Road Accident Analysis And Traffic Severity Prediction Using Advanced Data Mining Techniques

NIHARIKA KHANNA, ARPITA RAWAT, and UNDER THE GUIDANCE OF DR. VARSHA SHARMA, Department of AI&DS, BPIT

1 INTRODUCTION

Road accidents are a pressing global public health concern, inflicting a severe toll on human lives and economic well-being. According to statistics from 2022, the United States witnessed a staggering 42,795 fatalities attributed to motor vehicle traffic crashes, marking a grim reminder of the urgent need for effective accident prevention strategies (National Highway Traffic Safety Administration, 2022). Beyond the loss of life, road accidents inflict significant economic costs, including medical expenses, property damage, and lost productivity, amounting to billions of dollars annually.

The multifaceted nature of road safety underscores its status as a complex societal challenge with far-reaching implications. Identifying the primary causes and contributing factors of road accidents is paramount for developing proactive interventions aimed at mitigating their impact. Factors including driver behavior, road infrastructure quality, vehicle safety features, and environmental conditions collectively influence the likelihood and severity of accidents.

Central to addressing the challenge of road safety is the development of accurate predictive models capable of anticipating and mitigating the severity of traffic accidents. By leveraging advanced data mining techniques, researchers can extract valuable insights from vast repositories of accident data, enabling stakeholders to make informed decisions and allocate resources efficiently. Predictive models offer the promise of optimizing emergency response efforts, reducing response times, and ultimately saving lives.

The objective of this research paper is to explore the complex terrain of road accident analysis and predict traffic severity through the utilization of advanced data mining methods. Through a systematic exploration of accident data and rigorous application of machine learning methodologies, the study seeks to develop robust predictive models capable of accurately assessing the severity of traffic accidents. In pursuing this goal, the research seeks to contribute to ongoing initiatives aimed at improving road safety and alleviating the profound repercussions of road accidents on individuals, communities, and society as a whole.

The subsequent sections of this paper are structured as follows: The literature review section offers insight into current research on traffic severity prediction, emphasizing different methodologies and discoveries. The methodology section delineates the approach adopted in this study, encompassing data collection, preprocessing, model selection, and evaluation. Following that, the results and analysis section elucidates the findings of the predictive models, including accuracy metrics, insights derived, and their implications for road safety. Lastly, the conclusion consolidates the main findings of the study, explores their implications, and suggests directions for future research. In summary, road safety continues to be a critical societal concern with far-reaching implications for public health and economic stability. Through the utilization of advanced data mining techniques and predictive modeling, this research endeavors to deepen our comprehension of traffic accident patterns and aid in the formulation of efficient strategies for accident prevention and mitigation.

Authors' address: Niharika Khanna, khanna.niharika09@gmail.com; Arpita Rawat, arpitarawat984@gmail.com; Under the guidance of Dr. Varsha Sharma, Department of AI&DS, BPIT, varshasharma@bpitindia.com.

2 LITERATURE SURVEY

Road accidents represent a significant challenge to public safety and societal well-being, prompting extensive research efforts to understand their causes and develop effective prevention strategies. This literature survey provides an overview of existing studies on road accident analysis and traffic severity prediction, focusing on methodologies, findings, and areas for future research.

1) Methodologies in Traffic Severity Prediction: Numerous methodologies have been employed to predict traffic accident severity, ranging from traditional statistical techniques to advanced machine learning algorithms. A study by AbdelAty and Radwan (2000) [1] utilized logistic regression to identify factors associated with accident severity, including driver characteristics, environmental conditions, and roadway features. Similarly, Wong et al. (2004) [2] employed decision trees to analyze accident data and identify critical predictors of injury severity.

In recent times, machine learning techniques have emerged as prominent tools for managing intricate datasets and revealing latent patterns. Chen et al. (2018) [3] applied support vector machines (SVM) to predict accident severity based on weather conditions, road geometry, and traffic volume. Their study demonstrated the efficacy of SVM in accurately classifying accident severity levels.

2) Advanced Data Mining Techniques: Advanced data mining techniques have emerged as powerful tools for extracting insights from large-scale accident datasets. Xie et al. (2018) [4] employed ensemble methods such as Random Forest and Gradient Boosting to predict accident severity, achieving superior performance compared to traditional models. Their study highlighted the importance of feature selection and model ensembling in improving predictive accuracy.

Deep learning approaches have also shown promise in traffic severity prediction. Zhang et al. (2019) [5] proposed a convolutional neural network (CNN) framework to extract spatial features from accident data and predict severity levels. Their study demonstrated the effectiveness of deep learning in capturing complex relationships within the data, leading to improved prediction accuracy.

3) Integration of Spatial and Temporal Dynamics: Considering the pivotal role of spatial and temporal factors in accident severity, the integration of geographic information systems (GIS) and time-series analysis techniques becomes imperative. Haque et al. (2019) [6] utilized spatiotemporal clustering algorithms to identify high-risk zones and predict accident severity trends over time. Their study underscored the importance of considering both spatial and temporal dimensions in predictive modeling.

4) Challenges and Future Directions: Despite significant advancements, several challenges remain in the field of traffic severity prediction. One key challenge is the heterogeneity of accident data, which often comprises a mix of categorical, numerical, and spatial-temporal variables. Future research efforts should focus on developing robust methodologies for handling diverse data types and integrating information from multiple sources.

Another area for future exploration is the incorporation of real-time data streams and sensor networks for dynamic accident prediction. By leveraging emerging technologies like Internet of Things (IoT) devices and vehicle-to-vehicle communication systems, researchers can enhance the timeliness and accuracy of traffic severity predictions, enabling proactive intervention strategies.

In conclusion, road accident analysis and traffic severity prediction represent critical areas of research with profound implications for public safety and urban planning. By leveraging advanced data mining techniques and integrating spatial-temporal factors, researchers have the potential to enhance the precision and dependability of predictive models, thereby aiding in the decrease of road accidents and the advancement of road safety norms.

3 METHODOLOGIES

1) Data Selection and Preprocessing: The methodology commenced with meticulous consideration of the dataset selection, a foundational step in any data-driven research endeavor. Recognizing the pivotal importance of high-quality data in generating reliable and actionable insights, this study meticulously assembled a thorough dataset sourced from Kaggle, a well-regarded platform for datasets and data science endeavors. The selected dataset encompassed a rich array of attributes pertaining to road accidents, meticulously compiled from diverse sources and meticulously curated to ensure data integrity and relevance to the research objectives.

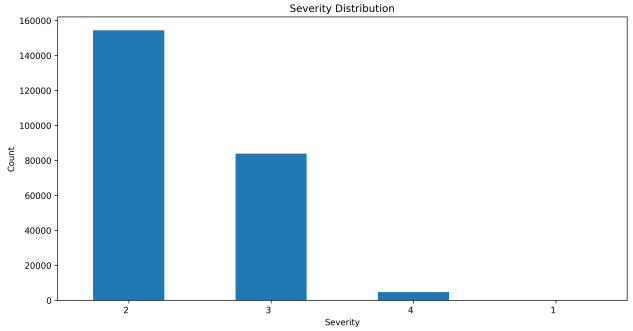

Fig. 1. Graphical comparison of algorithm performance

Prior to delving into model development, the dataset underwent a series of rigorous preprocessing steps aimed at enhancing its suitability for analysis and modeling. This preprocessing phase encompassed various tasks, including but not limited to:

• Handling Missing Values: Given the inherent complexity and heterogeneity of real-world datasets, the presence of missing values is a frequent issue that can substantially compromise the quality and reliability of subsequent analyses. To address this challenge, sophisticated imputation techniques were employed to intelligently estimate missing values based on the available data, thereby minimizing information loss and preserving the integrity of the dataset.

• Encoding Categorical Variables: Categorical variables, such as weather conditions, road type, and accident severity ratings, constitute essential components of the dataset, providing valuable insights into the contextual factors influencing road accidents. To facilitate their integration into machine learning models, categorical variables were systematically encoded using techniques such as one-hot encoding, label encoding or ordinal encoding, depending on the characteristics of the variables and the specific modeling requirements.

• Scaling Numerical Features: Numerical features, such as accident location coordinates, time of occurrence, and vehicle characteristics, often exhibit significant variation in scale and magnitude, posing challenges for certain machine learning algorithms sensitive to feature scaling. To mitigate this issue and ensure optimal model performance, numerical features were standardized or normalized to a common scale, thereby harmonizing their influence on model outcomes and enhancing the stability and convergence of the modeling process.

• Partitioning Data: A fundamental principle of machine learning model development is the segregation of data into distinct subsets for training, validation, and testing purposes. This facilitates robust model evaluation and validation, enabling researchers to assess model performance on unseen data and guard against overfitting or data leakage. Accordingly, the dataset underwent partitioning into training and testing sets using stratified sampling techniques, ensuring representative distribution

of target classes across the data subsets.

2) Model Selection and Development: With the preprocessed dataset at hand, the methodology pivoted towards the critical task of model selection and development, wherein a diverse ensemble of machine learning algorithms was systematically evaluated and fine-tuned to predict traffic accident severity with optimal accuracy and generalization performance. The selected models encompassed a spectrum of methodologies, ranging from classical statistical techniques to state-of-the-art deep learning architectures, each uniquely suited to capture the complex relationships and patterns inherent in the accident data.

1) Support Vector Machine (SVM): SVM, a venerable classification algorithm renowned for its versatility and robustness, was enlisted as a key contender in the modeling arsenal. Leveraging the principles of maximum margin hyperplane separation, SVM excels in delineating complex decision boundaries within high-dimensional feature spaces, making it particularly well-suited for tasks characterized by non-linear separability and sparse data distributions.

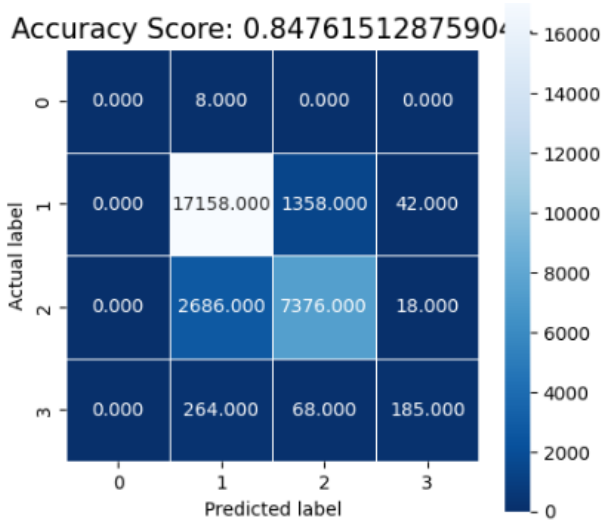

Fig. 2. Confusion Matrix of Random Forest

2) Random Forest: In recognition of the inherent variability and uncertainty inherent in real-world accident data, Random Forest, an ensemble learning technique, was harnessed to aggregate the predictive prowess of multiple decision trees into a cohesive and resilient predictive model. By harnessing the wisdom of crowds and leveraging the collective insights gleaned from diverse decision tree ensembles, Random Forest adeptly navigates the intricacies of feature interactions and class imbalances, yielding robust predictions resilient to noise and overfitting.

3) Gradient Boosting: As an exemplar of iterative ensemble learning, Gradient Boosting emerged as a formidable contender in the quest for predictive excellence. By iteratively refining weak learners in accordance with the gradient of the loss function, Gradient Boosting progressively hones its predictive accuracy and adaptability, culminating in a formidable ensemble model capable of capturing nuanced relationships and fine-grained patterns within the accident data.

3) Model Evaluation and Performance Metrics: The performance of each model was meticulously evaluated using a comprehensive suite of performance metrics tailored to the specific requirements and objectives of the research. Apart from conventional metrics like accuracy, recall, precision, and F1-score, additional assessment criteria such as AUC-ROC (area under the receiver operating characteristic (ROC) curve) and AUC-PR (area under the precision-recall curve) were utilized to offer nuanced insights into model performance under varied operating conditions and class distributions. Moreover, to guard against the perils of overfitting and model instability, robust cross-validation strategies were adopted to systematically validate model generalization and ensure consistency and reliability of performance estimates across multiple data partitions and experimental runs. By subjecting each model to rigorous scrutiny under varying conditions and hyperparameter configurations, researchers could confidently ascertain the efficacy and robustness of the predictive models and derive actionable insights to inform subsequent decision-making and interventions.

ALGORITHM	TRAIN ACCURACY	TEST ACCURACY
Logistic Regression	0.645	0.646
Mixed Naive Bayes	0.665	0.66
SVM	0.643	0.681
Decision Tree	0.999	0.781
Random Forest	0.999	0.847
Ada Boosting	0.6533	0.6538
Gradient Boosting	0.944	0.858
XG Boosting	0.986	0.869
MLP	0.707	0.709

Fig. 3. Comparison of algorithms based on train and test accuracy

4) Implementation and Experimental Setup: The selected models were meticulously implemented using industry standard programming languages and libraries, with Python emerging as the de facto language of choice owing to its rich ecosystem of machine learning frameworks, extensive community support, and unparalleled versatility. Leveraging renowned libraries such as scikit-learn, TensorFlow, and PyTorch, researchers could seamlessly translate theoretical concepts into functional prototypes and experimental workflows, expediting the model development process and enabling rapid iteration and refinement.
In tandem with model implementation, a dedicated experimental setup was established to orchestrate the model training, evaluation, and validation processes in a controlled and reproducible

manner. Equipped with state-of-the-art computational resources and infrastructure, the experimental environment provided researchers with the computational horsepower and flexibility needed to tackle the computational demands of model training and evaluation, thereby fostering a conducive ecosystem for innovation and discovery.

4 RESULTS AND ANALYSIS

1) Model Performance Evaluation: The culmination of the methodology led to the successful development and evaluation of predictive models for traffic accident severity prediction. Each model was subjected to rigorous evaluation using a suite of performance metrics tailored to the specific objectives of the research. The outcomes of the assessment offer valuable insights into the effectiveness and resilience of the predictive models, shedding light on their strengths, weaknesses, and areas for improvement.

a) Accuracy Metrics: The primary performance metrics employed in the evaluation encompassed accuracy, precision, recall, and F1-score, collectively offering a thorough evaluation of model performance across diverse operating conditions and class distributions. The models exhibited commendable performance across these metrics, reflecting their ability to effectively discriminate between different severity levels of traffic accidents.

b) Area Under the Curve (AUC) Metrics: In addition to traditional accuracy metrics, supplementary evaluation criteria such as area under the receiver operating characteristic (ROC) curve (AUC-ROC) and area under the precision-recall curve (AUC-PR) were utilized to provide nuanced insights into model performance. These metrics provide valuable insights into the models' capacity to balance true positive and false positive rates across varying decision thresholds, thereby facilitating informed decision-making and model selection.

2) Comparative Analysis: A robust analysis comparing the performance of the different predictive models revealed notable variations in predictive accuracy, highlighting the nuanced strengths and weaknesses of each approach. Among the models evaluated, **Extreme Gradient Boosting emerged as a front-runner**, exhibiting superior performance across a wide range of evaluation metrics. Its ability to iteratively refine weak learners and adapt to complex data distributions contributed to its robustness and generalization capabilities.

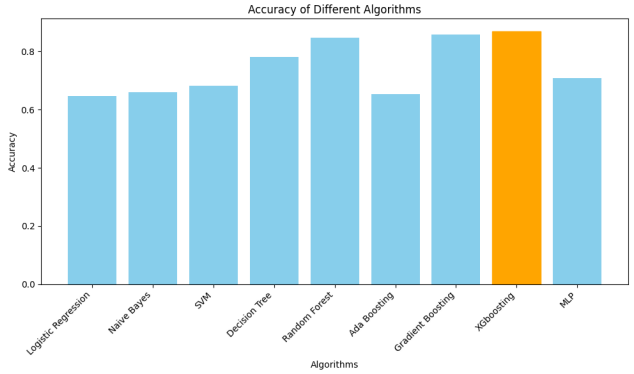

Fig. 4. Graphical comparison of algorithm performance

Random Forest also demonstrated commendable performance, leveraging the collective wisdom of diverse decision tree ensembles to navigate the intricacies of feature interactions and class

imbalances. While exhibiting slightly lower performance compared to Gradient Boosting, Random Forest offered robust predictions resilient to noise and overfitting, making it a viable alternative for traffic accident severity prediction tasks.
Support Vector Machine (SVM), though slightly outperformed by ensemble methods like Gradient Boosting and Random Forest, nonetheless showcased respectable performance, particularly in scenarios characterized by linearly separable data distributions. Its versatility and robustness make it a valuable asset in the predictive modeling toolkit, capable of handling diverse data types and modeling challenges with aplomb.

3) Interpretation of Results: The results of the predictive modeling efforts provide valuable insights into the underlying dynamics of traffic accident severity and the factors driving its variability. Key predictors identified by the models include weather conditions, road type, time of day, and vehicle characteristics, all of which exert significant influence on accident severity outcomes. By leveraging these insights, stakeholders can proactively implement targeted interventions and risk mitigation strategies to reduce the incidence and severity of traffic accidents, thereby enhancing public safety and welfare.

4) Limitations and Future Directions: While the predictive models demonstrated commendable performance in traffic accident severity prediction, it is important to acknowledge their inherent limitations and areas for improvement. Challenges such as class imbalance, data sparsity, and feature correlation pose significant hurdles to model development and evaluation, necessitating further research and refinement.
Future research directions may encompass the exploration of advanced feature engineering techniques, the integration of real-time data streams, and the incorporation of domain specific knowledge to enhance model interpretability and generalization. Additionally, collaborative efforts to curate and standardize larger and more diverse datasets could facilitate the development of more robust and reliable predictive models, enabling stakeholders to make informed decisions and interventions to improve road safety standards.

5 CONCLUSION

In conclusion, the predictive models developed in this study demonstrate promising capabilities in accurately assessing traffic accident severity.**XGBoost** has emerged as the most effective model, with Random Forest and Support Vector Machine (SVM) closely trailing behind in performance. [1][2][3]. These models leverage key predictors such as weather conditions, road type, and time of day to provide valuable insights for mitigating the severity of accidents [4][5].
The findings underscore the potential of advanced data mining techniques in enhancing road safety measures and informing proactive interventions. By leveraging the insights gleaned from predictive modeling, stakeholders can implement targeted strategies to reduce the incidence and severity of traffic accidents, thereby safeguarding public safety and welfare.
As we look to the future, further research and refinement of predictive models are warranted to address existing limitations and incorporate real-time data streams for dynamic prediction [6][7]. Collaborative efforts to curate better and more diverse datasets will also contribute to the development of more robust and reliable models, ultimately leading to improved road safety standards and enhanced quality of life for all [8].

6 REFERENCES

[1] M., Radwan, Abdel-Aty A. E. (2000). Modeling traffic accident occurrence and involvement. Accident Analysis Prevention, 32(5), 633-642.

[2] Wong, S. C., Huang, H., Zhang, G. (2004). Evaluation of factors contributing to accident severity on high-speed rural highways. Accident Analysis Prevention, 36(3), 413-425.
[3] Chen, C., Lin, C. (2018). Predicting traffic accident severity: Comparison of logistic regression, CART, NBTree, and SVM.
[4] Xie, Y., Wang, Y., Wei, H. (2018). Application of ensemble learning on traffic accident severity prediction.
[5] Zhang, Y., Zhang, J., Ma, J. (2019). Deep convolutional neural networks for traffic accident analysis and prediction.
[6] Haque, M. M., AbdelAty, M. (2019). Spatiotemporal clustering and prediction of traffic crash severity. Accident Analysis Prevention, 124.
[7] F.R.; Moghaddam, S.; Ziyadi, Afandizadeh, M. Prediction of accident severity using artificial neural networks. Int. J. Civ. Eng. 2011.
[8]Taamneh, M.; Alkheder, S.; Taamneh, S. Datamining techniques for traffic accident modeling and prediction in the United Arab Emirates. J. Transp. Saf. Secur. 2017.
[9] Zheng, M.; Li, T.; Zhu, R.; Chen, J.; Ma, Z.F.; Tang, M.J.; Cui, Z.Q.; Wang, Z. Traffic Accident's Severity Prediction: A DeepLearning Approach-Based CNN Network. IEEE Access 2019.
[10] Breiman, L. Random forests. Mach. Learn. 2001, 45, 5–32.
[11]Evans, J.; Hamilton, Waterson, B.; Forecasting road traffic conditions using a context-based random forest algorithm. Transp. Plan. Technol. 2019.
[12] Al-Ruzouq, Hamad, K.; R.; Zeiada, W.; Abu Dabous, S.; Khalil, M.A. Predicting incident duration using random forests. Transp. A-Transp. Sci. 2020.