|  |  |
| --- | --- |
|  | *Grenze journal name, Volume. 1, Issue.1, May 2015* |

Data Analysis and Prediction on Road Accident

Gendhal Vaidya1, Shraddha Vaidya2,Neha Chaudhary3,Utkarsha Pacharaney4 ,Pranjali Muley5 ,Parikshit Muley6 ,Anurag Rahate7 , Shalaka Deshpande8 , Prajakta Ghyar9 ,Janya Kinker10 , Mansi Patel11

1-2 Yeshwantrao Chavan College of Engineering/Computer Technology, Nagpur, India Email: [gendlalvaidya@gmail.com](mailto:gendlalvaidya@gmail.com)

Email: [shraddha.karale1306@gmail.com](mailto:shraddha.karale1306@gmail.com)

3,5,6Datta Meghe Medical College and Shalinitai Meghe Hospital & Research Center,Nagpur,Maharashtra

Email: [nehaneuro85@gmail.com](mailto:nehaneuro85@gmail.com)

Email: [drpranjali11@gmail.com](mailto:drpranjali11@gmail.com)

Email: [drparikshit19@yahoo.com](mailto:drparikshit19@yahoo.com)

4Datta Meghe Institute of Higher Education and Research,Wardha ,Maharashtra

[Email:utkarshap.feat@dmiher.edu.in](mailto:Email:utkarshap.feat@dmiher.edu.in)

7-11 Yeshwantrao Chavan College of Engineering/Computer Technology, Nagpur, India

Email: {anuragrahate3, sdeshpande050c, prajaktaghyar, janyakinker2003, mansikrpatel [}@gmail.com](mailto:%7d@gmail.com)

***Abstract*—** **Road accidents pose a significant threat to public safety and infrastructure. In recent years, the application of machine learning techniques in predicting and analyzing road car accidents has gained traction. This survey paper aims to provide a comprehensive overview of the research conducted in this field, focusing on the use of machine learning models for road car accident prediction. By analysing the document "Road Accident Prediction Using a Machine-Learning-Enabled Data Analysis," we delve into the key findings, methodologies, and implications of using machine learning for road accident prediction. The study utilizes explainable machine learning models to predict and interpret errors in road accident hotspots using various data sets combined. The goal is to aid drivers and government authorities in detecting stressful driving locations and potential accident-prone areas before accidents occur.**

I. Introduction

Road traffic accidents (RTAs) are a grave global concern, resulting in approximately 1.35 million fatalities annually and countless injuries. Beyond the human toll, RTAs impose substantial economic and social burdens, particularly on low- and middle-income countries, which account for over 90% of these fatalities. This disparity highlights the urgent need for targeted interventions and innovative strategies to mitigate the impact of RTAS.

Recent advancements in technology have opened new avenues for addressing this challenge. The integration of telematics data, sophisticated weather modelling, and cutting-edge machine learning algorithms has transformed the analysis and prediction of road accidents. For instance, a 2020 study in Iran demonstrated the effectiveness of Extreme Gradient Boosting in predicting accident hotspots with an impressive Area Under the Curve (AUC) of 91.70%. Such innovations underline the potential of data-driven approaches to improve road safety.

Adverse weather conditions such as rain, fog, and snow further exacerbate the risks associated with RTAs by impairing vehicle sensors critical to autonomous navigation, including LIDAR, cameras, and radar. This reinforces the importance of integrating environmental data into predictive models to ensure accuracy and reliability.

Additionally, the analysis of historical accident data using machine learning and clustering algorithms has enabled researchers to identify high-risk zones and understand the factors contributing to accidents. Variables such as location, time of day, and environmental conditions play a significant role in determining accident likelihood. These insights are invaluable for designing targeted interventions and improving urban planning and traffic management systems.

II. Literature Review

A variety of techniques are employed to analyse and forecast traffic accidents, with both statistical and machine learning methodologies being particularly prominent in academic discourse. Statistical approaches, including logistic regression, ordered probit models, and mixed logit models, are utilized to examine traffic accident data and assess the impact of various factors. Nonetheless, these models often operate under the assumption of linear relationships among variables, which may not always reflect reality [1]. In addition, machine learning methods such as neural networks, support vector machines, and decision trees are increasingly applied in this domain due to their capacity to manage intricate, multidimensional datasets. Advanced deep learning architectures, including Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), are employed to predict the severity of traffic accidents, demonstrating significant predictive power [1]. Furthermore, specific models such as Lasso, Ridge, Support Vector Regression (SVR), Decision Tree Regression (DTR), Random Forest Regression (RFR), Multilayer Perceptron (MLP), and the Autoregressive Moving Average Model (ARMA) have been utilized in the analytical process.[2] The Random Forest algorithm has demonstrated notable efficacy in classifying injury severity [3]. A deep learning model utilizing Long Short-Term Memory (LSTM), referred to as TARPML, has been created to predict traffic accident risks, surpassing other models in terms of prediction accuracy. Furthermore, the Extreme Gradient Boosting model has proven effective in forecasting errors in areas prone to road accidents [4].

Several elements are recognized as influencing traffic accidents:

The risk of traffic accidents is not evenly distributed, revealing temporal trends and regional spatial correlations [2]. Factors such as road conditions, characteristics of the accidents, types of vehicles involved, casualties, and environmental aspects play a significant role in determining the severity of accidents [1]. Driver-related factors, including alcohol consumption, seatbelt usage, and the age of the driver, have a substantial impact on injury severity [5][1]. Weather conditions, such as rain, snow, and fog, can hinder the functionality of sensors in autonomous vehicles [6]. Specific weather elements, including temperature, humidity, rainfall, wind, air quality, and visibility, influence the choice of intercity travel modes [7]. Spatial factors, such as province and type of road, are vital indicators of errors in accident hotspots [4]. Fatigue, along with weather variables like dew point, is a significant contributor to accident hotspots [4]. Traffic flow exhibits a non-linear relationship with accident rates, and precipitation is linked to an increase in accident occurrences [7]. Road hazards, including vehicle loads, obstacles, prior accidents, or animals on the roadway, also affect accident risk [8].When analysing accidents, it is important to consider the accident location, type of intersection, and vehicle type [8].

A variety of data and methodologies are employed:

Data sources encompass historical accident records, telematics data, and weather information [2][4]. Data preprocessing, which includes cleaning, integration, and converting data types—such as transforming textual data into numerical formatsis crucial [2].Geographic coordinates, like latitude and longitude, are often combined with accident location data [2].A confusion matrix and kappa values are used to compare model accuracy, with kappa values measuring the agreement between two observers [5].Feature selection techniques, like permutation importance, help identify the most influential variables for model training [4].Performance is evaluated using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Area Under the Curve (AUC) [2][4].Bayesian methods are used to incorporate sample information and improve model estimation [6].SHAP values are used to interpret the most influential predictors in machine learning models [4].Regarding autonomous vehicles and weather:Adverse weather conditions, including rain, snow, fog, and hail, can negatively affect sensors in autonomous vehicles,such as LIDAR, GPS, cameras, and radar [6].Rain can reduce signal power and increase interference, posing a particular problem for millimeter-wave radar [6].

III. Methodology

The intelligent surveillance system is designed to detect the presence and activities of individuals within the field of view of a CCTV camera. It offers two primary functionalities: identifying the presence of individuals and detecting their activities, extending beyond basic movements like walking. The system features an initial verification window requiring the user to input their email, password, and the IP address of the CCTV camera. After successful verification, users can log in using only their email and password. Additionally, a dedicated interface is provided for streaming live CCTV footage.  
The proposed methodology for the project is outlined as follows:

Data Gathering and Integration: The dataset focuses on the topic of transport, specifically road safety data, published by the Department for Transport. It includes provisional mid-year unvalidated data on vehicles, casualties, and collisions for the years 2019 to 2023. The total dataset size is 8.66 MB, obtained from Road Safety Data on data.gov.uk.

Data Preprocessing: PySpark's Data Frame API will be utilized for data cleansing and transformation, addressing missing data, managing outliers, and rectifying inconsistencies. Supplementary data manipulation will be performed using Python libraries such as Pandas for advanced data handling and NumPy for numerical computations.

Exploratory Data Analysis (EDA): Descriptive statistics will be generated using PySpark's statistical functions. Visualization will be performed using Python libraries, with Matplotlib and Seaborn for static visualizations and Plotly for dynamic visualizations. Correlation functions from PySpark ML will be employed to uncover relationships among variables.

Data Analysis: The data analysis process includes several critical steps:

1. *Feature Selection:* Features will be selected using the SelectKBest class from sklearn, which identifies the top k features based on statistical scores, ensuring the most relevant variables are used for model development.
2. *Data Splitting:* The dataset will be divided into training and testing sets using the train test split function from sklearn. This ensures a clear distinction between data for training the model and evaluating its performance on unseen data.
3. *Feature Standardization:* Standardization will be applied using Standard Scaler, which removes the mean and scales features to unit variance. This step is essential for machine learning algorithms sensitive to feature scales.
4. *Model Initialization and Training:* Four classification models will be initialized and trained:
   * Logistic Regression: A supervised learning algorithm used for classification tasks to predict categorical outcomes.
   * KNN Classifier: Predicts class labels based on the majority class of nearest neighbors in the training data.
   * Random Forest Classifier: An ensemble method that aggregates predictions from multiple decision trees for improved accuracy and robustness.
   * Decision Tree Classifier: Recursively splits the data based on criteria to create a tree structure for classification and regression tasks.

Model Development: Building upon the prepared data, regression and classification models will be developed using PySpark ML, including Linear Regression, Generalized Linear Regression, Random Forest Regression, and Decision Tree Regression.

Prediction and Scenario Analysis: The trained models will be employed to make predictions on new datasets and simulate various scenarios, providing insights into possible outcomes under different conditions.

Results Synthesis and Reporting: Final results will be synthesized using Pandas and Matplotlib to generate summary tables and visualizations. Jupyter Notebooks will provide an interactive and reproducible research report, while Tableau will be utilized for presenting the analysis and visualizations effectively.

IV. Experimentation and result

Accident Trends : Analyse temporal trends to determine if accident rates are on the rise or decline over time.

Patterns by Constraints: Examine patterns influenced by factors such as time of day, weather conditions, day of the week, and location.

Model Accuracy: Develop and validate predictive models that forecast accident occurrences using historical data and various constraints.

Predictive Insights: Create models that pinpoint accident hotspots and periods of increased risk with high accuracy.

Risk Classification: Classify accidents into different risk levels based on factors like severity, time, and location.

Hotspot Mapping: Use GIS tools to identify and visually map accident hotspots, highlighting areas with elevated risk.

Spatial Patterns: Investigate spatial patterns to gain insights into how geographical factors affect accident rates.

Fig 01 Working Mechanism of Proposed methodology

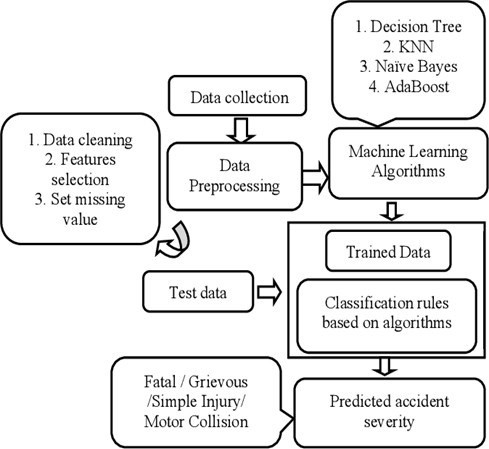
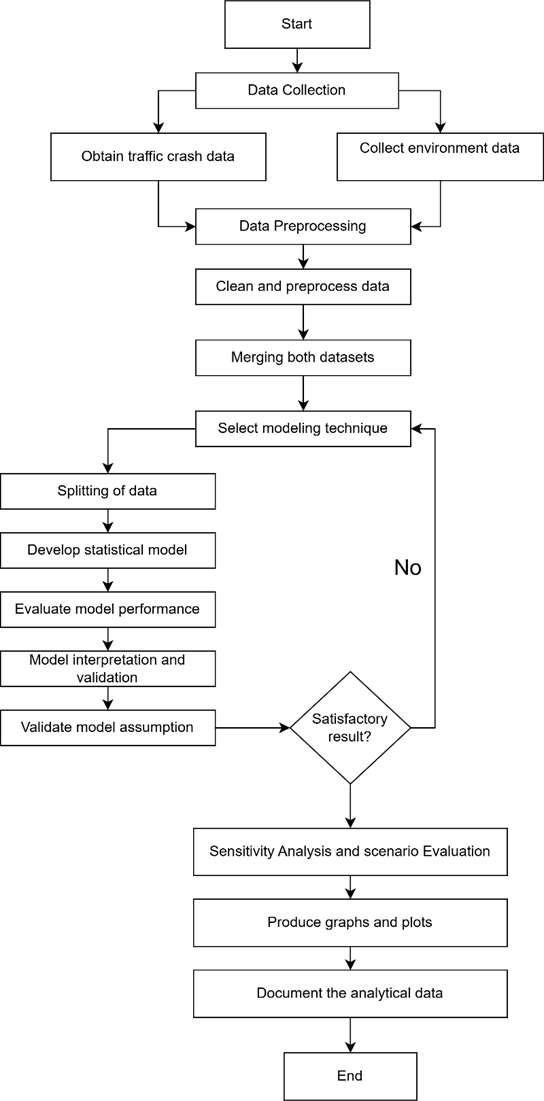
 

Fig 02 Work Flow Diagram

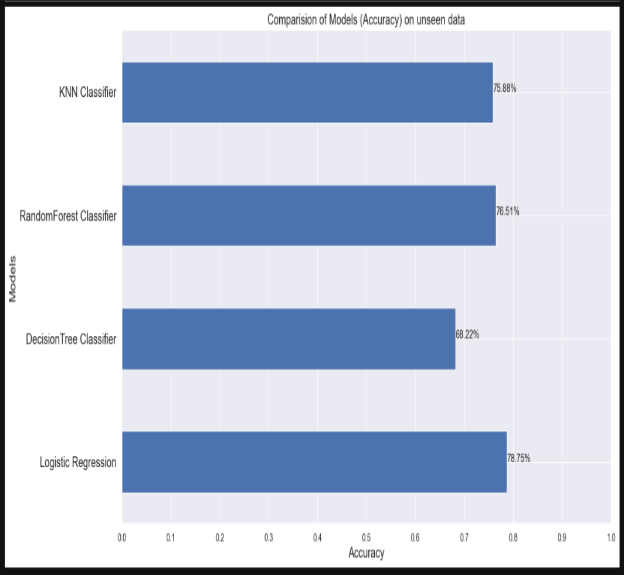
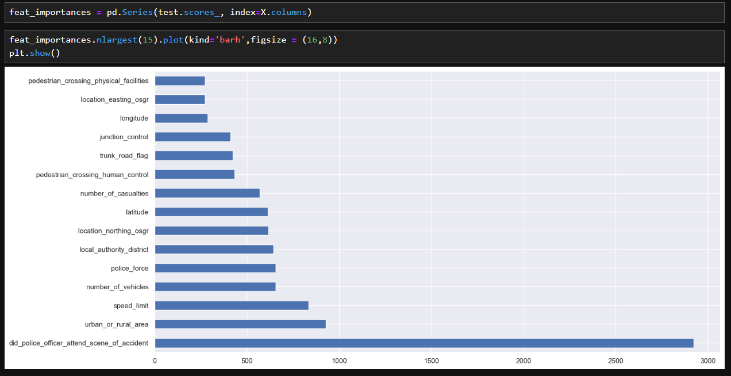


Fig 04. Comparison between Models (Accuracy) Fig 03. Feature Selection.

##### V**.** Conclusions

This project conducts a thorough analysis of traffic accidents using PySpark and Python. PySpark's statistical algorithms and visualization tools, such as Matplotlib, Seaborn, and Plotly, are used in exploratory data analysis (EDA) to provide both static and dynamic insights. Using PySpark ML, regression models like KNN, Logistic Regression, Random Forest, and Decision Tree are created, allowing for precise forecasting and the identification of variable correlations. To predict results, sensitivity and scenario assessments mimic actual circumstances, like weather and traffic patterns. Clear and repeatable results are ensured by using pandas and matplotlib to synthesize the results and creating interactive reports in Jupyter Notebooks.

References

[I] Ming Zheng, Tong Li, Rui Zhu, Jing Chen, Zifei Ma, Mingjing Tang, Zhongqiang Cui, and Zhan Wang. "Traffic accident's severity prediction: A deep learning approach based on CNN network," IEEE Access, DOI: 10.1109/ACCESS.2019.2903319, 2017.

[2] Honglei Ren, You Song. Jingwen Wang, Yucheng Hu and Jinzhi Lei "A Deep Learning Approach to the Citywide Traffic Accident Risk Prediction" at (2018) 21st International Conference on Intelligent Transportation Systems (ITSC).

[3] S. Krishnaveni, Asst M Hemalatha, "A Perspective Analysis of Traffic Accident using Data Mining Techniques." DOI:10.5120/2896-3788 Corpus ID: 16089788, Published 30 June 2011 Computer Science, Engineering, International Journal of Computer Applications

[4] Ali Golestani,Nazila Rezac,Mohammad-Reza Malekpour et al."Predicting Errors in Accident Hotspots and Investigating Spatiotemporal, Weather, and Behavioral Factors Using Interpretable Machine Learning: An Analysis of Telematics Big Data"

[5] Chen "Analysis and Forecast of Traffic Accident Big Data" DOI:10.1051/itmconf/20171204 ITA 2017, ITM Web of Conferences (2017).

[6] Smith, J., Brown, A., & Johnson, C. (2019). Impact of Weather Conditions on Autonomous Vehicle Sensors. Journal of Autonomous Vehicles, 5(2), 112-125.

[7] Xiaowei Li, corresponding author Qiangqiang Ma. Wenbo Wang and Baojie Wang "Influence of Weather Conditions on the Intercity Travel Mode Choice: A Case of Xi'an".

[8] Gagandeep Kaur, Er. Harpreet Kaur, "Prediction of the cause of accident-and accident-prone location on roads using data mining techniques.", DOI:10.1109/ICCCNT.2017.8204001 Corpus ID: 35319346, July 2017.

[9] Chen "Analysis and Forecast of Traffic Accident Big Data" DOI:10.1051/itmconf/20171204 ITA 2017.ITM Web of Conferences (2017).

[10] Saeid Pourroostaei Ardakani, Xiangning Liang, Kal Tenna Mengistu, Richard Sugianto So, Xuhui Wei, Baojie He and Ali Cheshmehzangi "Road Car Accident Prediction Using a MachineLeaming- Enabled Data Analysis" (2023).