Road Accidents Severity Prediction

Tamilarasi M
Department of Computer Science,
Bharathiar University, Coimbatore,
Tamilnadu, India 641046
Email:tamilarasi2002m@gmail.com

Geetha K
Assistant Professor
Department of Computer Science,
Bharathiar University, Coimbatore,
Tamilnadu, India 641046
Email: geethakab@gmail.com

Abstract— **Road accidents are a major public safety concern worldwide, causing significant loss of liinjuries,and financial implications. Annually, road accidents claim approximately 1.35 million lives and leave millions more injured, making it a leading cause of death, especially among younger populations. The need for effective prediction models to identify factors contributing to accident severity is critical in preventing serious outcomes. Predicting accident severity (slight, serious, or fatal) allows authorities to implement targeted safety measures, efficiently allocate emergency resources, and design preventive strategies for high-risk scenarios. Algorithms such as Decision Trees, Random Forest, and XGBoost are tested, with ensemble methods and class weighting employed to enhance performance. Metrics like accuracy, F1-score, and recall assess model effectiveness in distinguishing between severity levels. This work aims to improve road safety by leveraging machine learning to predict accident outcomes with greater accuracy.**

Index Terms— Decision Tree, Random Foreset, XGBoost, Road Accident, Road Safty.

I. INTRODUCTION

The need for effective prediction models to identify factors contributing to accident severity is critical in preventing serious outcomes. Predicting accident severity (slight, serious, or fatal) allows authorities to implement targeted safety measures, efficiently allocate emergency resources, and design

preventive strategies for high-risk scenarios. This project aims to develop a robust machine learning model to predict road accident severity based on historical data. By analysing factors like weather, road conditions, lighting, vehicle type, and location, the project seeks to identify patterns influencing accident severity.accident severity based on historical data. By analysing factors like weather, road conditions, lighting, vehicle type, and location, the project seeks to identify patterns influencing accident severity. Given the imbalanced dataset, with "slight" accidents outweighing "serious" and "fatal" ones, techniques like SMOTE are used to balance the data and improve predictions for minority classes.Road accident severity prediction has become a critical area of research, especially with the increasing rates of urbanization and the accompanying rise in vehicular traffic. Accidents in densely populated urban areas often result in severe outcomes, leading to economic losses, injuries, and fatalities. In high-traffic environments, the complexity of predicting accident severity increases due to various contributing factors, including driver behaviour, environmental conditions, road infrastructure, and socioeconomic factors.In urban areas, studies show that factors such as road lighting, weather, time of day, and traffic density significantly influence accident outcomes. The use of ensemble models, including Random Forests and Gradient Boosting, has proven beneficial in these settings due to their ability to integrate diverse, categorical data and reveal interactions among variables. Additionally, advancements in real-time data processing have allowed predictive models to incorporate dynamic traffic data, further improving their effectiveness in urban traffic environments. These predictive insights are valuable for policymakers and city planners aiming to design safer urban spaces and deploy emergency resources more efficiently. The ongoing development of machine learning techniques and data-driven approaches to severity prediction holds promise for reducing the impact of accidents in high-risk, high-density areas.

II. LITERATURE SURVEY

Road Accident Severity Prediction in High-Traffic Urban AreasKumar, R., & Singh, H. Accident Analysis & Prevention, This study by Kumar and Singh highlights the effectiveness of Support Vector Machines (SVM) in classifying accident severity in densely populated areas. By analysing variables such as road type, lighting, and time, the study demonstrates SVM’s ability to capture non-linear relationships in high-dimensional data, achieving high accuracy on large datasets. The authors emphasize that SVM-based predictions can enhance resource allocation by emergency teams, potentially lowering fatalities in severe accidents.

Comparing Machine Learning Algorithms for Predicting Accident SeverityAlam, M., & Rahman. Accident Analysis & Prevention, Alam and Rahman compare various algorithms, including Decision Trees, k-Nearest neighbours (k-NN), and Gradient Boosting, for predicting accident severity. Their findings suggest that Gradient Boosting outperforms others by effectively handling imbalanced datasets common in accident data. The study also underscores the importance of feature selection, as redundant features can degrade performance, especially in high-dimensional models like Decision Trees and k-NN.

Gradient Boosting for Severity Classification in Real-Time Accident Prediction Systems, Lee, H., & Park, J. International Journal of Environmental Research and Public Health,Lee and Park apply Gradient Boosting to classify accident severity in real-time traffic conditions, focusing on densely populated urban areas. The model integrates real-time data, such as traffic volume and weather conditions, achieving high accuracy and quick processing times. The study highlights that Gradient Boosting’s ability to handle both continuous and categorical data makes it well-suited for accident prediction in dynamic environments.

Analysing Road Accident Severity with Logistic Regression and Feature Engineering Techniques,Ali, S., & Qureshi. Accident Analysis & Prevention.Ali and Qureshi employ logistic regression combined with feature engineering to evaluate the impact of engineered features like accident location and time on severity. Their analysis shows that engineered features significantly improve logistic regression's predictive power, providing insights into how context-specific variables influence accident severity.

III. METHODOLOLGY

A. Data Collection

The platform also offers Kaggle Notebooks (formerly Kaggle Kernels), where users can write, test, and execute Python or R code in the cloud directly from a browser, eliminating the need for powerful local hardware. For computationally intensive tasks, Kaggle provides free access to GPU and TPU resources, making it easier to work with deep learning models and large datasets.

B. Dataset Overview

Driver Information: Age, gender, education level, driving experience, and relationship to the vehicle (owner or employee).Vehicle Information: Type of vehicle, ownership status, and service years Accident Conditions: Time, day of the week, road and weather conditions, light conditions, and road alignment Collision and Casualty Data: Type of collision, number of vehicles involved, number of casualties, and casualty class (e.g., driver, passenger, pedestrian).Causation Factors: Specific causes such as careless driving, improper vehicle distancing, and failure to yield right-of-way.

C. Data Preprocessing

Data preprocessing is a fundamental step in preparing raw data for analysis and modelling. It involves a series of operations aimed at cleaning, transforming, and organizing the data to make it suitable for further processing.The dataset contains several columns with missing values, particularly in fields related to vehicle service years, vehicle defects, and casualty information. Columns with a high proportion of missing values, such as vehicle service year, vehicle defects, casualty work, fitness of casualty, and time, were dropped, as they offered limited value after significant data loss.Encoding categorical variables were encoded to make them suitable for machine learning models. For nominal variables, such as Educational level and Vehicle driverrelation, one-hot encoding was applied to create binary columns for each category. Ordinal variables, like Drivingexperience, were encoded using label encoding to assign integer values based on their inherent order.The dataset contains numerous categoricalcolumns,such as Day_of_week, Age_band_of_driver, Sex_of_driver, and Type_of_vehicle. These fields must be encoded numerically for the machine learning models to process them effectively.

D. Exploratory Data Analysis (EDA)

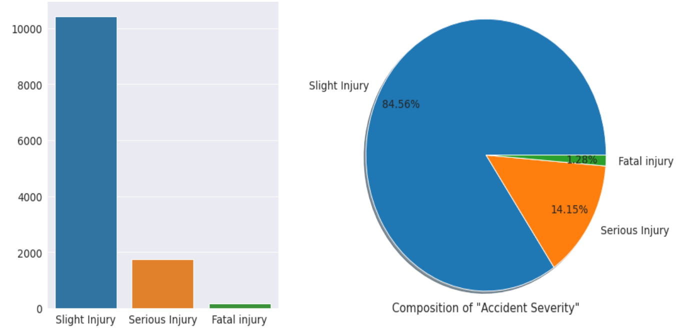

Figure1. Bar and Pie Chart

This figure shows the distribution of road accident severity, with two different visualizations: a bar chart and a pie chart.

Bar Chart:

The bar chart clearly shows the number of accidents for each severity level.The majority of accidents resulted in "Slight Injury," followed by "Serious Injury," and a very small number in "Fatal Injury."

Pie Chart:

The pie chart provides a visual representation of the proportions of each severity level.It emphasizes that "Slight Injury" accounts for the largest portion (84.56%) of accidents."Serious Injury" makes up 14.15%, and "Fatal Injury" is the smallest category at 1.28%.Overall, the figure demonstrates that most road accidents have relatively minor consequences, with a small percentage resulting in serious or fatal outcomes.

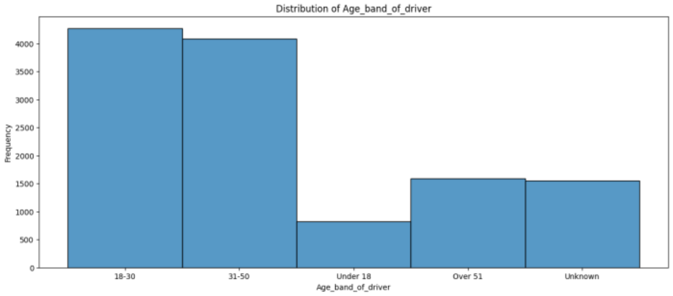

Figure2.Bar Plot

This bar chart shows the percentage of total accidents by age band. The age band with the highest percentage of accidents is 18-30, followed by 31-50. The age band with the lowest percentage of accidents is Under 18.

E. Data Splitting

The dataset was divided into training and test sets using an 70-30 split. This allows the model to learn patterns from the training data and be validated on unseen data. The test set, which represents real-world data, helps assess model generalizability and prevents overfitting.

F. Detection and Removal

Outliers, particularly in features like vehicle speed or traffic volume, were carefully examined. While certain outliers might represent genuine rare events, removing extreme values in features like accident severity or vehicle count prevents these points from unduly influencing the model. Outlier detection methods, such as z-scores or IQR, can help flag values that deviate significantly from the norm.

G. Handling Class Imbalance

Class imbalance is a significant issue in accident severity prediction, where severe accidents (e.g., fatal or serious) are often underrepresented compared to minor accidents. This imbalance can lead to biased models that tend to predict the majority class (minor accidents) more frequently, neglecting the minority class (severe accidents). SMOTE (Synthetic Minority Over-sampling Technique), which generates synthetic samples of the minority class by interpolating between existing minority class instances. This increases the representation of severe accidents in the dataset without duplicating the original data.In addition to SMOTE, random oversampling and under sampling methods are applied. Random oversampling replicates instances of theminority class, while random under sampling reduces the number of majority class instances. This reduces the risk of bias and allows the model to better predict rare events, like fatal accidents, which are crucial for public safety. These methods not only improve model accuracy but also enhance performance metrics like recall, precision, and the F1-score for the minority class, ensuring that severe accidents are identified more reliably. Proper handling of class imbalance results in a more robust and fair predictive model, capable of making accurate predictions for both common and rare accident types.

IV. MODEL BULIDING

Decision Tree

Decision Tree algorithms are powerful tools for predicting road accident severity by utilizing features such as weather, road type, time of day, and vehicle characteristics. The model recursively splits data based on the most influential feature at each node, forming a tree structure that categorizes accident severity levels (e.g., fatal, major, or minor). For instance, it may assign higher severity to accidents occurring on highways or in poor weather.Its interpretability is valuable for understanding decisions, as each split represents a clear choice based on feature values.Decision Trees can handle both numerical and categorical data, adapting well to accident-related datasets.

Random Forest

Random Forest is a powerful ensemble learning algorithm that significantly enhances road accident severity prediction by combining the outputs of multiple decision trees. Each decision tree in the forest is trained on a random subset of both the data and the features, which reduces the risk of overfitting that can occur with a single decision tree. This process improves the model's ability to generalize to new, unseen data. Random Forest takes into account various factors such as weather conditions, road type, traffic density, and vehicle speed to predict accident severity, ensuring accurate classifications of accidents as minor, major, or fatal. This ensemble approach allows Random Forest to capture complex interactions between variables that may not be easily detected by simpler models. It also excels in handling noisy, missing, or incomplete data, which is common in accident datasets, making it highly robust for real-world applications. Random Forest is particularly effective in dealing with imbalanced datasets, a common challenge in accident severity prediction, where severe accidents are less frequent than minor ones. Random Forest can efficiently process large datasets, ensuring that it provides reliable road safety predictions that can inform decision-making and improve traffic management systems.

Support Vector Machine

Support Vector Machine (SVM) is a robust supervised learning algorithm widely used for classifying road accident severity into categories such as fatal, major, or minor. It operates by identifying the optimal hyperplane that maximizes the margin between different classes, enhancing generalization and preventing overfitting. makes SVM highly effective in predictive modeling, especially for tasks like accident severity prediction where accuracy is critical.SVM is particularly well-suited for high-dimensional datasets, which often include a variety of features like weather conditions, road type, traffic volume, and vehicle speed. These features are essential for understanding the severity of accidents and can vary widely, making high-dimensional space a natural fit for SVM.

XGBoost

XGBoost is a powerful ensemble learning algorithm that excels in predicting road accident severity by boosting the performance of multiple weak learners, typically decision trees, through iterative training. Each tree in XGBoost is sequentially trained to correct the errors of its predecessor, which allows the model to reduce bias and improve predictive accuracy over time. In the context of road accident severity, XGBoost can handle a wide range of influential features, such as weather conditions, time of day, traffic density, and vehicle characteristics, to accurately classify accident outcomes as minor, serious, or fatal. By leveraging techniques such as gradient boosting and regularization, XGBoost effectively minimizes overfitting, making it a suitable choice for complex,real-world datasets that may contain noise or imbalanced classes

V. MODEL EVALUATION

A.Accuracy analysis

Accuracy is a fundamental evaluation metric, representing the overall proportion of correct predictions made by the model. In road accident severity prediction, accuracy indicates how well the model can differentiate between various severity levels, such as fatal, major, and minor accidents. For instance, a high accuracy percentage on the training and test datasets would suggest that the model is effective at identifying the correct severity class.

B.Precision and Recall analysis

Precision and recall are essential for evaluating a model’s effectiveness in predicting accident severity. Precision focuses on minimizing false positives, ensuring that severe accidents are correctly identified without over-predicting. Recall, on the other hand, measures the model’s ability to capture all actual severe accidents, which is particularly important for rare events like fatal or major accidents. In areas with higher accident risks, recall ensures that no severe accidents are missed. Together, precision and recall provide a comprehensive understanding of the model’s predictive performance.

C.F1-Score

The F1-score combines precision and recall into a single metric, providing a balanced view of the model’s performance. In cases of imbalanced accident data, it prevents bias toward the majority class. The F1-score ensures the model is both sensitive to severe accidents (recall) and accurate (precision). This makes it crucial for real-world applications where accurate predictions of severe accidents are essential for resource allocation and emergency response.

D. Confusion Matrix

A confusion matrix for road accident severity prediction typically consists of a 3x3 matrix (for a three-class classification problem), where each cell represents the count of instances for a specific combination of predicted and actual values. For instance, the rows of the matrix correspond to the true classes (actual accident severity), while the columns represent the predicted classes. Here's a simplified example of a confusion matrix for a model that classifies accidents into three severity categories.**True Positive (TP)**: The number of correctly predicted accidents for each severity level (e.g., the number of correctly identified fatal accidents).**False Positive (FP)**: The number of times an accident was incorrectly predicted as a more severe category than it actually was (e.g., predicting a fatal accident when it was actually minor).**False Negative (FN)**: The number of times an accident was incorrectly predicted as a less severe category than it actually was (e.g., predicting a minor accident when it was actually fatal).**True Negative (TN)**: The number of correctly predicted instances of other severity levels (e.g., correctly predicting non-fatal accidents).making it a reliable tool for predicting road accident severity and assisting in timely all.

Decision Tree Classifier

The Decision Tree Classifier achieved an accuracy of 85% in predicting road accident severity. It demonstrated a high precision of 0.94 for the Fatal Injury class, reliably predicting fatalities. For Serious Injury, the precision was 0.77, indicating the model's capability to identify severe accidents, though improvements are possible. The recall for Fatal Injury was 0.99, capturing almost all fatal accidents. The recall for Slight Injury was slightly lower at 0.71 but still strong. The model's overall F1-score of 0.85 reflects balanced performance across all accident severity categories. This makes the Decision Tree Classifier a reliable tool for road accident prediction.

Support Vector Machine

The Support Vector Machine (SVM) model achieved an accuracy of 87.3% in predicting road accident severity. It showed high precision for the Fatal Injury class with a score of 0.92, accurately identifying fatal accidents. Its recall for fatal accidents was strong at 0.96, detecting96% of actual fatal incidents. For the Serious Injury class, the precision was 0.88, while the recall was 0.74, indicating some room for improvement. The Slight Injury class had a precision of 0.82 and recall of 0.92, effectively recognizing minor accidents. With an overall F1-score of 0.87, the model performed well across all categories. This demonstrates SVM's reliability for real-time accident severity prediction and resource allocation.

XGBosstClassifier

The XGBoost (XGBClassifier) model achieved an accuracy of 83.8% in predicting road accident severity. It demonstrated strong precision for the Fatal Injury class with a score of 0.88, reliably identifying fatal accidents. Its recall for fatal accidents was 0.93, capturing 93% of actual fatal incidents. For the Serious Injury class, precision was 0.80, with a recall of 0.73, indicating room for improvement in capturing all serious injuries. The Slight Injury class had a precision of 0.82 and recall of 0.85, effectively identifying minor accidents. With an overall F1-score of 0.84, XGBoost showed balanced performance across all categories. This model is well-suited for real-time applications in intelligent transportation systems.

Random Forest

The Random Forest model achieved an impressive accuracy of 90.7% in predicting road accident severity. It showed exceptional precision for the Fatal Injury class with a score of 0.99, accurately predicting fatal accidents. Its recall for fatal accidents was also 0.99, identifying nearly all severe incidents. For the Serious Injury class, precision was 0.85, and recall was 0.89, reflecting solid classification performance. In the Slight Injury class, the model achieved precision of 0.89 and recall of 0.83, effectively identifying minor injuries. With an overall F1-score of 0.91, Random Forest demonstrated balanced performance across all categories. This makes it an excellent choice for real-time accident severity prediction.

Table 5.1 Performance Analysis

MODEL	ACCURACY
Random Forest	90.7%
Decision Tree	85%
Support Vector Machine	87.3%
XG Boost	83.8%

VI. CONCLUSION

The analysis on road accident data reveals significant insights into accident severity and contributing factors. The data, consisting of variables like driver age, vehicle type, and road conditions, shows distinct trends. Young to middle-aged drivers and male drivers are predominantly involved in accidents, often with severe or fatal outcomes. The most frequent accident causes include "no distancing" and "careless driving," highlighting a need for stricter road safety education. Accidents primarily occur under "daylight" and "normal" weather, suggesting human error as a major factor rather than environmental conditions. Additionally, data indicates that accident severity is heightened in cases of vehicle collisions and rollovers. Various predictive models, such as SVM model and Random Forest, were applied, with accuracy scores improving significantly after data balancing and feature selection. These findings suggest a multifaceted approach to accident prevention, encompassing both driver behaviour and improved vehicle safety protocols. This dataset provides a foundation for policymakers to address critical areas in road safety and implement targeted interventions.

The feature scope for predicting road accident severity involves a comprehensive set of variables that capture driver attributes, vehicle characteristics, accident specifics, and casualty information. Driver attributes include factors such as age, gender, education level, and driving experience, which provide insight into possible risk profiles. Vehicle characteristics encompass the type, ownership status, service years, and any known defects, which may affect accident severity. Accident-specific factors include the time, day, location, road alignment, surface type, light, and weather conditions, along with the type of collision and number of vehicles involved, all of which influence the likelihood and impact of severe outcomes. Additionally, casualty information details the classification and severity of injuries, differentiating between minor, serious, and fatal injuries. This comprehensive set of features allows for a detailed predictive analysis, improving the ability to assess and mitigate road accident severity in various conditions.

REFERENCES

[1] Chawla, N. V., Bowyer, K. W., Hall, L. O., &Kegelmeyer, W. P. (2002). *SMOTE: Synthetic Minority Over-sampling Technique.* Journal of Artificial Intelligence Research, 16, 321–357.

[2]Breiman,L.(2001).*Random Forests.* Machine Learning,45(1), 5–32.

[3] Chen,T.,&Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System.* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

[4] Garcia, S., Luengo, J., & Herrera, F. (2015). *Data Preprocessing in Data Mining.*Springer.

[5] Zhou, Z. H., & Liu, X. Y. (2006). *Training Cost-Sensitive Neural Networks with Methods Addressing the Class*

Imbalance Problem. IEEE Transactions on Knowledge and Data Engineering, 18(1), 63–77.

[6] López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). *An Insight into Classification with Imbalanced Data: Empirical Results and Current Trends on Using Data Intrinsic Characteristics.* Information Sciences, 250, 113–141.

[7] Austroads. (2015). *Road Safety Engineering Risk Assessment Part 6: Crash Prediction Models.*AustroadsPublication..

[8] Peden,M.,(2004). *World Report on Road Traffic Injury Prevention.* World Health Organization.

[9] National Highway Traffic Safety Administration (NHTSA). (2018). *Traffic Safety Facts Annual Report.* U.S. Department of Transportation.

[10] Yan, X., Radwan, E., & Abdel-Aty, M. (2005). *Characteristics of Rear-End Accidents at Signalized Intersections Using Multiple Logistic Regression Model.* Accident Analysis & Prevention, 37(6), 983–995.

[11] Kockelman, K. M., & Kweon, Y. J. (2002). *Driver Injury Severity: An Application of Ordered Probit Models.* Accident Analysis & Prevention, 34(3), 313–321.

[12] Lahoti, N., & Huang, B. (2015). *Predicting Injury Severity of Motor Vehicle Accidents Using Machine Learning Techniques.* Transportation Research Record, 2513(1), 9–17.

[13] Santos, A., & Shah, A. (2004). *An Overview of U.S. Road Safety Trends.* Journal of Safety Research, 35(2), 141–145..

[14]El-Basyouny, K.,&Sayed,T.(2006).Comparison of Two Negative Binomial Regression Techniques in Developing Accident Prediction Models. Transportation Research Record, 1950(1), 9–16.

[15] Hossain, M., &Muromachi, Y. (2012). A Bayesian Network-Based Framework for Real-Time Crash Prediction on the Freeway. IEEE Transactions on Intelligent Transportation Systems, 13(2), 914–927.

[16] Washington,S.P., Karlaftis, M. G., & Mannering, F. L. (2010). Statistical and Econometric Methods for Transportation Data Analysis. CRC Press.

[17] Hosseinpour, M., Yahaya, A. S., &Norghani, M. H. (2013). Accident Prediction Models for Urban Roads Using Statistical Methods. International Journal of Crashworthiness, 18(1), 37–48.

[18] Shankar, V., Milton, J., & Mannering, F. (1997). Modeling Accident Frequencies as Zero-Altered Probability Processes: An Empirical Inquiry. Accident Analysis & Prevention, 29(6), 829–837.