

ROAD ACCIDENT ANALYSIS

S. Kishore Babu¹, B. Geethika Sindhu², P. Priya Darsini³, G. Srilekha⁴

Associate Professor, B. Tech Students

* Department of Information Technology*

** ANDHRA LOYOLA INSTITUTE OF ENGINEERING & TECHNOLOGY **

Abstract-In India, most people are willing to buy two-wheelers as they are cheaper than any other vehicle. In the same way, Engineers and researchers in the automobile industry have tried to design and build safer automobiles, but traffic accidents are unavoidable. Patterns involved in dangerous crashes could be detected if we develop accurate prediction models capable of automatic classification of type of injury severity of various traffic accidents. These behavioral and roadway accident patterns can be useful to develop traffic safety control policies. We believe that to obtain the greatest possible accident reduction effects with limited budgetary resources, it is important that measures be based on scientific and objective surveys of the causes of accidents and severity of injuries. This work summarizes the performance of various machine learning paradigms applied to modeling the severity of injury that occurred during traffic accidents.

Keywords:- Machine learning, RF, DTC, S VM, Naive Bayes, KNN, validation, heat map correlation.

I. INTRODUCTION

The costs of fatalities and injuries due to traffic accidents have a great impact on the society. In recent years, researchers have

II. AIM & OBJECTIVE

The behavioral and roadway accident patterns can be useful to develop traffic safety control policies. We believe that to obtain the greatest possible accident reduction effects with limited budgetary resources, it is important that measures be based on scientific and objective surveys of the causes of accidents and severity of injuries.

The Our objective to detect and classify the data, whether its severity or not. It can be done through comparative analysis of group of classification techniques.

III. EXISTING SYSTEM & ITS LIMITATIONS

Features of road accident data signals are distinctive and collection of the signals is cost-efficient. From road accident (accident, casualties, vehicles data) and its features for the development of the model. For the implementation of a supervised learning method in MATLAB. In this study, Supervised Machine Learning has been used to classify the road accident. Road accident that was chosen to classify Road are interval, and analyzing them, the model will decide if the subject is stressed or relaxed. As there are exactly two class labels in our data, (i.e Road accident or relaxed) we have chosen different ML models for classification and detection of road accident.

➤ Unable to control the learning process.

➤ Feature Selection, was not included.

IV. PROPOSED SYSTEM

The to develop an application that can be used to predict the risk of accidents. The data sets are collected from various websites such as Kaggle, data towards science etc., The next step is to clean the data and transform it into the desired format by feature selection method. Applying data mining techniques to model traffic accident data records can help to understand the characteristics of drivers' behavior, roadway condition and weather condition that were causally connected with different injury severity. This can help decision makers to formulate better

traffic safety control policies. The behavioral and roadway accident patterns can be useful to develop traffic safety control policies. We believe that to obtain the greatest possible accident reduction effects with limited budgetary resources, it is important that measures be based on scientific and objective surveys of the causes of accidents and severity of injuries.

- It helps to provide an understanding of most effective solutions.
- Essential for monitoring and evaluating safety of road network.
- By using ML techniques, we can easily identify trends and patterns

V. STUDY OF THE SYSTEM

1. Matplotlib: Matplotlib can be used in Python scripts, the Python and I Python shells, the Jupiter notebook, web application servers, and four graphical user interface toolkits. Matplotlib tries to make easy things easy and hard things possible. Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. We can able to generate plots, histograms, power spectra, bar charts, error charts, scatter plots, etc., with just a few lines of code. For examples, see the sample

plots and thumbnail gallery. For simple plotting the pyplot module provides a MATLAB-like interface, particularly when combined with I Python. For the power user, you have full control of line styles, font properties, axes properties, etc., via an object-oriented interface or via a set of functions familiar to MATLAB users.

2.Torch: PyTorch is an open-source machine learning framework based on the Torch library, used for applications such as computer vision and natural language processing, primarily developed by Facebook's AI Research lab. It is free and open-source software released under the Modified BSD license.

3.TensorFlow: TensorFlow offers multiple levels of abstraction so you can choose the right one for your needs. Build and train models by using the high-level Keras API, which makes getting started with TensorFlow and machine learning easy. For more flexibility, eager execution allows for immediate iteration and intuitive debugging. For large ML training tasks, use the Distribution Strategy API for distributed training on different hardware configurations without changing the model definition.

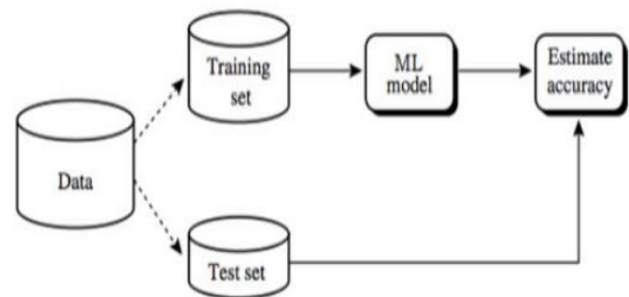
4.Numpy: NumPy is a general-purpose array-processing package. It provides a high performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python. It contains various features including these important ones: • A powerful N-dimensional array object Sophisticated (broadcasting) functions • Tools for integrating C/C++ and Fortran code • Useful linear algebra, Fourier transform, and random number capabilities NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary datatypes can be defined using NumPy which allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

5.Pandas: Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures. Python was majorly used for data munging and preparation. It had very little contribution towards data analysis. Pandas solved this problem. Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data load, prepare, manipulate, model, and analyze. Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc.

VI.SYSTEM DESIGN

System design shows the overall design of system. In this section we discuss in detail the design aspects of the system.

VII. SYSTEM ARCHITECTURE



VIII. METHODOLOGY INVOLVED IN THIS PROJECT

There are 3 modules we are using in this project. They are:

- 1) User Activity
- 2) Training Dataset
- 3) Text Vectorizer

1.User Activity: User login our system or mobile phone or any text devices. our clarify questions and response the admin the every time in same questions but not answer the same answer. The admin replies for the user.

2.Training Dataset: We have to create the captions and objects for every dataset and we have to split the dataset for training, testing and splitting.

3.Text Vectorizer: We have to convert the words from training data to vectorizer It transforms a batch of strings (one example = one string) into either a list of token indices (one example = 1D tensor of integer token indices) or a dense representation (one example = 1D tensor of float values representing data about the example's tokens). This layer is meant to handle natural language inputs. To handle simple string inputs (categorical strings or pre-tokenized strings)

Machine Learning Tools:

There are many different software tools available to build machine learning models and to apply these models to new, unseen data. There are also a large number of well-defined machine learning algorithms available. These tools typically contain libraries implementing some of the most popular machine learning algorithms. They can be categorized as follows: Pre-built application-based solutions. Programming languages which have specialized libraries for machine learning. Using programming languages to develop and implement models is more flexible and gave us better control of the parameters to the algorithms. It also allows us to have a better understanding of the output models

produced. Some of the popular programming languages used in the field of machine learning are:

PYTHON: Python is an extremely popular choice in the field of machine learning and AI development. Its short and simple syntax make it extremely easy to learn and use.



MATLAB: MATLAB is a programming language developed by MathWorks. Created primarily for numerical computing, MATLAB is also extremely popular among machine learning programmers. It is heavily used in statistical analysis and complex systems. MATLAB excels at handling matrices, making it especially useful in image recognition. It is an extremely versatile language which allows matrix manipulations, implementation of algorithms and creation of user interfaces. It is also able to interface with other programming languages like C, Fortran and Java.

SCIKIT-LEARN: SciKit learn is an open source machine learning library built for python. Since its release in 2007, Scikit-learn has become one of the most popular open source machine learning libraries. Scikit-learn (also called sklearn) provides algorithms for many machine learning tasks including classification, regression, dimensionality reduction and clustering. It also provides utilities for extracting features, processing data and evaluating models. It provides in-built code for many of the popular machine learning algorithms. The reasons for this are: We already have some familiarity and exposure to Python, and thus have a smaller learning curve. Both Python and Scikit-learn have excellent documentation and tutorials available online. The number of classic machine learning algorithms that come with Scikit-learn, and the consistent patterns for using the different models i.e., each model can be used with the same basic commands for setting up the data, training the model and using the model for prediction. This makes it easier to try a range of machine learning algorithms on the same data. The machine learning algorithms included with ski-learn have modifiable parameters known as hyper-parameters that effect the performance of the model. These usually have sensible default values, so that we can run them without needing a detailed knowledge or understanding of their semantics. Data Preprocessing Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a method used to resolving such issues. Real world data is generally incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate

data. It can also contain noise such as errors or outliers. The steps involved in data preprocessing are:

1. Data Cleaning: The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data and noise. Missing Data:

a. Missing Data: Remove tuples with missing values: We simply remove those tuples which contain missing values. This is only feasible if we have a large dataset which will not be affected by losing rows.

b. Imputation: It involves filling the missing values manually. It can be done by taking the mean, median or most probable value using the values in the same column. Creation of Synthetic Dataset for Haze Removal and Haze Classification. Noisy Data:

a. Binning: This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.

b. Regression: Here data can be made smooth by fitting it to a regression function. The regression used may be linear or multiple.

c. Clustering: This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

2. Data Transformation: In this step we transform the data into a form suitable for the data mining process.

a. Normalization: It is a method used to scale all the data values to a specified derange. This helps standardize all the values.

b. Feature Selection: It is the process of reducing the size of the input by selecting only a subset of features from all the originals. It helps reduce the degree of cardinality.

c. Discretization: It is used to change continuous values into discrete intervals.

d. Concept Hierarchy: Here we create a hierarchy that allows attributes to get converted into an attribute which is higher in the hierarchy.

3. Data Reduction: While working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we uses data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs. The various steps to data reduction are:

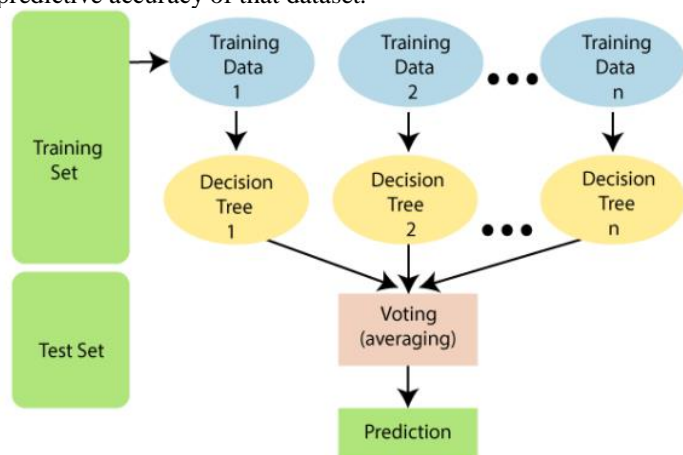
a. Data Cube Aggregation: Aggregation operation is applied to data for the construction of the data cube. Raw data is gathered and expressed in a summary form for statistical analysis.

b. Attribute Subset Selection: The highly relevant attributes should be used, rest all can be discarded. For performing attribute selection, one can use level of significance and p- value of the attribute. The attribute having p value greater than significance level can be discarded. Creation of Synthetic Dataset for Haze Removal and Haze Classification.

c. Dimensionality Reduction: This reduce the size of data by encoding mechanisms. lossy or lossless. If after reconstruction from compressed data, original data can be retrieved, such reduction are called lossless reduction else it is called lossy reduction. The two effective methods of dimensionality reduction

are: Wavelet transforms and PCA (Principal Component Analysis).

RANDOM FOREST ALGORITHM: Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.



SUPPORT VECTOR MACHINES (SVM): A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two-dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.

VIII.SYSTEM REQUIREMENTS

SOFTWARE REQUIREMENTS

- OS: Windows
- Python Version: python 3.7.0
- Language: Python

HARDWARE REQUIREMENTS

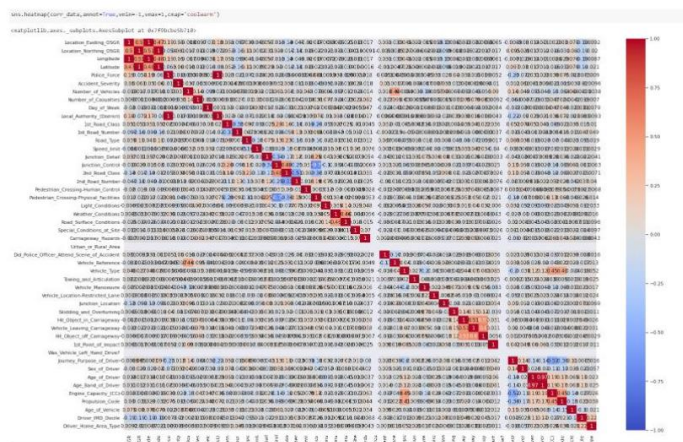
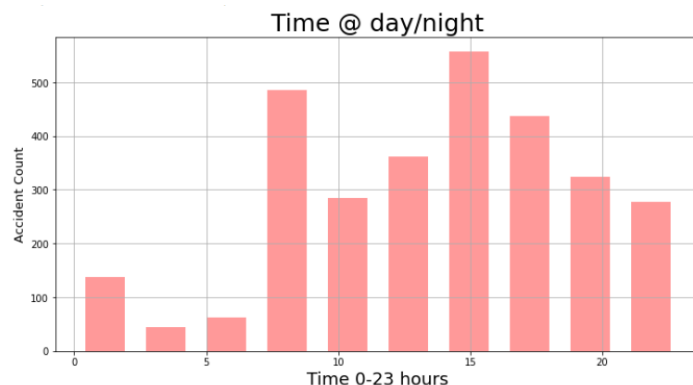
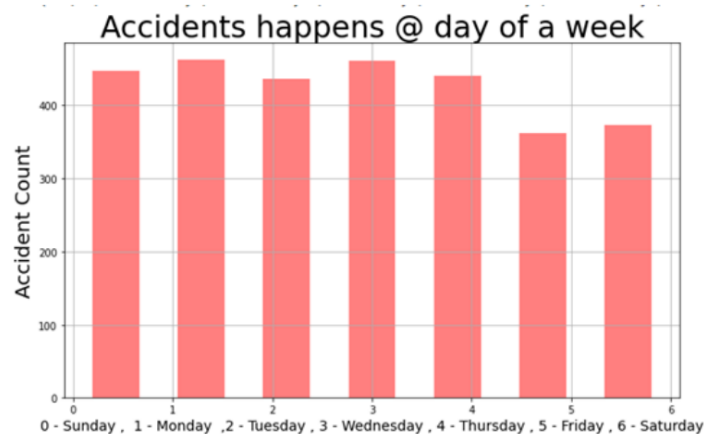
- RAM: 8GB and Higher
- Processor: Intel i7
- Hard Disk: 500GB: Minimum

CONCLUSION

In our work, we used correlation heat map feature selection process to get the important features, and its applied to the classifier. Here it's an comparative analysis of road accident analysis, their classifiers like RF, DTC, KNN, SVM, NB, Logistic Reg etc. When compare with other classifier Support Vector classifier performance is good and its accuracy level is around

(89.24%). In future implement the road accident is real-time data acquisition-based analysis.

FINAL OUTPUT



	model	feature_count	acc	prc	rec	f1
0	rfc_model_1	10	0.888018	0.888018	0.888018	0.888018
1	dtc_model_1	10	0.790594	0.790594	0.790594	0.790594
2	nbc_model_1	10	0.857783	0.857783	0.857783	0.857783
3	svc_model_1	10	0.892497	0.892497	0.892497	0.892497
4	knc_model_1	10	0.883539	0.883539	0.883539	0.883539

REFERENCES

- https://www.researchgate.net/publication/31749569Analysis_of_Datamining_Technique_for_Traffic_Accident_Severity_Problem_A_Review

- https://www.researchgate.net/publication/24359517_Kernel_density_estimation_and_Kmeans_clustering_to_profile_road_accident_hotspots
- https://www.researchgate.net/publication/322314901_Intelligent_Alarm_System_for_Road_Collision
- https://www.sciencedirect.com/science/article/pii/S0001457511000765?casa_token=_1gKMfbkmM8AAAAA:s7dVvFQ2-4h7eje6HyuOSKUVMi4poCTqaRZzRkgymDOWyHE9_INYjjGpjopgbm4LnAoHvaYgRhY

AUTHORS

S. Kishore Babu¹ M.Tech., Associate Professor,
Department, IT.
B.Geethika Sindhu² B.Tech, Andhra Loyola Institute
of Engineering & Technology
P. Priya Darsini³ BTech, Andhra Loyola
Institute of Engineering & Technology
G. Srilekha⁴ B.Tech, Andhra Loyola Institute
of Engineering & Technology