

Road Accident Prediction using Machine Learning Techniques.

Gauri Mane

Department of computer Engineering

Vijay U. Rathod

Department of computer Engineering

Abstract:- The road has been remodelled into a complex building in style and management areas due to the rise within the variety of vehicles. This situation has known the matter of road accidents, contributed to public health and therefore the country's economy, and conducted studies on the answer to the current problem. Huge knowledge integration has been enlarged due to technological advances and knowledge retention at a lower cost. The emergence of the necessity for knowledge retrieval from this huge knowledge scale has found a cornerstone of the info mine. During this study, the allocation of the foremost relevant machine classification strategy for road accident measure by knowledge mining is intended. Due to exponentially increasing number of vehicles on the road, number of accidents occurring on a daily basis is also increasing at an alarming rate. The capacity to predict the number of traffic accidents over a specific period of time is crucial for the transportation department to make informed judgments given the large number of traffic incidents and fatalities these days. In this situation, it will be beneficial to assess the frequency of accidents so that we may utilise this information to further help in the development of strategies to reduce them. Even though the bulk of accidents are characterised by unpredictability, there is a degree of regularity that may be observed when incidents are observed in a specific location over time. Making accurate forecasts about the chances of accidents occurring in a given location and creating accident prediction models can both benefit from this regularity. In this paper, we've examined the connections between traffic accidents, road conditions, and the impact of environmental factors on the possibility of an accident. We have made use of Machine Learning techniques in developing an accident prediction model using classification Model. Traffic accident prediction is useful for several stakeholders, including government public works agencies, contractors, and other automotive industries to better design roads and vehicles based on the obtained estimates.

Keywords: Big data, Machine Learning, Logistic Regression, SVM, Decision Tree, Naïve Bayes, Adaboost Algorithm, Rule Mining

Date of Submission: 14-07-2022

Date of acceptance: 28-07-2022

I. INTRODUCTION

According to the death statistics published by the World Health Organization, the number of traffic accidents that occur annually in the world is alarming. "Traffic accidents kill 1.2 million people and injure 50 million people every year. Approximately 3,300 people were killed and 137,000 people were injured each day. Direct economic losses amounting to \$43 billion, the frequent occurrence of road accidents directly threatens human lives and property safety." [3] Road accidents are a serious cause of concern across the Indian mainland. In 2019 alone, the country claimed more than 151 thousand deaths due to traffic accidents every year, which represents 3 to 5 percent of the country's gross domestic product invested in traffic accidents. Notably, while the Republic of India has approximately one percent of the world's vehicle population, it also accounts for approximately six percent of the world's traffic accidents, with virtually seventy accidents involving young Indians. Traffic studies killed this traffic accident and death. - the quantitative ratio of the laceration may increase. Traffic design and management using advanced systems are available for important needs, traffic risk beliefs and laws, and interventions on top of these assumptions can reduce traffic accidents..Associate in nursing assumption system which will be ready with accessible information and new risks are going to be advantages. Road traffic accidents square measure one amongst the foremost fatal hazards to individuals. Predicting potential traffic accidents will facilitate to avoid them, decrease harm from them, and provides drivers alerts to potential dangers, or improve the emergency management system. a discount in reaction time is also earned if authorities in a section receive advance notice or warning on which parts of the district's roads square measure a lot of doubtless to possess AN accident at numerous times of the day. The work and approach delineated during this paper square measure supported the extraction of information from various sources and making an integrated information, mistreatment AI methodologies to form new models, group action and evaluating totally different AI approaches (machine learning), and assessment of the prophetic power of the models and their validation.

II. LITERATURE REVIEW

No specific approach available for the traffic police to predict which area is accident prone at a specific time. The traditional Back propagation network has defects. It has a 17% lower accuracy than the proposed model. We propose the use of a machine learning technique. Machine learning has the ability to model complex non-linear phenomenon. [1] To predict the traffic accident severity by using convolution neural Network. [1] Traditional way of linear analyses can not reveal the really situation the result of prediction is not satisfactory. Compares traditional BP network with its proposed solution [3] In order to increase the precision of predictive modelling, this research suggests an evolutionary cross validation approach for locating optimal folds in a dataset. This paper presents a study on the Random Forest (RF) family of ensemble methods. [4] This paper investigates how sensitive decision trees are to a hyper-parameter optimization process. Four different tuning techniques were explored. The number of road accidents that occur each year around the world is alarming, according to the World Health Organization's death figures. 1.2 million people die and 50 million people are injured in traffic accidents every year. Approximately 3,300 people were killed and 137,000 people were injured each day. Traffic accidents occur frequently, directly threatening human lives and property safety, with direct economic damages amounting to \$43 billion. Traffic accident prediction is one of the important research contents of traffic safety. The occurrence of traffic accidents is mainly influenced by the geometric characteristics of the road, the traffic flow, the characteristics of the drivers and the road environment [5]. Many studies have been conducted to predict accident rates and analyze the characteristics of traffic accidents, including studies on the identification of danger spots/hot spots.

III. BACKGROUND

The most important background of machine learning algorithms their technique and mathematical formulation are outlined in this section. Analysing the accident data used these algorithms. To boost machine learning algorithm by using hyper-parameter optimization, first we need to find out what key component hyper-parameter are need to hyper tune machine learning model for solving specific problems or datasets.

Machine learning model type can be classified into Supervised learning, Unsupervised learning, and Reinforcement learning algorithm based on labeled datasets, unlabeled datasets, reward system. Supervised learning algorithm classified into Naïve Bayes (NB), Support Vector machine (SVM), K – Nearest neighbor (KNN), Decision Tree (DT), Logistic Regression (LR).

1. Supervised Machine learning algorithm:

In supervised machine learning algorithm independent variable (x as input) and dependent variable (as output) are available, the goal of predictive modeling technique f^* to minimize cost function $f(x, y)$ that model error between estimated output and actual output.

I - SUPPORT VECTOR MACHINE

A support vector machine is a supervised machine learning algorithm is used for solving both classification and regression problem. Support vector machine learning algorithm uses key concept of high dimensionality and low dimensionality for map data points, hyperplane i the model generate the classification boundary to classification data points of datasets.

$$\arg \min_w \left\{ \frac{1}{n} \sum_{i=1}^n \max\{0, 1 - y^i f(x_i)\} + C w^T w \right\}$$

Where,

w – is normalization vector

c – Is penalty parameter or error term(1)

The kernel function $f(x)$ is used to compare similarity between two data points (x_i, y_j) . In the algorithm kernel type hyper-parameter would be tuned.

Different kernel function in the support vector machine is as follows.

1. Linear Kernel
2. RBF Kernel
3. Polynomial Kernel
4. Sigmoid Kernel

In the support vector machine learning algorithm need to tuned other different parameters after tuning of kernel type. In support vector machine learning algorithm different hyperparameter such as epsilon, it indicates the distance error of its loss function.

II – NAÏVE BAYES

Naïve bayes is a supervised machine learning algorithm is used for classification and regression problem. Naïve bayes is based on bayes theorem, objective function of naïve bayes can be denoted by

$$\hat{y} = \underset{y}{\operatorname{arg\,max}} P(y) \prod_{i=1}^n P(x_i|y) \dots(2)$$

Where,

$P(y)$ is the probability of a value y

$P(x_i|y)$ is the posterior probabilities of x_i given the value of y . Four types of naïve bayes classifiers model are:

1. Bernoulli Naïve Bayes
2. Gaussian Naïve Bayes
3. Multinomial Naïve Bayes
4. Complement Naïve Bayes

For Gaussian Naïve Bayes, the likelihood of feature is assumed to follow Gaussian distribution.

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right) \dots(3)$$

Maximum likelihood method is used for calculate the mean value. The accuracy and efficiency of depend on the datasets follows the multinomial naïve bayes distribution.

III – NAÏVE BAYES

Naïve Bayes is the machine learning algorithm based on bayes theorem and used for classification model that optimizes joint likelihood. Assume there are n features X_1, X_2, \dots, X_n and a target variable y , the objective function of naïve bayes can be denotes as,

$$\hat{y} = \underset{y}{\operatorname{arg\,max}} P(y) \prod_{i=1}^n P(x_i|y) \dots(4)$$

Where,

$P(y)$ is the probability of a value of y

$P(x_i|y)$ is the posterior probabilities of x_i

There are different types of Naïve Bayes algorithm. The four main types of naïve bayes algorithm are as follows,

1. Bernoulli Naïve Bayes
2. Multinomial Naïve Bayes
3. Gaussian Naïve Bayes
4. Complement Naïve Bayes

IV. Evaluation Matrix:

1. Accuracy:

it is measured how many true positive and true negative cases is correct. Mathematically it is defined as

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \dots(5)$$

2.Sensitivity or Recall:

1. Recall: Tells us how many of the actual positive cases we were able to predict correctly with our model. Mathematically it is defined as

$$\text{Recall} = \frac{TP}{TP+FN} \dots(6)$$

2. Specificity: Tells us how many times classifier gets true negative correct value, mathematically it is defined as

$$\text{Specificity} = \frac{TN}{TN+FP} \dots(7)$$

3. Precision:

Precision tells us how many of the correctly predicted cases actually turned out to be positive. Mathematically it is defined as,

$$\text{Precision} = \frac{TP}{TP+FP} \dots(8)$$

PROPOSED SYSTEM

In this article, we have created an application that is able to predict the possibility of accidents based on the available traffic accident data. Data pre-processing is performed on this traffic accident data to obtain a dataset. The data preprocessing step includes cleaning to remove null and garbage values and data normalization, followed by feature selection, where only relevant features from the original dataset are selected to be included in the final dataset. The acquired raw traffic accident data is pre-processed to form a dataset to be fed into the model. The model is further trained using the training data and made to predict the possible risk of accidents for the area specified by the user. Based on the obtained statistics, a graphical representation is displayed to the user. A traffic accident prediction model has been developed and implemented, which takes into account various possible causative factors. The number of factors selected for the study is mainly limited to the condition of the road, weather effects and the nature of the cause of the accident. This system refers to the following methods.

- Data Collection
- Data Pre-processing
- Model Selection
- Model Evaluation
- Classification Result (output)

PROPOSED ALGORITHM

The following shows the pseudocode for the proposed accident prediction method

1. Accident data
2. Determine the results of training and testing
3. Data cleaning and preprocessing.
 - a) Fill in the missing values with the mean values related to the numerical values.
 - b) Fill in the missing values with the mode values related to the categorical variables.
 - c) Remote treatment.
4. Use modeling for prediction
 - a) Removal of Cargo Identifier
 - b) Create a target variable (based on the request). In this approach, the target variable is accident rate
 - c) Create a dummy variable for the categorical variable and split the training and test data for validation.
 - d) Implement the model using NB or SVM methods.
5. Determine the precision followed by the Confusion Matrix.

IMPLEMENTATION

1. Data Collection:

This is the first step to actually developing a machine learning model, collecting data. This is a critical step that will change with how good the model gets as we get more and better data; the better our model will perform. The data comes from government website <https://data.gov.in/> Indian police forces collect the accidents data using the form called Stats19. The data consists of all kind of vehicle collisions from 2005 to 2015. Every column of the dataset is in numerical format. A supporting document to understand each numerical category in accidents dataset is provided on the <https://data.gov.in/> website.

As well as data comes from UCI repository <https://archive.ics.uci.edu/ml/datasets/UrbanGB>

2. Gather data and prepare it for training.

Clean up what this may require (remove duplicates, fix errors, deal with missing values, normalizations and data type conversions, etc.) Randomized data to erase the effects of the particular order in which we collected or otherwise prepared our data. Visualize data to help uncover relevant relationships between variables or class imbalances, or perform other exploratory analysis. Split into training and evaluation sets.

3. Model Used: We used Decision tree, AdaBoost, Naive Bayes, Support Vector Machine, KNN Classifier algorithm.

SVM: during this approach, every knowledge item is aforesought in associate degree n dimensional area, wherever n represents the amount of options with every feature delineated during corresponding co- ordinates. A hyper plane is decided to tell apart the categories (possibly two) supported their options.

Naïve Thomas (NB) Model: The idea behind the NB model is Bayes' Theorem (BT), where the squared event rate is reciprocally excluded just like a dice roll. In addition, BT assumes that entry options collectively referred to as predictor square measure freelance. Similarly, NB jointly assumes that the square of the input options measures the independent nature. However, this is not possible within realistic procedures. Because this

assumption leads to naivety, this rule is called the naive mathematician's rule. Thus, NB can be a probability rule wherever probability decisions are made about input options. On the other hand, in the whole situation of dependent input options, the probability is calculated twice, leading to incorrect results. For higher prediction results in relation to the NB model, square measures of external input options are determined and processed. Dataset collected from Kaggle feed. Functions within the scope of the data set.

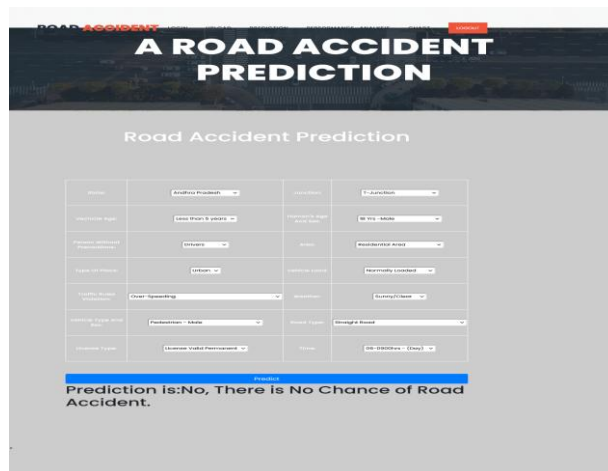
4. Validation:

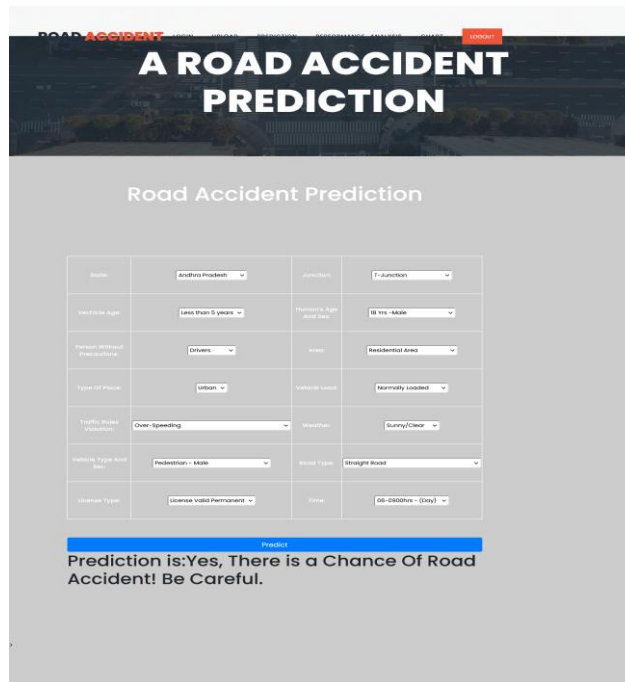
Machine learning, especially supervised learning techniques such as classification and regression, require training data to build a model. The training data consists of labeled data, i.e. datasets that are complete with the target value along with the input feature vectors. A good classification or regression model can be built if a significant amount of training data is provided during the training process. This is followed by a validation process where test data is fed into the trained model to evaluate its predictive accuracy.

5.Screenshots of Application:



Figure 1: Application Interface





Input	Value	Input	Value
Driver's Age	Andhra Pradesh	Gender	T-Junction
Vehicle Age	Less Than 5 years	Driver's Age and Sex	18 Yrs - Male
Person Without License	Others	Area	Residential Area
Type of Road	Urban	Vehicle Load	Normally Loaded
Traffic Rules Violation	Over-Speeding	Location	Sunny/Clear
Vehicle Type and Size	Pedestrian - Male	Road Type	Straight Road
License Type	License Valid (Permanent)	Time	06:00hrs - (Day)

Prediction
Prediction is: Yes, There is a Chance Of Road Accident! Be Careful.

Figure 2: Prediction Result

BLOCK DIAGRAM

The proposed module can be divided into different sections, machine learning, Flask, HTML, CSS, Anaconda-Jupyter notebook. Architecture used in proposed system are given below.

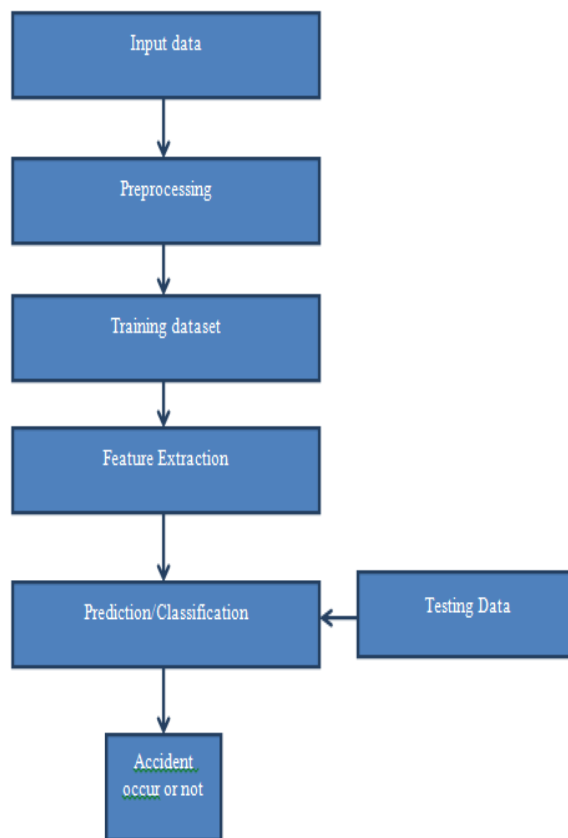


Figure 3: Architecture of System Design

V. RESULT

This section shows a comparative study of all the models that were built. These models are evaluated through accuracy, precision, and f1-score.

The table below presents the values obtained for different metrics from different models. Since the f1-score and accuracy values of most models are similar except for the logistic regression model, we chose to measure model performance using accuracy. It shows that the accuracy of logistic regression is less than other models. The accuracy of Random Forest is also quite low than Gradient Boosting, KNN Classifier and Naïve Bayes. Therefore, we can infer that Gradient Boosting and KNN Classifier do prediction well for our dataset.

Sr. no	Model Used	Accuracy	Precision	F1 score
01	Logistic Regression	83.21	0.86	0.91
02	Naïve Bayes	90.46	0.94	0.95
03	Gradient Boosting	90.35	0.93	0.96
04	Random Forest	85.45	0.88	0.91
05	Knn Classifier	89.78	0.93	0.94

Figure 4: Performance Evaluation of models

We observe that, stage how many people every year died in road accident we analysis this basis on some state of India

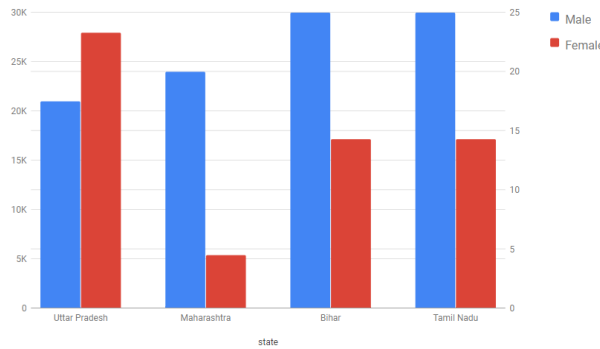


Figure 5: Plot Graph 1-Total Accident in Residential Area

We analysis how many accident in per week in India in various environment

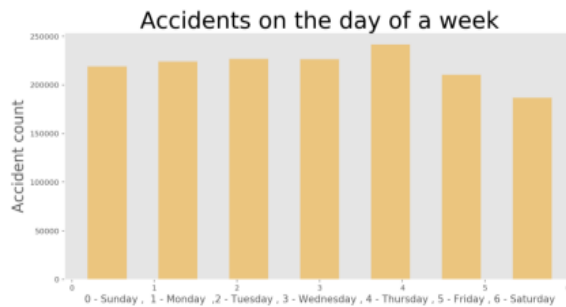


Figure 6:Plot Graphs 2- Accident on the day of week

VI. CONCLUSION

This project aims at using Machine Learning classification techniques to predict severity of an accident at any particular location. Machine Learning has enabled us to analyze meaningful data to provide solutions with a greater accuracy than with humans. . A web-based app using the most accurate algorithm has been developed which can be accessed through the domain name. This project can be used by governments to prevent accidents.

REFERENCES

- [1]. Moprevis. Available online: <https://moprevis.uevora.pt/en/> (accessed on 2 August 2021).
- [2]. Hébert, A.; Guédon, T.; Glatard, T.; Jaumard, B. High-Resolution Road Vehicle Collision Prediction for the City of Montreal. In Proceedings of the 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, 9–12 December 2019.
- [3]. Siam, Z.S.; Hasan, R.T.; Anik, S.S.; Dev, A.; Alita, S.I.; Rahaman, M.; Rahman, R.M. Study of Machine Learning Techniques on Accident Data. In Advances in Computational Collective Intelligence; Hernes, M., Wojtkiewicz, K., Szczerbicki, E., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 25–37.
- [4]. Xu, C.; Wang, W.; Liu, P. Identifying crash-prone traffic conditions under different weather on freeways. *J. Saf. Res.* 2013, 46, 135–144.
- [5]. Yu, R.; Xiong, Y.; Abdel-Aty, M. A correlated random parameter approach to investigate the effects of weather conditions on crash risk for a mountainous freeway. *Transp. Res. Part C Emerg. Technol.* 2014, 50, 68–77
- [6]. Theofilatos, A.A.; Yannis, G. Investigation of Powered-Two-Wheeler accident involvement in urban arterials by considering real-time traffic and weather data. *Traffic Inj. Prev.* 2016, 18, 293–298.
- [7]. Theofilatos, A.; Graham, D.; Yannis, G. Factors Affecting Accident Severity Inside and Outside Urban Areas in Greece. *Traffic Inj. Prev.* 2012, 13, 458–467
- [8]. Iranitalab, A.; Khattak, A. Comparison of four statistical and machine learning methods for crash severity prediction. *Accid. Anal. Prev.* 2017, 108, 27–36
- [9]. Lin, L.; Wang, Q.; Sadek, A. A Novel Variable Selection Method based on Frequent Pattern Tree for Real-time Traffic Accident Risk Prediction. *Transp. Res. Part C Emerg. Technol.* 2015, 55, 444–459
- [10]. DhanyaViswnath,preethi K, NandiniR,Bhuvaneshwari R “ Road Accident Prediction Model Using Data Mining techniques” Proceedings of the Fifth International Conference on Computing Methodologies and Communication(ICCMC 2021)IEEE Xplore Part Number:CFP21K25-ART