UNIVERSITY OF BARISHAL

COMPUTER SCIENCE AND ENGINEERING

A

Thesis Report

On

Traffic Accident Severity Prediction Using Ensemble Machine Learning Methods for Taking Precautions Accordingly

Submitted by:

Mir Samiur Rahim

Roll: 09-001-29

Submitted to:

Dr. Tania Islam

Associate Professor

Department of Computer Science and Engineering

University Of Barishal

Contents

Chapter 1: Overview

1.1 Introduction

Road accidents are a major public health and safety problem. The severity of a road accident is influenced by a number of factors, including speed, vehicle type, seat belt usage, road design, and alcohol and drug impairment etc. There are a number of things that can be done to reduce the severity of road accidents, including enforcing speed limits, promoting seat belt use, improving road design, and combating alcohol and drug impairment. In addition to these general measures, there are also a number of specific things that can be done to reduce the severity of road accidents involving different types of road users, such as pedestrians, cyclists, and motorcyclists. Road accidents can have devastating consequences for the victims, their families, and society as a whole. The most immediate consequence is death or injury. Road accidents can also lead to significant property damage, including damage to vehicles, infrastructure and buildings. This can have a significant financial impact on individuals and businesses. In addition to the physical and financial consequences, road accidents can also have a significant psychological impact on the victims. Overall, the consequences of road accidents are far-reaching and devastating. It is important to take all necessary precautions to prevent road accidents from happening. This includes wearing seat belts and helmets, obeying the speed limit, not driving under the influence of alcohol or drugs, and being aware of your surroundings. The rapid increase in traffic volume on urban roads has changed the global traffic scenario, leading to more frequent and severe road accidents. To improve traffic safety and management, it is essential to predict the severity level of accidents. Machine learning models offer a promising approach to this challenge. In this study the goal is to find an efficient and better classification model that would predict accident severity in most accurate way.

1.2 Problem Statement

In today's world, the number of road accidents is steadily increasing. Accidents in the transportation sector frequently result in fatalities and injuries. There are many countries that have been found where the road accident rate is high. Traffic accidents are a major problem, resulting in millions of deaths and injuries each year. Machine learning algorithms can be used to predict and analyze highway crashes based on the roadway, human and environmental factors.

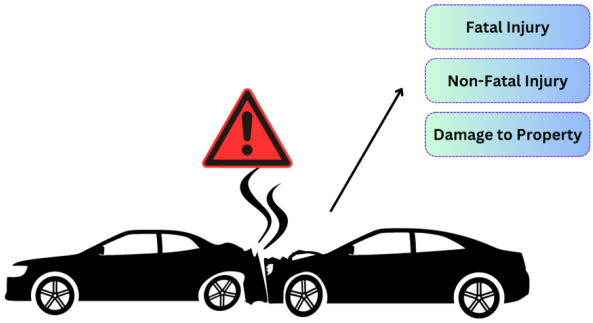

Figure 1.1: Accident Consequences

1.3 Objective

This study aims to achieve the accuracy and identify the factors behind Traffic Accident Severity that could be helpful to reduce accident frequency and severity, thus saving many lives and wealth. Additionally, the study aimed to establish models to select a set of influential factors and to build up a model for classifying the severity of injuries. In simpler terms, traffic accidents are a big problem, and machine learning can help us predict and prevent them. This study is using machine learning to identify the factors that contribute to traffic accidents and to develop models that can be used to predict the severity of injuries. This information can be used by traffic agencies to make highways safer. That's one of the main aspects which makes the system more effective.

This research will make the following key contributions:

1. Comparative analysis of tree-based and regression-based ensemble learning classifiers: Logistic Regression (LR), Naive Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbour (KNN), XGBoost.

2. Analysis of the influence of significant variables on evaluation measures such as accuracy, precision and recall.

Chapter 2: Methodology

2.1 Literature Review

The paper [1] introduces a novel geospatial analysis approach to pinpoint the areas with high road traffic accidents (RTA) in Dhaka Metropolitan Area (DMA). The methodology involves utilizing spatial analysis and spatial statistics tools to scrutinize the spatial distribution of accident data. The results derived from the analysis of accident data within DMA demonstrate the effectiveness of these methods in statistically detecting spatial patterns and clusters of traffic accidents. This research provides a sound basis for identifying and analyzing traffic accident hotspot zones, contributing to enhanced road safety initiatives.

The study [2] explores the application of machine learning models to predict road accident severity, specifically focusing on tree-based ensemble models like Random Forest, AdaBoost, Extra Tree, and Gradient Boosting, as well as an ensemble of two statistical models, Logistic Regression Stochastic Gradient Descent (LR+SGD). The research identifies significant features strongly correlated with accident severity, with Random Forest proving to be the most effective in this regard. Comparatively, the experimental results demonstrate that Random Forest outperforms other methods. The most significant features identified by Random Forest are also used as inputs in ensemble models, resulting in improved accuracy, precision, recall, and F-scores across all models. However, Random Forest consistently emerges as the top-performing model in predicting accident severity.

The central aim of this study [3] is to achieve high accuracy in predicting traffic accident severity and to identify the contributing factors. This information can prove invaluable in reducing the frequency and seriousness of accidents, thereby safeguarding lives, wealth, and other vital resources. Moreover, this study seeks to establish models for selecting influential factors and constructing a severity classification model that can be utilized by Michigan Traffic Agencies (MTA). The resulting model will empower MTA and other responsible authorities in Michigan to proactively address high-risk areas on freeways. This research investigates the effectiveness of four distinct machine learning algorithms in creating precise and reliable classifiers. As per the confusion matrix F1-Score, the test results demonstrate that the Random Forest model appears to outperform the other models. The research findings indicate that these algorithms can predict accidents with an high accuracy rate.

The primary objective of the paper [4] is to shed light on the general characteristics of fatal car accidents and identify the most common types of fatal accidents along with their underlying causes. Fatal car accidents encompass more than just injuries and collisions; they have a broader scope. Notably, in urban areas where car traffic is substantially higher than in rural regions, the accident rate is nearly double. This underscores the importance of comprehensive training for city-based car drivers. This study delves into the realm of car accidents, with a specific focus

on the detailed categorization of accidents based on factors such as location, accident types, collision types, severity, and casualties. These categorizations provide valuable insights into the current landscape and the role of cars in road traffic accidents. Additionally, the paper explores potential accident prevention measures related to engineering, enforcement, and education.

2.2 Background Study

In this research, road accident severity analysis is performed by using ensemble learning models such as Logistic Regression (LR), Naive Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbour (KNN), XGBoost.

2.2.1 Logistic Regression

Logistic regression is a supervised machine learning algorithm that is used for classification tasks. It works by fitting a logistic function to the training data. The logistic function is a sigmoid function that maps the input values to a probability. The output of the logistic regression model is the probability that a data point belongs to a particular class. Logistic regression is a simple and effective algorithm that is easy to implement and understand. Logistic regression is often used in tasks such as:

- Risk assessment: Used to assess the risk of a particular event happening, such as the risk of a customer defaulting on a loan or the risk of a patient developing a particular disease.
- Fraud detection: Used to detect fraudulent transactions.
- Medical diagnosis: Used to diagnose diseases.
- Marketing: Used to predict which customers are likely to respond to a particular marketing campaign.

Logistic regression is a powerful and versatile machine learning algorithm that is well-suited for a variety of classification tasks.

2.2.2 Naive Bayes

Naive Bayes is a classification algorithm that is based on Bayes' theorem. Bayes' theorem is a mathematical formula that allows you to calculate the probability of an event happening, given that you know the probabilities of other related events happening.It is a simple and effective algorithm that is easy to implement and understand. It is also a very interpretable algorithm, meaning that it is possible to understand how the model makes its predictions. Naive Bayes is being used in the real world as:

- Spam filtering: Naive Bayes is used by email providers to filter out spam emails.
- Sentiment analysis: Naive Bayes is used by companies to analyze the sentiment of customer reviews and social media posts.
- Medical diagnosis: Naive Bayes is used by doctors and hospitals to diagnose diseases.
- Fraud detection: Naive Bayes is used by banks and other financial institutions to detect fraudulent transactions.

Naive Bayes is a powerful tool that can be used to improve the performance of many different applications.

2.2.3 Support Vector Machine

Support vector machines (SVMs) are a type of supervised machine learning algorithm that can be used for both classification and regression tasks. SVMs are a very powerful and versatile machine learning algorithm. They are able to handle high-dimensional data and they are also robust to outliers. SVMs are often used in tasks such as:

- Image classification: SVMs are used by companies like Google and Facebook to classify images.
- Text classification: SVMs are used by companies like Gmail and SpamAssassin to classify emails as spam or not spam.
- Medical diagnosis: SVMs are used by doctors and hospitals to diagnose diseases.
- Fraud detection: SVMs are used by banks and other financial institutions to detect fraudulent transactions.

SVMs are a powerful and versatile machine learning algorithm that can be used for a variety of tasks.

2.2.4 Decision Tree

Decision trees are a type of supervised machine learning algorithm that can be used for both classification and regression tasks. Decision trees are a powerful and versatile machine learning algorithm. They are easy to understand and interpret, and they can be used to solve a wide variety of problems. Decision trees are being used in the real world as:

- Fraud detection: Decision trees are used by banks and other financial institutions to detect fraudulent transactions.
- Medical diagnosis: Decision trees are used by doctors and hospitals to diagnose diseases.

- Risk assessment: Decision trees are used by insurance companies and other financial institutions to assess risk.
- Marketing: Decision trees are used by companies to predict which customers are likely to respond to a particular marketing campaign.

Decision trees are a powerful tool that can be used to improve the performance of many different applications.

2.2.5 Random Forest

Random forests is a type of ensemble machine learning algorithm that uses multiple decision trees to make predictions. Random forests is a very powerful and versatile machine learning algorithm. It is able to handle high-dimensional data and it is also robust to outliers. Random forests are often used in tasks such as:

- Classification: Random forests can be used to classify data into different categories, such as spam or not spam, cancer or not cancer.
- Regression: Random forests can be used to predict numerical values, such as the price of a house or the number of customers who will visit a store on a given day.
- Ranking: Random forests can be used to rank items, such as the relevance of search results or the quality of product reviews.

Random forests is a powerful tool that can be used to improve the performance of many different applications.

2.2.6 K-Nearest Neighbour

K-nearest neighbors (KNN) is a simple, supervised machine learning algorithm that can be used for classification and regression tasks. It works by finding the K nearest points in the training dataset and using their class or value to predict the class or value of a new data point. KNN is also a very easy algorithm to implement and understand. Some examples of how KNN is being used in the real world is given below:

- Image classification: Used by companies like Google and Facebook to classify images.
- Medical diagnosis: Used by doctors and hospitals to diagnose diseases.
- Recommendation systems: Used by companies like Netflix and Amazon to recommend products and content to users.
- Fraud detection: Used by banks and other financial institutions to detect fraudulent transactions.

KNN is a powerful and versatile machine learning algorithm that can be used for a variety of tasks. It is a good choice for beginners and for tasks where interpretability is important.

2.2.7 XGBoost

XGBoost is a machine learning algorithm that uses ensemble learning to make predictions. XGBoost is one of the most popular machine learning algorithms because it is very fast and accurate, and it can be used for a variety of tasks, including classification, regression and ranking. Some examples of how XGBoost is being used in the real world is given below:

- Fraud detection: Used by banks and other institutions to detect fraudulent transactions.
- Recommendation systems: Used by companies like Netflix and Amazon to recommend products.
- Medical diagnosis: Used by doctors and hospitals to diagnose diseases
- Risk assessment: Used by insurance companies and other institutions to assess risk.

XGBoost is a powerful and versatile machine learning algorithm that can be used for a variety of tasks. It is particularly well-suited for tasks involving large datasets and for tasks where accuracy is critical.

2.3 Methodology

Incidents of deaths and injuries in road accidents Roads are a regular topic of discussion among locals road users. That's why predicting the severity of traffic accidents is still an active area of research. Machine learning models have shown promise in improving classification accuracy, but there is a lack of comparison between state-of-the-art machine learning models and hybrid models. Finding the best approach and identifying the factors that affect traffic accidents can improve prediction accuracy.

Proposed system will use ensemble learning models to predict the severity of traffic accidents. The ensemble learning models used in this experiment are Logistic Regression (LR), Naive Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbour (KNN), XGBoost. The proposed model will work on following way:

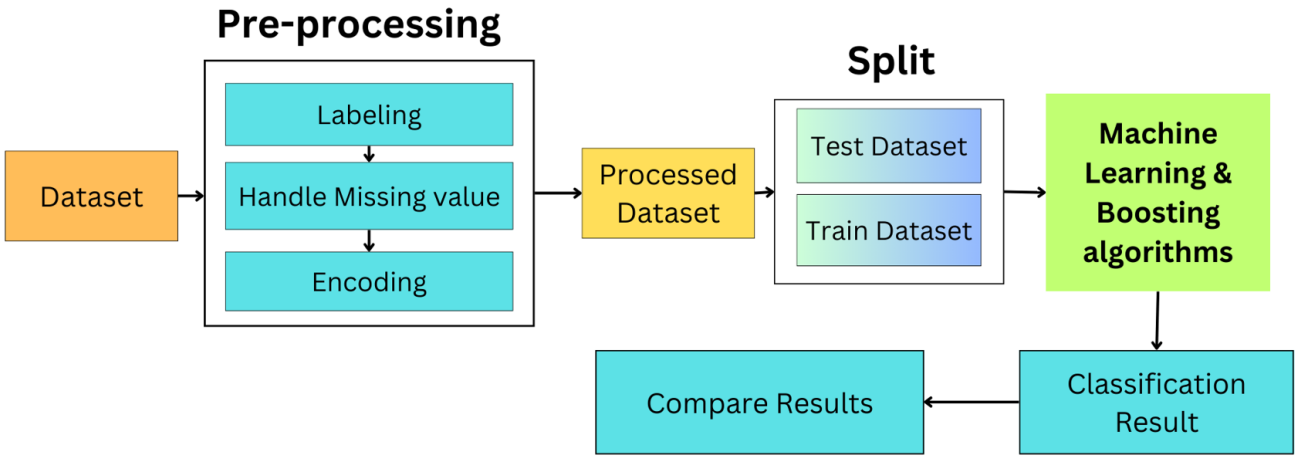

Figure 2.1: Proposed Methodology

2.4 Dataset Description

For checking the applicability of proposed model the most used "US Accidents" is used. This dataset contains car accident records from all 49 states in the USA, from February 2016 to June 2020. There are about 3.5 million records in the dataset, each with 49 columns of data. The columns include the ID of the accident, the source of the data, the traffic management center (TMC) code, the severity of the accident, the start and end times of the accident, the start latitude and longitude, the end latitude and longitude, the number of fatalities, the number of injuries, the types of vehicles involved, the weather conditions and other relevant information.

2.5 Tools to be used

- OS: Windows/Linux/MacOS
- Language: Python
- IDE: Jupyter Notebook/Visual Studio Code/Colab
- Deep learning libraries: TensorFlow/PyTorch
- Additional libraries and packages: NumPy, SciPy, Pandas, scikitlearn

2.6 Verification and Validation

Evaluating machine learning models is important for classification tasks. There are many different evaluation metrics, but accuracy, precision, recall and F-score are among the most common.

True Positive (TP): The number of positive predictions that are correctly predicted.

True Negative (TN): The number of negative predictions that are correctly predicted.

False Positive (FP): The number of negative predictions that are incorrectly predicted as positive.

False Negative (FN): The number of positive predictions that are incorrectly predicted as negative.

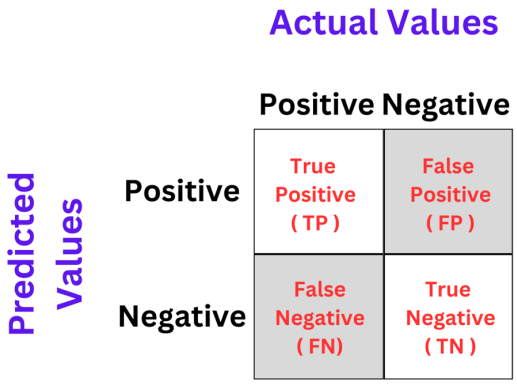

Figure 2.2: Confusion Matrix

Precision is the percentage of all positive predictions that are actually positive.

$$Precision = \frac{TP}{TP + FP}$$

Recall is the percentage of true positive predictions that are correctly predicted.

$$Recall = \frac{TP}{TP + FN}$$

F-score is a measure of both precision and recall. It is calculated as the harmonic mean of precision and recall.

$$F\_score = \frac{Precision.Recall}{Precision + Recall}2$$

And finally Accuracy tells you how often the model makes correct predictions overall.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Chapter 3: Final Discussion

3.1 Rationale of The Research

Accident severity prediction systems use machine learning to identify patterns in historical accident data that can be used to predict the severity of future accidents. This information can be used to prevent accidents and to improve the response of emergency services. Accident severity prediction systems are a valuable tool for improving road safety. By using machine learning to identify patterns in historical accident data, these systems can help to prevent accidents and to improve the response of emergency services.

3.2 Future Work

As there's no precise dataset of Bangladesh accidents is available, that can create a huge impact if we can collect that and predict our local issues with more accuracy.

3.3 Conclusion

Traffic accidents represent one of the most pressing global concerns due to their significant contribution to annual fatalities, injuries and economic burdens. Developing precise models for predicting the severity of traffic accidents stands as a pivotal challenge for transportation systems. This research endeavor aims to create models that identify key influencing factors and construct a framework for categorizing injury severity. Various machine learning techniques are employed to formulate these models.

References

[1] A. Razi, X. Chen, H. Li, B. Russo, Y. Chen, and H. Yu, “Deep learning serves traffic safety analysis: A forward-looking review”, Mar. 2022.

[2] M. Umer, S. Sadiq, A. Ishaq, D. S. Ullah, N. Saher, and H. Madni, “Comparison analysis of tree based and ensembled regression algorithms for traffic accident severity prediction”, Oct. 2020.

[3] R. E. AlMamlook, K. M. Kwayu, M. R. Alkasisbeh, and A. A. Frefer, “Comparison of machine learning algorithms for predicting traffic accident severity”, pp. 272–276, 2019.

[4] H. Ahsan, M. A. Raihan, and M. Rahman, “A study on car involvement in road traffic accidents in bangladesh.”, Dec. 2011.