Road Accident Analysis

DEKKA SRIDHAR
B.E. CSE AIML
APEX INSTITUTE OF TECHNOLOGY, CHANDIGARH UNIVERSITY
NH-95, Ludhiana - Chandigarh State Hwy, Punjab 140413, India
sreedharsmartshii@gmail.com
Supervisor : Arun Mittal

ISHIKA JAIN
B.E. CSE AIML
APEX INSTITUTE OF TECHNOLOGY, CHANDIGARH UNIVERSITY
NH-95, Ludhiana - Chandigarh State Hwy, Punjab 140413, India
ishikajaindec02@gmail.com
Supervisor : Arun Mittal

AKHIL KUMAR
B.E. CSE AIML
APEX INSTITUTE OF TECHNOLOGY, CHANDIGARH UNIVERSITY
NH-95, Ludhiana - Chandigarh State Hwy, Punjab 140413, India
akhilkumar3837@gmail.com
Supervisor : Arun Mittal

ABSTRACT:

Traffic accidents are one of the world's most serious concerns, as they cause a large number of casualties, injuries, and deaths each year, as well as significant financial losses. Accidents on the road are caused by a multitude of factors. If these aspects can be better understood and predicted, measures to decrease the harms and their severity may be conceivable. The purpose of this project is to develop models that can choose a set of influential factors for use in classifying the severity of an accident and supporting data analysis. Furthermore, this study gives a forecast model for future traffic accidents based on previous data. Number of automobile accidents that occur each year in India, it has become a major issue. It is both unacceptable and terrible to allow civilians to perish in road accidents. As a result, in order to deal with this overcrowded situation, a precise analysis is required. These patterns of behavior and roadway usage can be used to establish traffic safety policies. Measures must be based on scientific and objective research on the causes of accidents and the severity of injuries. Using machine-learning methodologies, the system gives several models for predicting the severity of injuries sustained during traffic accidents. We looked at networks that have been trained utilizing learning methods. The outcomes of the number of automobile accidents that occur each year in India, it has become a major issue. Experiments show that many methods to machine learning paradigms were look into traffic accidents in greater detail in order to assess the severity of accidents using machine learning approaches. We also highlight the key factors that have a direct impact on road accidents and offer some practical advice on the subject. Despite the efforts of automobile engineers and researchers to create and build number of automobile accidents that occur each year in India, it has become a major issue.

Keywords:

machine learning(ML); , Analysis; Prediction; data analysis; road accident data; Data pre-processing; decision trees (DT); random forests (RF); logistic regression (LR).

INTRODUCTION:

The number of road accidents that occur each year around the world, according to the World Health Organization's fatality numbers, is scary. Every year, 1.2 million people are murdered and 50 million are injured in traffic accidents. Over 3,300 people are killed and 137,000 are injured every day. The high frequency of traffic accidents puts human life and property safety at jeopardy, resulting in direct economic losses of 43 billion dollars. Road accident prediction is one of the most important research fields in transportation safety. Road geometry, traffic flow, driver characteristics, and the road environment all play a role in the occurrence of road traffic accidents. There have been numerous research on danger location/hot spot detection, accident injury severity analysis, and accident length analysis.

One of the most important research disciplines in transportation safety is road accident prediction. The occurrence of road traffic accidents is influenced by factors such as road geometry, traffic flow, driver characteristics, and the road environment. In order to anticipate accident frequency and study traffic accident aspects, several studies on hazardous location/hotspot detection, accident injury severity analysis, and accident investigation have been done. Take a look at the timeline. Accident causes are the subject of some studies. The weather, as well as the lighting conditions along the journey, are also factors to consider. According to the World Health Organization's fatality statistics, the number of traffic accidents that occur each year around the world is frightening. Because various nonlinear components, such as people, cars, roads, and climate, have a role in traffic accidents, traffic accident prediction is critical in integrated traffic planning and management. Due to noise pollution and a lack of data, the usual approach of linear analysis is unable to depict the genuine situation, resulting in an unacceptable prediction result. The problem of accident is especially acute in highway transportation due to the complex flow pattern of vehicle traffic and the existence of mixed traffic with pedestrians. Accidents occur for a variety of reasons and can occur at any time. Traffic accidents result in the loss of lives and property. As a result, traffic engineers bear a substantial amount of responsibility for guaranteeing road user safety through ensuring safe traffic movements. The accident rate can be reduced even with minimal resources by using good traffic engineering and administration. As a result, it is critical to conduct a thorough examination of traffic incidents. The development of preventative

design and control measures will be aided by a detailed examination of the accident's cause. Crash data gathered on the site by police personnel is a valuable source of highway safety statistics. Police have an exceptional capacity to gather crash data on the scene as soon as it happens, as well as ephemeral data that may deteriorate (such as tyre marks) or be destroyed. Police personnel are in a unique position to collect crash data, but this is not their only responsibility. On-scene responsibilities include securing the incident site, caring for injured persons, and re-establishing traffic flow. As a result, while adopting new technologies, on-scene data collection systems must account for officers' needs. After a set of data is acquired, Accident 5 sites are chosen for prospective treatment, and more information is usually needed before a decision can be made about which sites will be addressed and what type of improvement work is needed. This additional information should be tied to both the site accident data and other components that may aid in determining the nature of the problem at the site, and should be acquired during site visits.

Due to the complex flow pattern of vehicle traffic and the presence of mixed traffic with pedestrians, the problem of accident is especially acute in highway transportation. Accidents happen for a variety of reasons and at random. Accidents in traffic result in the loss of lives and property. As a result, traffic engineers must shoulder a significant amount of responsibility for ensuring the safety of road users by providing safe traffic movements. Although achieving zero causality is challenging, and some consider it unattainable, it is possible to lower it to fractions per 100,000 population using the latest technologies and advances in the field. This will necessitate a big expenditure.

Uncertainty and unpredictability are inherent in road and traffic accidents. These accidents are caused by a variety of factors, including: no priority for pedestrians, no priority for cars, unlawful pedestrian crossings, bicycle rider deviations, speed not appropriate to road conditions, driver deviations of animal traction vehicles, and so on. It might be claimed that road and traffic accidents are defined by a number of variables, some of which are well-known and others which are less so. Accident reduction and road safety are key public health problems. This assertion is backed up by data: Every day, about 3000 individuals from all over the world die as a result of traffic congestion. These vehicle accidents result in $518 billion in annual global economic losses due to road traffic damage. The country's economic imbalance is exacerbated by these massive losses. Road accidents are projected to cost $100 billion annually in developing countries. Road accidents have an impact on the demography of any country in addition to their economic impacts. In this context, each country should focus on developing methods to counteract these effects. Injury reporting in official accident statistics is incomplete at all levels of severity of injuries, according to 49 research from 13 nations. These reports discovered discrepancies when compared to the actual scenario. It was discovered that 95 percent of the reported cases had the following lesions: 70% major injuries, 20% minor injuries, and 5% extremely small injuries.

Accident-related deaths and injuries are expected to be a worldwide problem. Since the dawn of the motor age, almost a century ago, traffic safety has been a major concern. Every year, it is estimated that over 300,000 people die and 10 to 15 million people are injured in traffic accidents all over the world. Statistics also demonstrate that road accident mortality is very high among young individuals, who make up the majority of the workforce. Various road safety tactics, procedures, and countermeasures are required to address this issue. The survey looked into a variety of injury-related deaths. According to research published by the World Health Organization (WHO), road traffic accidents account for the majority of mortality among those aged 15 to 29, with more than 1.25 million people dying each year. According to a WHO assessment, some of the fundamental causes include a lack of training institutes, bad road conditions, and inadequate traffic management. To address this problem, a systematic approach and firmly founded solution with efficient and effective methods is required. As a result, when our system encounters such parameters, it provides a systematic and visual representation of how to overcome and explain the situation. Automobile engineers and researchers have attempted to design and build safer vehicles, yet traffic accidents are inescapable. The development of a prediction model that automatically determines the type of injury severity of diverse traffic accidents could discover patterns involved in severe wrecks. The formulation of traffic safety control policies can benefit from these behavioral and roadway trends. Measures must be based on scientific and objective research on the causes of accidents and the severity of injuries. Using machine-learning methodologies, the system gives several models for predicting the severity of injuries sustained during traffic accidents. We looked at networks that have been trained

utilizing learning methods. The outcomes of the experiments show that many methods to machine learning paradigms were investigated.

METHODOLOGY:

Machine Learning Model:

We employed decision trees, random forests, and logistic regression, as well as hyperparameter tweaking, to improve its efficiency. The Random Forest approach, with an accuracy of 86.86 percent, was chosen as our model. The model has been run, and the severity has been calculated. There are three severity levels: 1 (fatal), 2 (severe), and 3 (very serious) (minor). The output is visible to the user and is sent back to the front-end. The police receive an SMS with the accident's geographical coordinates and severity so that they can take precautionary action on the scene. Creating a Virtual Machine: It uses a machine learning algorithm that has been taught and tested. Decision Tree, Random Forest, and Logistic Regression Classification Algorithms: The decision tree, random forest, and logistic regression classification algorithms have all been implemented. Hyperparameter tweaking was performed to attain the best level of precision. Because random forest has a maximum accuracy of 86 percent, it was chosen as the model for the web app. Applied technologies include:

➢ ***ANACONDA:*** Anaconda is a Python and R programming language distribution for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, and so on), with the goal of making package administration and deployment easier. Data-science packages for Windows, Linux, and macOS are included in the release. Anaconda, Inc. is responsible for its development and upkeep.

➢ ***VISUAL STUDIO CODE:*** Visual Studio Code is a lightweight but capable supply code editor for Windows, macOS, and Linux that runs on your desktop. It has JavaScript, TypeScript, and Node built-in support.

➢ ***PYTHON:*** Python is an interpreted high-level programming language that may be used for a wide range of activities. Python features an automatic memory management system and a dynamic type of system. It gives for a wide range of programming possibilities. There are 11 of them, including object-oriented, imperative, functional, and procedural paradigms. It also includes a substantial standard library. Software developers, analysts, data scientists, and machine learning engineers all use it, making it the world's most popular and fastest-growing programming language.

➢ ***NUMPY:*** NumPy is the most important Python package for scientific computing. It includes a powerful N-dimensional array object as well as sophisticated (broadcasting) features. helpful linear algebra, Fourier transform, and random number capabilities tools for integrating C/C++ and Fortran code NumPy can be used as a multidimensional container of generic data in addition to its apparent scientific applications. It is possible to define any number of data kinds. This enables NumPy to work with a wide range of databases with ease and speed. Because they are both interpreted, NumPy in Python provides functionality comparable to MATLAB, and they both allow the user to construct fast programs if most operations are performed on arrays or matrices rather than scalars. SciPy is a library that adds more MATLAB-like functionality and Matplotlib is a plotting package that provides MATLAB-like plotting functionality. NumPy is licensed under the BSD license, enabling reuse with few restrictions.

➢ ***SCIKIT-LEARN:*** Scikit-learn is a Python-based machine learning library that is available for free. It includes support vector machines, random forests, gradient boosting, k-means, and DBSCAN, among other classification, regression, and clustering techniques, and is designed to work with the Python numerical and scientific libraries NumPy and SciPy. - Simple and effective data mining and data analysis tools that are accessible to everyone and can be reused in a variety of situations. - Built using NumPy, SciPy, and Matplotlib - Commercially useable, open source - BSD license

➢ ***API:*** Application Programming Interface (API) In basic terms, APIs just allow applications to communicate with one another and data to one another. The following APIs are used:

1. The Geolocation API returns a location and accuracy radius based on data from cell towers and Wi-Fi nodes detected by the mobile client. This page explains the protocol 12 for sending data to the server and receiving a response from the server. POST is used to communicate over HTTPS. Both the request and the answer are in JSON format, with application/json as the content type.
2. Open Weather Map's Weather API gives you access to current weather data, 5- and 16-day forecasts, UV Index, air pollution, and weather conditions, among other things. Text

Local provides the SMS API. It's simple to integrate with any application and you can start sending SMS in minutes.

➢***GOGGLE MAPS:*** We can use Google Maps to discover a popular restaurant, calculate the shortest driving route and travel time, and so on. You're probably wondering how to accomplish it programmatically in Python if you've been provided a huge amount of location data to check and validate against a map and then select the optimal routes. We'll look at how to do all of these things with Python Google Maps APIs in this tutorial.

You can use the Google Static Maps API to embed a Google Maps image on your website without having to use JavaScript or load a dynamic page. The Google Static Maps API creates a map using URL parameters from a standard HTTP request and returns it as an image that can be displayed on a website.

ROAD ACCIDENT DECISION TREES AND RANDOM FORESTS:

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (for example, whether a coin flip will come up heads or tails), each branch reflects the test's conclusion, and each leaf node represents a class label (decision taken after computing all attributes).

The categorization rules are represented by the pathways from root to leaf.

A decision tree consists of three types of nodes:

Decision nodes – typically represented by squares

Chance nodes – typically represented by circles

End nodes – typically represented by triangles

Machine Learning can be used to create high-performing classification systems from a group of representative samples of a data population. Combining an ensemble of individual classifiers to produce a unique classification system, known as Classifier Ensemble, is an efficient technique to deal with this type of challenge. Since the early 1990s, research has demonstrated that various combination principles, such as boosting (or arcing), bagging, random subspaces, and, more recently, Random Forests, are extremely efficient. Combining classifiers efficiently relies on the capacity to account for complementarity between individual classifiers in order to increase the ensemble's generalization performance as much as feasible. The diversity property is frequently used to define this ability. Although there is no universally accepted definition for diversity, it is widely acknowledged as one of the most essential properties for improving generalization performance in a group of classifiers. It can be defined as an ensemble's individual classifiers' capacity to agree on good predictions while disagreeing on prediction errors. Ensemble approaches from the Random Forest (RF) family. A fixed number of randomized decision trees are inducted to form an ensemble in a "traditional" RF induction method. There are two major downsides to this type of algorithm:

(i) the number of trees has to be fixed a priori

(ii) Because of the randomization principle, decision tree classifiers lose their interpretability and analytical capabilities. This type of method, in which trees are added to the ensemble in a piecemeal fashion, does not guarantee that all of those trees will work together efficiently in the same committee.

HYPER-PARAMETER TUNING:

HYPER-PARAMETER tuning has a significant impact on the prediction performance of machine learning algorithms. Choosing an appropriate configuration for an ML algorithm's HPs is usually done through trial and error. Finding a decent set of values manually can take a long time, depending on how long the ML system takes to train. As a result, recent HP for ML algorithm research has centered on the development of better HP tuning strategies. The HP process is typically viewed as a Blackbox optimization problem, with the objective function linked to the model's predictive performance caused by the method.

DECISION TREE HYPERPARAMETER TUNING:

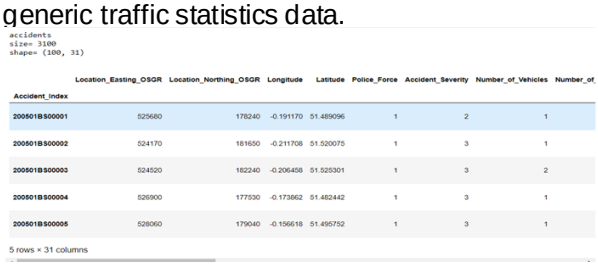
The most researched problem in Machine Learning is supervised categorization. Decision Tree algorithms are a popular choice among the various algorithms used in such tasks since they are both robust and efficient to create. Furthermore, they offer the benefit of developing understandable models with acceptable accuracy levels in a variety of application domains. These algorithms, like most Machine Learning methods, contain several hyperparameters whose values have a direct impact on the performance of the induced models. Because there are so many possible hyper parameter values, some research use optimization approaches to identify a decent set of solutions in order to construct classifiers with good predictive performance. To change Decision Tree algorithm hyper parameters, four distinct tuning strategies were investigated. Overall, experiments with varied datasets were used to investigate the tuning effect on induced

models. The experimental results reveal that, despite a low average improvement across all datasets, the improvement is statistically significant in the majority of situations. Because supervised classification is one of the most common Machine Learning (ML) tasks, there is a wide range of classification algorithms to choose from. Decision Tree (DT) induction algorithms have been widely employed among them. DTs are classifiers that are represented by rules in the form of a tree. They are commonly utilized owing to their intelligible character, which is similar to human thinking. According to several writers, DTs are among the most commonly employed data mining algorithms by researchers and practitioners, reinforcing their prominence in the ML field. DT induction techniques have a number of advantages over many other machines learning algorithms, including noise resistance (missing data, imbalanced classes), minimal computational cost, and the ability to handle redundant attributes. Quinlan's method and Classification and Regression Tree are two well-known DT induction algorithms in the literature (CART). The hyper-parameters (HPs) of the ML algorithm have a direct impact on the predictive performance of the models created by them. As a result, a fair selection of these values has been studied in ML for many years.

EXPERIMENTAL ANALYSIS:

1)DATA IMPORTING:-

To analyse the data, we need to import three files. This information is divided into three files: accidents, casualties, and vehicles. However, we have one more file that contains generic data on traffic counts from 2000 to 2015. For the machine learning aspect, we can use generic traffic statistics data.



accidents
size= 3100
shape= (100, 31)

Accident_Index	Location_Easting_OSGR	Location_Northing_OSGR	Longitude	Latitude	Police_Force	Accident_Severity	Number_of_Vehicles	Number_of
200501BS00001	525680	178240	-0.191170	51.489096	1	2	1	
200501BS00002	524170	181650	-0.211708	51.520075	1	3	1	
200501BS00003	524520	182240	-0.206458	51.525301	1	3	2	
200501BS00004	526900	177530	-0.173862	51.482442	1	3	1	
200501BS00005	528060	179040	-0.156618	51.495752	1	3	1	

5 rows × 31 columns

Fig 1.1 Importing

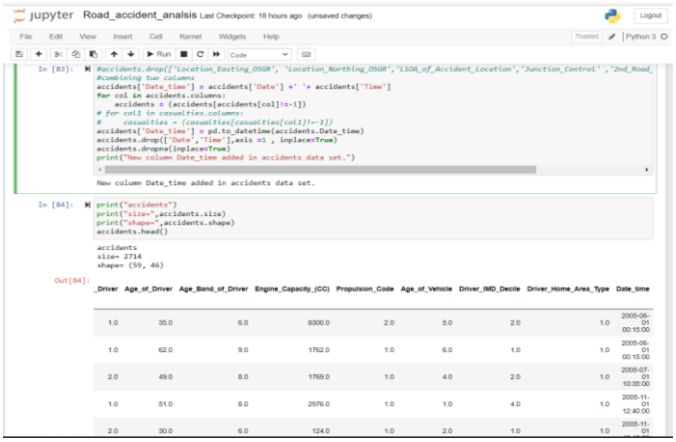
2) PRE-PROCESSING OF DATA:

Data Cleaning :-

Here we identify noisy, irrelevant data. We also understand through visualization which factors are more important.

Identifying Missing Values :-

There are two types of missing values in this dataset: '-1' and 'Nan'. We'll look into each column that has a total of missing values. Because the dataset is large enough to do analysis, we will not be imputing any mean or median values.



jupyter Road_accident_analysis Last Checkpoint: 18 hours ago (unsaved changes)

In [83]:
```
#accidents.drop(['Location_Easting_OSGR', 'Location_Northing_OSGR','LSOA_of_Accident_Location','Junction_Control' ,'2nd_Road_
#combining two columns
accidents['Date_time'] = accidents['Date'] +' '+ accidents['Time']
for col in accidents.columns:
    accidents = [accidents[accidents[col]!=-1]
# for col1 in casualties.columns:
#     casualties = (casualties[casualties[col1]!=-1])
accidents['Date_time'] = pd.to_datetime(accidents.Date_time)
accidents.drop(['Date','Time'],axis =1 , inplace=True)
accidents.dropna(inplace=True)
print("New column Date_time added in accidents data set.")
```

New column Date_time added in accidents data set.

In [84]:
```
print("accidents")
print("size=",accidents.size)
print("shape=",accidents.shape)
accidents.head()
```

accidents
size= 2714
shape= (59, 46)

Out[84]:

_Driver	Age_of_Driver	Age_Band_of_Driver	Engine_Capacity_(CC)	Propulsion_Code	Age_of_Vehicle	Driver_IMD_Decile	Driver_Home_Area_Type	Date_time
1.0	35.0	6.0	6300.0	2.0	5.0	2.0	1.0	2005-06-01 00:15:00
1.0	62.0	9.0	1762.0	1.0	6.0	1.0	1.0	2005-06-01 00:15:00
2.0	49.0	8.0	1769.0	1.0	4.0	2.0	1.0	2005-07-01 10:35:00
1.0	51.0	8.0	2976.0	1.0	1.0	4.0	1.0	2005-11-01 12:40:00
2.0	30.0	5.0	124.0	1.0	2.0	1.0	1.0	2005-11-01

Fig 1.2 joining

3.DATA VISUALIZATION:

The first thing we may do is look at accident times to gain insight and the ages of some of the drivers involved in the collision.

-We can determine the number of accidents on different days of the week.

-We can figure out the number of accidents by looking at the hours of the day.

-We can learn more about the incidents by looking at the age of the driver.

Accidents Occurring on a Specific Weekday

The number of accidents on specific days of the week can be determined. From 2005 to 2015, we can observe that Saturday has the largest number of accidents in this dataset. We must keep in mind that the number of accidents may vary depending on the quantity of traffic on any given day.

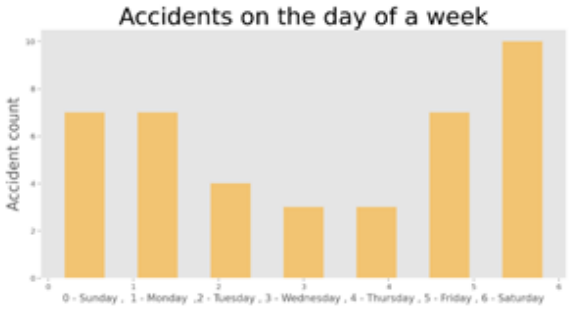


Fig 1.3 Accidents

Time of Accident:-

We discovered that the majority of the accidents occurred after noon. We can infer that the most traffic is flowing during this time of day, as people are leaving work.

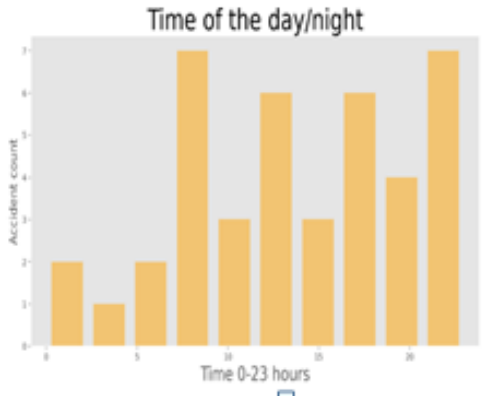


Fig 1.4 Time

Fig 1.6 Correlation of Accident

Age Band of Casualties :-

The age bands in this dataset are divided into 11 distinct codes. We'll make the labels and give them to the plot as x ticks to get a sense of how the bins are represented.

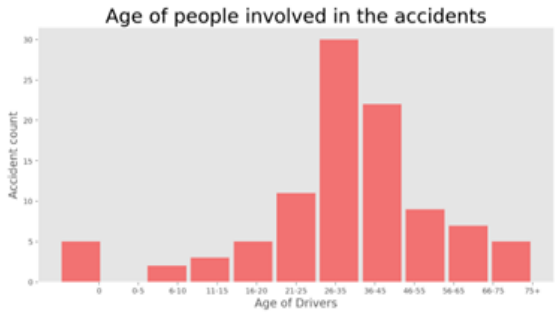


Fig 1.5 Age

Co-relation between variables :-

Because our dataset is made up of numeric values. We can determine if there is a correlation between columns. As can be seen, there aren't many strong relationships between any of the variables. Only one major positive association exists between speed limit and urban or rural area.

Plotting accidents Location on Google Maps :-

Classifying locations based on severity:

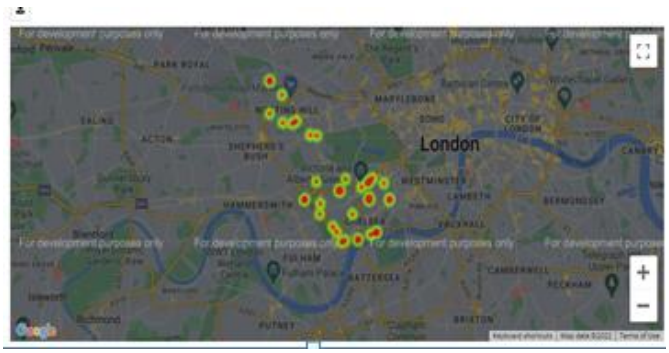


Fig 1.7 Heatmap(ROADMAP)

Fig 1.8 Heatmap(HYBRID)

4.NORMALIZE THE DATA:

We will standardize a few columns so that our machine learning algorithms are not severely impacted. The driver's age ranges from 18 to 88 in the dataset, which we can normalize.

Also, vehicle age ranges from 0 to 100, which can bias the efficiency of your machine learning model, so we'll normalize this prediction as well.

Before Normalization:-

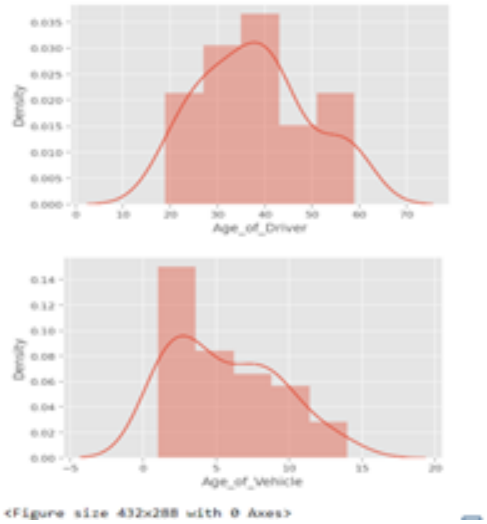


Fig 1.9 Representation

After Normalization:-

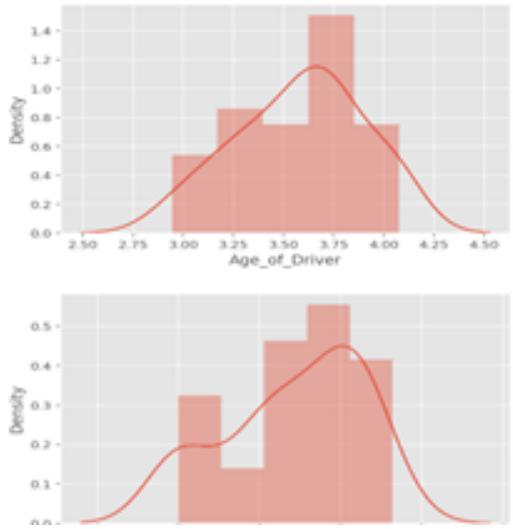


Fig 2.0 Normalization

5.MACHINE LEARNING:

We'll look at several columns to see if we can forecast the severity of the disaster. We can provide some recommendations to law enforcement for looking into this and being prepared for the future once we can estimate the severity of the event. The packages listed below are being imported.

6.SPLITTING THE DATA INTO TRAINING AND TEST DATA:

X is the input data and Y is the class label. 45% of the data is for testing and 50% for training.

```
In [164]: # Split the data into a training and test set.
          X_train, X_test, y_train, y_test = train_test_split(accident_ml.values,accidents['Accident_Severity'].
                                                              values,test_size=0.45, random_state=50)
```

Fig 2.1 Splitting Data

7.ALGORITHMS AND TECHNIQUES :

Algorithms implemented with accuracy and confusion matrix:-

LOGISTIC REGRESSION:-

```
print("Accuracy", round(accuracy_score(y_pred, y_test)*100,2))
Accuracy 78.95
print(sk_report)
              precision    recall  f1-score   support

           2   0.000000  0.000000  0.000000         1
           3   0.937500  0.833333  0.882353        18

    accuracy                       0.789474        19
   macro avg   0.468750  0.416667  0.441176        19
weighted avg   0.888158  0.789474  0.835913        19
```

Fig 2.2 Accuracy: Logistic regression

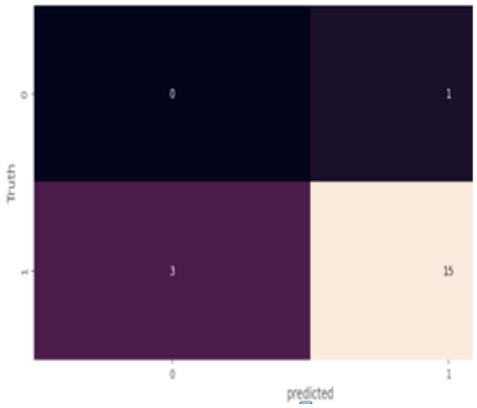


Fig 2.3 Confusion matrix: Logistic Regression

DECISION TREE:-

```
In [132]: y_pred = decision_tree.predict(X_test)
          y_pred = lr.predict(X_test)
          dtree_y_pred=y_pred
          y_pred
Out[132]: array([3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3],
                dtype=int64)

In [91]: print("Accuracy", acc_decision_tree1)
         Accuracy 78.95

In [92]: sk_report = classification_report(digits=6,y_true=y_test, y_pred=Y_pred)

In [93]: print(sk_report)
                       precision    recall  f1-score   support

                    2   0.000000  0.000000  0.000000         1
                    3   0.947368  1.000000  0.972973        18

             accuracy                       0.947368        19
            macro avg   0.473684  0.500000  0.486486        19
         weighted avg   0.897507  0.947368  0.921764        19

In [94]: ### Confusion Matrix
         pd.crosstab(y_test, Y_pred, rownames=['Actual'], colnames=['Predicted'], margins=True)
Out[94]:
```

Predicted	3	All
Actual		
2	1	1
3	18	18
All	19	19

Fig 2.4 Accuracy: Decision Tree

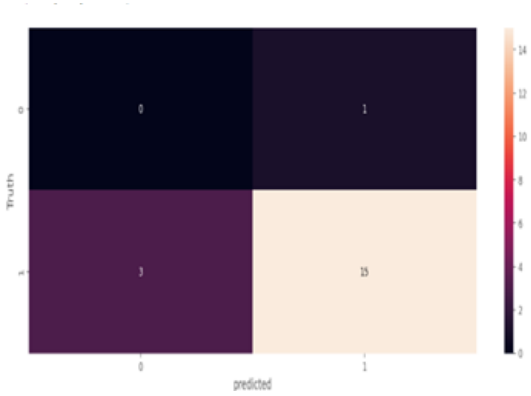


Fig 2.5 Confusion matrix: Decision Tree

RANDOM FOREST:-

```
In [126]: y_pred = random_forest.predict(X_test)
          RdmF_y_pred=y_pred
          y_pred
Out[126]: array([3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3],
                dtype=int64)

In [76]: print("Accuracy" , acc_random_forest1)
         Accuracy 94.74

In [77]: print(sk_report)
                       precision    recall  f1-score   support

                    2   0.000000  0.000000  0.000000         1
                    3   0.947368  1.000000  0.972973        18

             accuracy                       0.947368        19
            macro avg   0.473684  0.500000  0.486486        19
         weighted avg   0.897507  0.947368  0.921764        19

In [78]: pd.crosstab(y_test, Y_pred, rownames=['Actual'], colnames=['Predicted'], margins=True)
Out[78]:
```

Predicted	3	All
Actual		
2	1	1
3	18	18
All	19	19

Fig 2.6 Accuracy: Random Forest

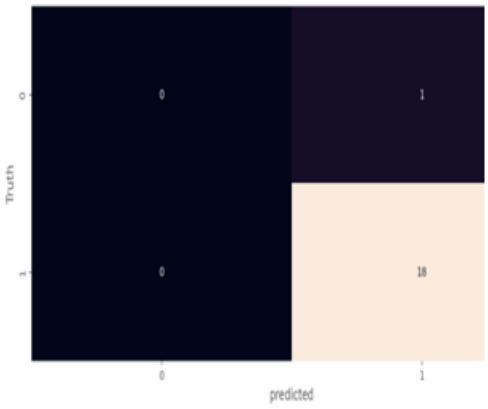


Fig 2.7 Confusion matrix: Random Forest

HYPERPARAMETERS TUNING FOR LOGISTIC REGRESSION:-

```
In [98]: y_pred = lr.predict(X_test)
         logHyp_y_pred=y_pred
         y_pred
Out[98]: array([3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3],
               dtype=int64)

In [99]: print("Accuracy", round(accuracy_score(y_pred, y_test)*100,2))
         Accuracy 94.74

In [100]: sk_report = classification_report(digits=6,y_true=y_test, y_pred=y_pred)

In [101]: print(sk_report)
                       precision    recall  f1-score   support

                    2   0.000000  0.000000  0.000000         1
                    3   0.947368  1.000000  0.972973        18

             accuracy                       0.947368        19
            macro avg   0.473684  0.500000  0.486486        19
         weighted avg   0.897507  0.947368  0.921764        19

In [102]: pd.crosstab(y_test, y_pred, rownames=['Actual'], colnames=['Predicted'], margins=True)
Out[102]:
```

Predicted	3	All
Actual		
2	1	1
3	18	18
All	19	19

Fig 2.8 Accuracy: hyperparameters tuning logistic regression

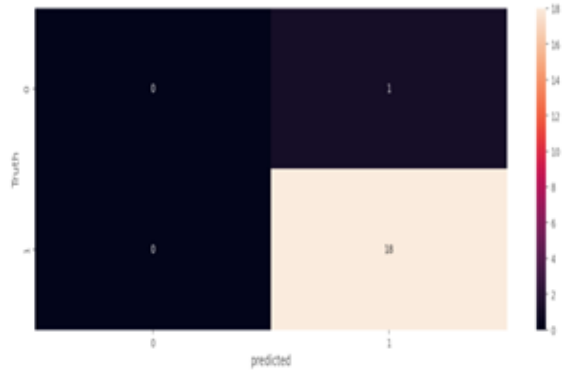


Fig 2.9 Confusion matrix: hyperparameters tuning logistic Regression

HYPERPARAMETERS TUNING FOR DECISION TREE:-

```
In [153]: y_pred = decision_tree.predict(X_test)
          dtreeHyp_y_pred=y_pred
          y_pred
Out[153]: array([3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3],
                dtype=int64)

In [154]: acc_decision_tree1 = round(decision_tree.score(X_test, y_test) * 100, 2)

In [155]: print("Accuracy", acc_decision_tree1)
          Accuracy 94.74

In [156]: sk_report = classification_report(digits=6,y_true=y_test, y_pred=y_pred)

In [157]: print(sk_report)
                        precision    recall  f1-score   support

                     2   0.000000  0.000000  0.000000         1
                     3   0.947368  1.000000  0.972973        18

              accuracy                       0.947368        19
             macro avg   0.473684  0.500000  0.486486        19
          weighted avg   0.897507  0.947368  0.921764        19

In [158]: # Confusion Matrix
          pd.crosstab(y_test, Y_pred, rownames=['Actual'], colnames=['Predicted'], margins=True)
Out[158]:
```

Predicted	3	All
Actual		
2	1	1
3	18	18
All	19	19

Fig 3.0 Accuracy: hyperparameters tuning Decision Tree

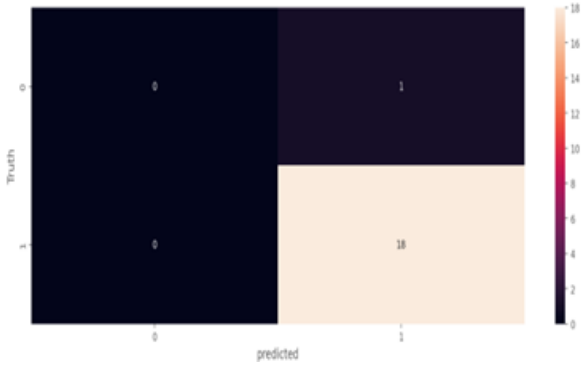


Fig 3.1 Confusion matrix: hyperparameters tuning Decision Tree

HYPERPARAMETERS TUNING FOR RANDOM FOREST:-

```
In [118]: y_pred = grid_search.predict(X_test)
          RdmHyp_y_pred=y_pred
          y_pred
Out[118]: array([3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3],
                dtype=int64)

In [119]: acc = round(grid_search.score(X_test, y_test) * 100, 2)
          acc
Out[119]: 94.74

In [120]: sk_report = classification_report(digits=6,y_true=y_test, y_pred=y_pred)

In [121]: print(sk_report)
                        precision    recall  f1-score   support

                     2   0.000000  0.000000  0.000000         1
                     3   0.947368  1.000000  0.972973        18

              accuracy                       0.947368        19
             macro avg   0.473684  0.500000  0.486486        19
          weighted avg   0.897507  0.947368  0.921764        19

In [122]: pd.crosstab(y_test, Y_pred, rownames=['Actual'], colnames=['Predicted'], margins=Tru
Out[122]:
```

Predicted	3	All
Actual		
2	1	1
3	18	18
All	19	19

Fig 3.2 Accuracy: hyperparameters tuning Random Forest

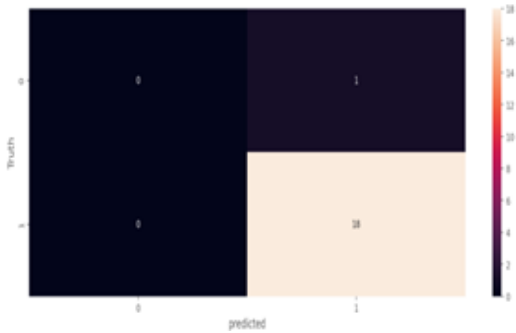


Fig 3.4 Confusion matrix: hyperparameters tuning Random Forest

	Rand_Forest	Log_Reg	D_tree	Log_Reg_hyp	D_tree_hyp	Rand_Forest_hyp
0	3	2	2	3	3	3
1	3	3	3	3	3	3
2	3	3	3	3	3	3
3	3	2	2	3	3	3
4	3	3	3	3	3	3
5	3	3	3	3	3	3
6	3	3	3	3	3	3
7	3	3	3	3	3	3
8	3	3	3	3	3	3
9	3	3	3	3	3	3
10	3	3	3	3	3	3
11	3	3	3	3	3	3
12	3	3	3	3	3	3
13	3	3	3	3	3	3
14	3	3	3	3	3	3
15	3	2	2	3	3	3
16	3	3	3	3	3	3
17	3	3	3	3	3	3
18	3	3	3	3	3	3

```
In [ ]: print("\n\nAccuracy:","  Random_Forest\t","Logistic_reg\t ","Decision_Tree
        print("Accuracy:",acc_rmF,acc_LgR,acc_DT,acc_H_LgR,acc_H_DT,acc_H_RmF)

Accuracy:    Random_Forest      Logistic_reg     Decision_Tree     Hyp_lo
Accuracy: 0.9473684210526315 0.7894736842105263 0.7894736842105263 0.94736
```

Fig 3.5 In Fig 3.5 we compare al the outputs and the accuracy of the different learning tecniques used for finding the severity of the Road accident analysis.

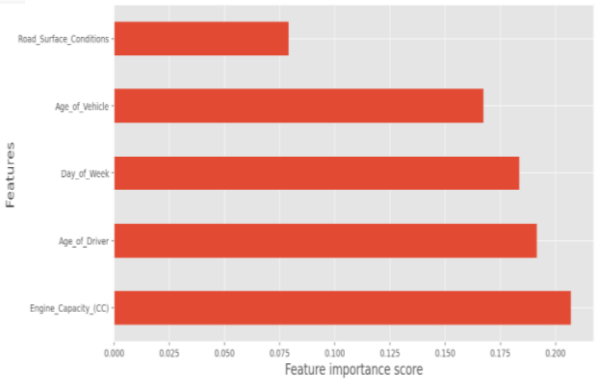


Fig 3.6 Feature importance

CONCLUSIONS:-

Road accident losses are unacceptably high, both for society and for a developing country like ours. As a result, it has become a necessity to use a sophisticated system to control and organize traffic in order to reduce the number of road accidents in our country. Traffic accidents can be avoided by taking modest precautions based on the predictions or warnings of a sophisticated system. Furthermore, there is now a critical requirement for our country to address the problem in which so many people are killed in traffic accidents every day, and this number is increasing day by day. The use of machine learning is a practical and effective way to make an accurate judgement based on past experience in order to manage the current situation, and the results of the analysis can be recommended to traffic authorities in order to reduce the number of accidents. Because of their established and higher accuracy in predicting traffic accident severity, we can apply the presented methodologies to deploy machine learning here. Furthermore, in order to make it more feasible, we will try to create a recommender system utilizing these methodologies, which will be able to predict traffic accidents and alert road users. In the future, we hope to develop a mobile application using this technology that will deliver an accurate prediction to the user while also being very useful and beneficial.

This project's purpose is to apply Machine Learning classification techniques to predict the severity of an accident at a certain location. Thanks to machine learning, we can now analyze large amounts of data and produce solutions that are more accurate than those supplied by humans. We created a model that is 17% more accurate than the traditional method .

REFERENCES:-

[1] Chong, Miao, Ajith Abraham, and Marcin Paprzycki. "Traffic accident analysis using machine learning paradigms." *Informatica* 29.1 (2005).

[2] Chong, M., Abraham, A., & Paprzycki, M. (2005). Traffic accident analysis using machine learning paradigms. *Informatica*, *29*(1).

[3] Sridevi, N., M. V. Keerthana, Monisha V. Pal, T. R. Nikshitha, and P. Jyothi. "Road accident analysis using machine learning." *International Journal of Research in Engineering, Science and Management* 3, no. 5 (2020): 859-861.

[4] Labib, Md Farhan, et al. "Road accident analysis and prediction of accident severity by using machine learning in Bangladesh." *2019 7th International Conference on Smart Computing & Communications (ICSCC)*. IEEE, 2019.

[5] Nandurge, Priyanka A., and Nagaraj V. Dharwadkar. "Analyzing road accident data using machine learning paradigms." In *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, pp. 604-610. IEEE, 2017.

[6] Karimnezhad, A. and Moradi, F., 2017. Road accident data analysis using Bayesian networks. *Transportation letters*, *9*(1), pp.12-19.

[7] Patil, Jayesh, Mandar Prabhu, Dhaval Walavalkar, and Vivian Brian Lobo. "Road accident analysis using machine learning." In *2020 IEEE Pune Section International Conference (PuneCon)*, pp. 108-112. IEEE, 2020.

[8] Vasavi, S. (2018). Extracting hidden patterns within road accident data using machine learning techniques. In *Information and Communication Technology* (pp. 13-22). Springer, Singapore.

[9] Ghandour, Ali J., Huda Hammoud, and Samar Al-Hajj. "Analyzing factors associated with fatal road crashes: a machine learning approach." *International journal of environmental research and public health* 17.11 (2020): 4111.

[10] Bülbül, Halil İbrahim, Tarık Kaya, and Yusuf Tulgar. "Analysis for status of the road accident occurance and determination of the risk of accident by machine learning in istanbul." *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2016.

[11] AlMamlook, R. E., Kwayu, K. M., Alkasisbeh, M. R., & Frefer, A. A. (2019, April). Comparison of machine learning algorithms for predicting traffic accident severity. In *2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)* (pp. 272-276). IEEE.

[12] Malik, S., El Sayed, H., Khan, M. A., & Khan, M. J. (2021, December). Road Accident Severity Prediction—A Comparative Analysis of Machine Learning Algorithms. In *2021 IEEE*

Global Conference on Artificial Intelligence and Internet of Things (GCAIoT) (pp. 69-74). IEEE.

[13] Karimnezhad, Ali, and Fahimeh Moradi. "Road accident data analysis using Bayesian networks." *Transportation letters* 9, no. 1 (2017): 12-19.

[14] Shahzad, Monib. "Review of road accident analysis using GIS technique." *International journal of injury control and safety promotion* 27, no. 4 (2020): 472-481.

[15] Zhang, Xue-Fei, and Lisa Fan. "A decision tree approach for traffic accident analysis of Saskatchewan highways." *2013 26th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*. IEEE, 2013.