 78 percent for the selected test data and other model we built using Decision Tree gives accuracy of approximately 72 percent.

REFERENCES

[1] Prediction and Analysis of the Severity and Number of Suburban Accidents Using Logit Model, Factor Analysis and Machine Learning: A case study in a developing country
[2] Road Accident Analysis and Prediction of Accident Severity by Using Machine Learning in Bangladesh
[3] DNN Classifier and Decision Tree-Based Novel Methodology for Analyzing Road Accidents
[4] Predicting Road Traffic Accident Severity using Decision Trees and Time-Series Calendar Heatmaps
[5] The study of technological prevention method of road accident related to driver and vehicle
[6] Survey of accident detection systems