Analysis and Prediction of Road Accidents *By Graphical Visualization using Machine Learning*

G. Mani Kandan[1], Sarilya Jaiswal[2], Rahul Mishra[3] & Mrs. Steffina Muthukumar[4]
[1,2,3] Student, [4]Assistant Professor
Computer Science and Engineering,
SRM Institute of Science and Technology, Chennai, India

Received: January 31, 2019 Accepted: March 16, 2019

ABSTRACT: *According to national statistics, thousands of people have been killed and millions injured in road accidents in the United Kingdom from the year 2014 to 2016. There may be a variety of reasons possible that are responsible for these crashes, such as, internal reasons like driver's fault. However, the statistics showed that the factors responsible for these accidents are mostly external, which include adverse weather conditions, poor road conditions, fog, rainfall, partial visibility etc. This paper attempts to create a model to analyse all these factors (spatial and temporal) that are mainly responsible for the cause of an accident and to determine likely road crash conditions. Results from this paper may help the civic authorities to take actions on crash prone weather and traffic conditions.*

I. Introduction

Accidents occur due to various kinds of factors which can be broadly classified into two categories: internal and external factors. These factors determine the reason behind the occurrence of the accident. Internal factors include driver's fault, driver under influence, distribution, etc. External factors include the environment in which the car was being driven, that is, weather conditions, poor road conditions, fog, rainfall, partial visibility etc. To develop a system that analyses all these factors and predicts these occurrences, machine learning concepts have been used.

Machine Learning or ML is a study of algorithms and models that a computer uses to perform specific tasks without further need of instructions, making use of patterns and draws inferences. The algorithms used build mathematical model of given data, also known as 'training data' or 'training sets', in order to make predictions or decisions without being specifically programmed for it.

Supervised Learning algorithms create a mathematical model of a training set that contains one or more inputs and a desired output. These algorithms include classification and regression. Classification algorithms are used when the outputs are restricted to limited values, and regression algorithms are used when the outputs have any numerical value within a range.

Unsupervised Learning algorithms use a data set that has only inputs and figure out a structure from such data, like grouping or clustering. These learn from test data and identify commonalities and give results based on the existence of such commodities in the piece of data.

Semi-supervised learning algorithms make use of data sets in which some of the training examples are missing the desired outputs. Here, the mathematical model represents each example as an array or a vector, and training data by a matrix. By iterative optimization of an objective function, the algorithms learn outputs that can be associated with new inputs.

Reinforcement Learning is a type of machine learning that works upon maximizing the notion of cumulative reward of the actions of software agents in an environment. This field is studied in many disciplines such as game theory, operations theory, information theory, swarm intelligence, statistics, genetic algorithms, etc., hence has a lot of generality.

II. Proposed System

The proposed system is a middleware which uses techniques that involves data slicing, data analysis and data pre-processing for secured and optimized results. Data pre-processing is an important and mandatory step for any machine learning model because it involves steps like feature scaling to get exact values. As machine learning models deals with values with close proximity, splitting the dataset in training is necessary as the dataset contains huge amount of features within which various unwanted features are also present that are not required. After pre-processing the important features are extracted for the problem and

based on the dataset they are analysed graphically. Further the model is trained with three algorithms one by one using the important features and at last accuracy is compared for the three algorithms based on true and predicted results.

III. System Architecture

The proposed system is a flow diagram architecture with multiple constraints in it. The architecture involves dataset, Middleware System and the Accuracy prediction. The mode of communication takes place between the dataset and, prediction and analysis models.

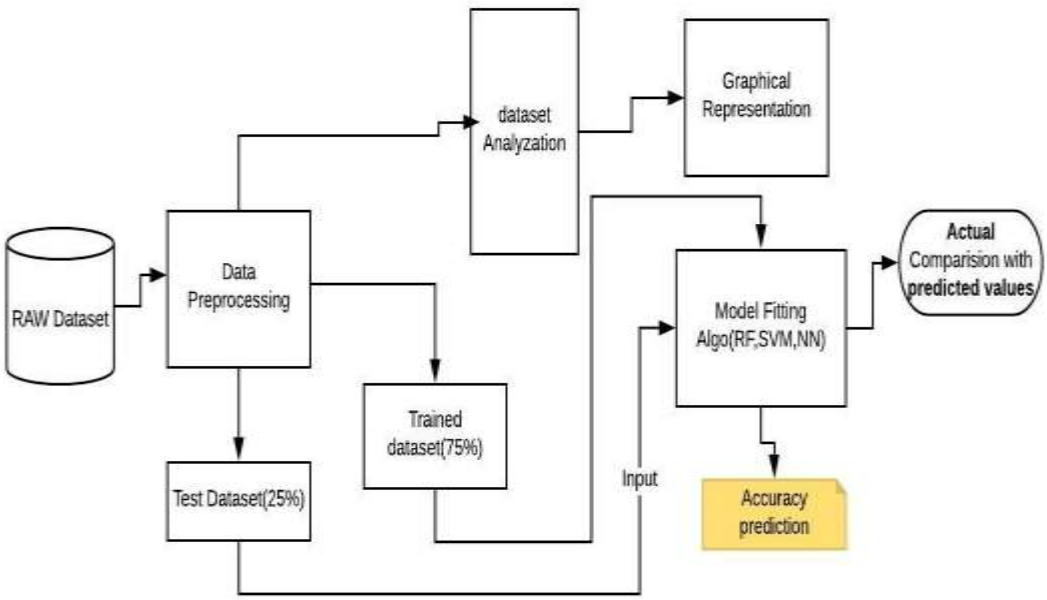

Fig 1: System Architecture

3.1 Dataset

The dataset contains very large amount of features and values arranged in rows and columns. Some of the features are not needed for our models only important features are extracted after analyzing graphically.

3.2 Middleware

The middleware system consists of data pre-processing which is actually mandatory for both analysis and prediction as it takes care of missing data, splitting our data set into training and test part, takes care of categorical data and applies feature scaling which is needed for accurate results.

3.2.1 Data Pre-Processing

As machine learning always works upon numerical values hence, categorical values are being encoded into dummy variables.

3.2.1.1 Feature Scaling

Feature scaling is used to reduce the data points values, for any particular feature if they are vast in range. Two methods have been employed, namely:
Standardization

$$x' = \frac{x - \bar{x}}{\sigma}$$

Where $\bar{x}$ =average(x) is the mean of that feature factor, x is the original feature factor, σ is its standard deviation.

3.3 Data Analysis

After pre-processing, important features are extracted from the dataset and visualized graphically based on the true results from the past data. This will help in analyzing the cause of the problem based on various scenarios and according to our dataset and finally use that scenario in our prediction model.

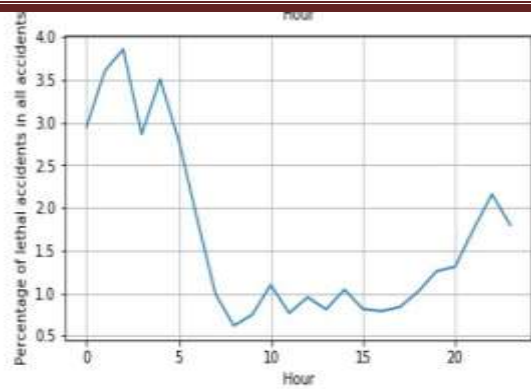

Fig 2: Percentage of lethal accidents vs hour

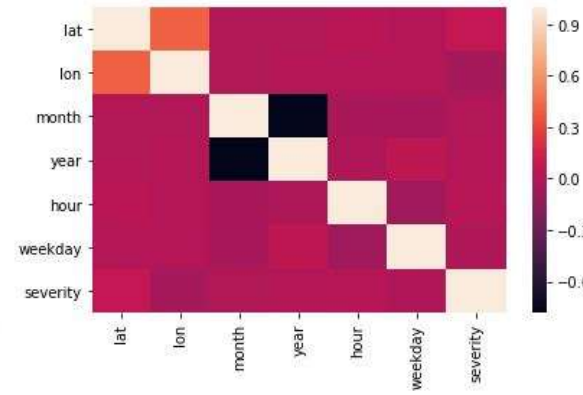

Fig 3: Correlation heatmap

The most dangerous hour to drive, when most fatal accidents happened in all accidents, is 2 o'clock

3.4 Prediction and Accuracy Calculation

The prediction model is used in the analysis of the important features. This model comprises of three algorithms which are random forest, artificial neural networks, support vector machines. These algorithms are helpful in comparing the results with true and predicted ones. True results based on our dataset and predicted results from the algorithms will help in determining the accuracy of the prediction.

Random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing multiple decision trees during training and outputting the class that is a mean prediction (regression) of the individual trees. The system uses a scikit learn library, through which the random forest class is called. The pre-processed dataset is given as an input and the output is received in the form of accuracy prediction. For the visualization in 2-D, kernel PCA is being used in dimensionality reduction of the dataset.

A neural network is an artificial replication of the system of biological neurons, to form a network composed of artificial neurons and nodes, used for solving artificial intelligence problems. These may be used for predictive modeling, adaptive control and in applications where they can be trained using a dataset. They can self-learn resulting into a development of experience within the network which can be used to derive solutions of seemingly complex problems. The NN in the system uses keras library as tenser flow backend which helps in accuracy prediction.

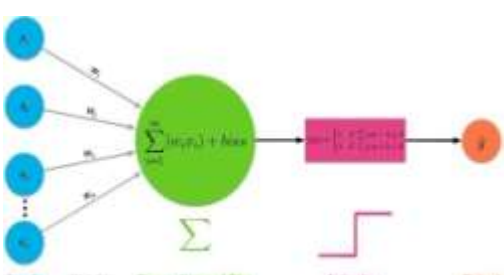

Fig 4: Artificial neural network structure

Fig 5: Rectifier function

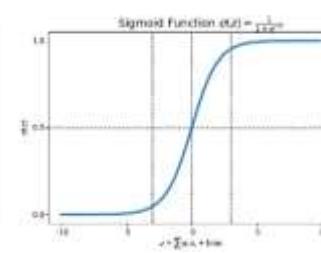

Fig 6: Sigmoid function

SVMs are supervised learning models with associated learning algorithms that analyze data used for analysis of classification and regression. It builds a model that assigns new examples to one category or another. A SVM model represents the training set as points in space, mapped so that the examples of different kinds are separated by a clear gap. As in this case non-linear dataset is being used, hence kernel-SVMs are being employed for accuracy prediction. For the visualization, kernel PCA is being used to sharpen the dataset features.

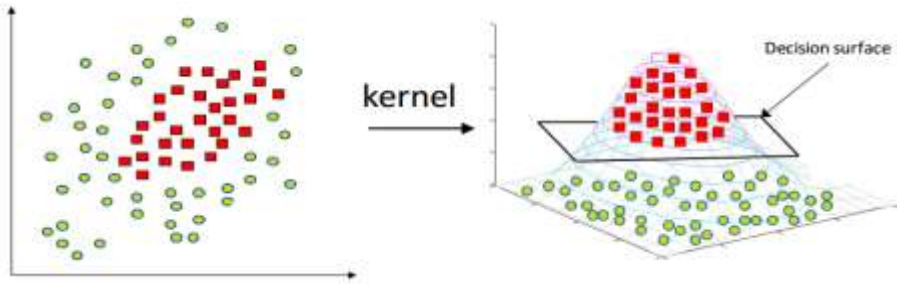

Fig 7: Kernal SVM

XG boost is an extreme Gradient Boosting technique in applied machine learning which encapsulate the high performance and better speed over the other classification and regression models. Gradient boosting is a method where new models are createdthatpredict the errors of earlier models and then combined together to make a final prediction. It helps in both classification and regression models in order to apply ensemble learning to optimize the performance and accuracy.
Accuracy is calculated for each algorithm and finally it is compared that which model is much accurate for our problem definition. Our model is trained in our training set and the results are tested in our test set which tells the accuracy of our model.
Accuracy validation is validated by the k-fold validation technique where accuracy is calculated k-times in training set and mean value are compared against the test set accuracy.

IV. Conclusion AND FUTURE Works

In this paper we have proposed a model which will help the data scientists that how a problem definition can be analyzed then extracting the important features that will actually help in any model and as of data pre-processing is the mandatory step for each and every machine learning model because it will help us to analyze the problem definition as our dataset contains huge amount of data which will have the important features and also the unwanted features that are not required for our problem. With the help of such analysis based on true results anyone can make a model that will be helpful for various business problems descriptions and other problem definitions etc. Future work may be after studying these models anyone can relate the problem definition that what are the features that are actually causing a problem(here accident) and based on that he/she can take precautionary measures to avoid the problem in future.

References

1. Sachin Kumar, DurgaToshniwal, "A data mining approach tocharacterize road accident locations", J. Mod. Transport.(2016)24(1):62–72.
2. S. Shanthi and Dr. R. GeethaRamani, "Gender SpecificClassification of Road Accident Patterns through Data Mining Techniques", IEEE-International Conference on Advances inEngineering, Science and Management (ICAESM -2012) March 30, 31, 2012.
3. Tessa K. Anderson, "Kernel density estimation and K-meansclustering to profile road accident hotspots", Accident Analysisand Prevention 41 (2009) 359–364.
4. Miao Chong, Ajith Abraham and Marcin Paprzycki, "TrafficAccident Data Mining Using Machine Learning Paradigms", Oklahoma State University, USA.
5. G. Andrienko, N. Andrienko, P. Bak, D. Keim, S. Kisilevich, and S. Wrobel, "A conceptual framework and taxonomy of techniques for analyzing movement," *J. Vis. Lang. Comput.*, vol. 22, no. 3, pp. 213–232, Jun. 2011