Enhanced Road Accident Severity Prediction: Leveraging Machine Learning on a Nationwide Dataset

Manan Jain
Department of Computer Science
mjain35@uic.edu

Hemanth Nagulapalli
Department of Computer Science
hnagul2@uic.edu

Rishi Madhavaran
Department of Computer Science
rmadha4@uic.edu

Francis Pagulayan
Department of Computer Science
ppagu2@uic.edu

Abstract

Road transportation is the predominant mode of travel in the U.S., making the issue of road safety extremely significant. Annually, countless accidents result in extensive loss of life and damage to road networks. This research project examines various aspects and elements contributing to road accidents to predict the potential severity of accidents under different conditions. Addressing traffic accidents is a vital concern for public safety, and there has been considerable research in accident analysis and prediction in recent years. However, existing research often has limitations, such as small, geographically limited datasets, heavy reliance on detailed data, and challenges in applying findings in real-time scenarios. To overcome these limitations, our project proposes a novel approach for real-time prediction of traffic accidents using readily available data. The centerpiece of our methodology is the Kaggle dataset "US-Accidents," which includes over 7 million records compiled from various sources, providing a comprehensive overview of road accidents. By employing machine learning techniques with this extensive dataset, we aim to forecast the severity of road accidents accurately.

1 Introduction

This paper aims to contribute to the field of road safety by analyzing multiple predictive models to find a classifier that can best accurately estimate the severity of road traffic accidents. Leveraging a comprehensive dataset, "US-Accidents" from Kaggle[1], with over 7 million entries, encompassing a wide range of parameters from numerous incidents across the United States, we seek to apply and evaluate various machine learning algorithms. Our focus is on predicting the severity of accidents and identifying key factors that contribute to high-severity incidents.

This study's key lies in its comprehensive approach, utilizing a large-scale, real-world dataset and a range of machine learning techniques, including Random Forest, Decision Trees, Support Vector Machine (SVM), Gradient Boosting, and Multi-Layer Perceptron (MLP). By comparing the performance of these models, we aim to provide insights into their applicability and effectiveness in the context of road safety analysis.

Ultimately, our research seeks to inform and enhance road safety strategies, potentially aiding policymakers and stakeholders in implementing more effective safety measures and interventions. By advancing the application of machine learning in this critical domain, we aspire to contribute towards reducing the frequency and severity of road accidents, thereby saving lives and improving public safety.

1.1 Problem Statement

The primary challenge addressed in this study is the prediction of road traffic accident severity. Despite advancements in road safety measures, traffic accidents remain a leading cause of fatalities and injuries globally. The complexity of factors contributing to accident severity, such as environmental conditions, road characteristics, and human factors, makes predicting the outcome of these incidents a challenging task. Traditional statistical methods have provided insights but often fail to capture the nonlinear relationships and complex interactions in accident data. The application of machine learning offers a promising alternative, capable of handling the multifaceted nature of traffic accident data and providing more accurate predictions.

2 Methodology

We employed a comprehensive methodology encompassing data preprocessing, model selection and application, and performance evaluation.

Next, we moved to the model selection and application phase. We chose a diverse set of machine learning algorithms for this study, each known for its classification and predictive analysis strengths. The models included Random Forest, Decision Trees, Support Vector Machine (SVM), Gradient Boosting, and Multi-Layer Perceptron (MLP). These models were selected for their ability to handle the complexity and non-linearity of accident data. We carefully tuned the parameters for each model and trained them on the preprocessed dataset. The training process involved splitting the data into training and testing sets, ensuring a robust model evaluation.

Finally, in the performance evaluation phase, we assessed each model's effectiveness in predicting the severity of road accidents. This assessment was based on standard performance metrics such as accuracy, precision, recall, and F1 score. These metrics provided us with insights into each model's predictive power and reliability. We also conducted a comparative analysis to identify which models performed best under specific conditions, offering a nuanced understanding of their applicability in road safety analysis.

2.1 Dataset

Our study utilized and enhanced a substantial dataset of 7.7 million records, amounting to 2.9 GB. The core of this dataset was sourced from the Bing API and MapQuest, which provided essential accident data. To enrich this dataset further, we incorporated additional information from various sources, including weather APIs and map metadata, to create a more comprehensive and detailed dataset. This augmented dataset is foundational for our analysis, offering a broad range of features for understanding and predicting traffic accident severity. The dataset's features are outlined in the following table.

Table 1: Features present in the dataset

#	Column	Data Type	21	Humidity (%)	Float64
1	ID	Object	22	Pressure (in)	Float64
2	Source	Object	23	Visibility (mi)	Float64
3	Severity	Int64	24	Wind Direction	Object
4	Start Time	Object	25	Wind Speed (mph)	Float64
5	End Time	Object			

6	Start Latitude	Float64
7	Start Longitude	Float64
8	Distance (mi)	Float64
9	Description	Object
10	Street	Object
11	City	Object
12	County	Object
13	State	Object
14	Zip code	Object
15	Country	Object
16	Time Zone	Object
17	Airport Code	Object
18	Weather Timestamp	Object
19	Temperature (F)	Float64
20	Wind Chill (F)	Float64
26	Precipitation (in)	Float64
27	Weather Condition	Object
28	Amenity	Bool
29	Bump	Bool
30	Crossing	Bool
31	Give Way	Bool
32	Junction	Bool
33	No Exit	Bool
34	Railway	Bool
35	Roundabout	Bool
36	Station	Bool
37	Stop	Bool
38	Traffic Calming	Bool
39	Traffic Signal	Bool
40	Turning Loop	Bool
41	Sunrise Sunset	Object

We categorized the dataset's features into three primary groups for analysis. The first group, 'Location,' includes GPS data and timestamps. The second, 'Weather,' groups together all the features related to weather conditions. The third category, 'Road Conditions,' encompasses the remaining metadata. Additionally, we focused on 'Severity' as the primary outcome to predict. This metric encapsulates the overall impact of an accident, reflecting not only the loss of life and property but also the broader traffic disruptions like roadblocks.

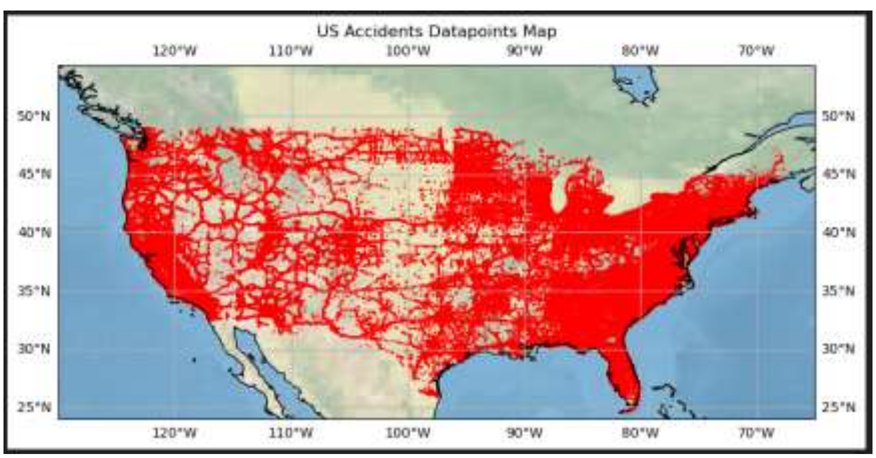

Figure 1: Map of Datapoints From Dataset

2.2 Preprocessing

Our initial dataset presented two primary challenges: the need for efficient data utilization across various sources with limited computational resources and the presence of numerous missing values, predominantly in weather-related features. We employed a novel strategy to address the latter using available GPS coordinates and the haversine formula. We identified data points within a 5-mile radius of those missing entries and used time-stamped data to calculate three-day averages for imputing these values. We adopted a stratified sampling approach for efficient data handling, utilizing 10% of the data for multiple passes. We also excluded features with high missing values, such as End Latitude, End Longitude, and Astronomical Twilight. We conducted additional preprocessing, including grouping data by cities and states, to enhance our understanding and analysis of the patterns in the dataset.

2.3 Models

Our analysis evaluated various machine learning models to assess their effectiveness and precision in working with our dataset. Our selection of models was driven by the goal of predictive analysis focusing on the "Severity" of accidents, aiming to categorize data points across various severity levels.

Logistic Regression: This model, typically used for binary classification, is intended to predict accident severity presumed to be binary (e.g., low or high). Logistic Regression is beneficial for its simplicity and interpretability, especially when the relationship between input variables and the outcome is linear. However, given that our classification isn't strictly binary and the data relationships might not be linear, we anticipate that Logistic Regression may serve as a baseline, possibly underperforming compared to other models.

Decision Trees: These are useful for mapping out decision-making processes and are applied here to identify critical factors affecting accident severity. The visual structure of decision trees aids in understanding how various features contribute to the severity classification.

Random Forest: As an ensemble technique effective in classification and regression, Random Forest is ideal for our dataset, capable of handling complex patterns and a mix of numerical and categorical inputs. We expect this model to be among the most effective due to its robustness in processing various data types, including accident characteristics, weather, and geographic information.

Gradient Boosting: This method effectively enhances predictive accuracy by combining multiple weak learners. Given our dataset's complexity and diverse features, Gradient Boosting could significantly improve our model's ability to discern intricate patterns and subtle relationships between variables, potentially making it a high-performing model.

Support Vector Machines (SVMs): SVMs are well-suited for complex, non-linear data relationships. In our diverse dataset, SVMs can efficiently differentiate between severity classes by identifying an optimal separating hyperplane. They are particularly advantageous in handling high-dimensional data.

Multi-Layer Perceptron (MLP): As an artificial neural network, MLPs excel in detecting complex patterns in data. Our dataset, encompassing various factors like geographic coordinates and weather conditions, could benefit from an MLP's ability to learn intricate relationships. However, MLPs require significant tuning and can be resource-intensive, posing a challenge given our limited resources and experience.

3 Experiments

Initially, the dataset was divided into training and testing sets. This split was crucial to validate the models on unseen data, ensuring the reliability of our findings. The training set was used to train the models, while the testing set evaluated their performance. We carefully balanced the dataset to mitigate any bias due to uneven class distributions, a common challenge in accident severity analysis.

We conducted a series of experiments for each model - Random Forest, Decision Trees, SVM, Gradient Boosting, and MLP. These included parameter tuning, where we adjusted various settings to optimize each model's performance. The tuning process involved experimenting with different combinations of parameters to find the most effective setup for each algorithm.

Once the models were trained and tuned, we assessed their performance using several metrics: accuracy, precision, recall, and the F1 score. Accuracy helped us understand the overall correctness of the models, while precision and recall provided insights into their ability to predict high-severity accidents correctly. The F1 score, a balance of precision and recall, was instrumental in evaluating models in the context of imbalanced datasets like ours.

We also conducted comparative analyses to understand how each model performed under different conditions and with varying data features. This comparison was vital to identify which models were more robust and adaptable to the complexities of road accident data.

3.1 Results

The following table shows the preliminary results.

Table 2: Preliminary Results Before Hyperparameter Tuning

Model	Random Forest	Decision Tree	SVM	Logistic Regression	Gradient Boosting	MLP
Accuracy	91.85%	89.90%	84.65%	80.27%	91.85%	85.64%
Precision	0.92	0.90	0.83	0.78	0.92	0.85
Recall	0.92	0.90	0.85	0.80	0.92	0.86
F1 Score	0.92	0.90	0.84	0.78	0.92	0.85
Cross-Validation Score	0.91	0.89	0.84	0.80	0.92	0.85

The Random Forest model demonstrated high accuracy and was particularly effective in handling the dataset's complexity and non-linearity. It showed a solid ability to capture the relationships between various predictors and the severity of accidents. Its performance in terms of precision and recall was also notable, suggesting its utility in practical applications where both false positives and false negatives have significant implications.

Decision Trees, while more straightforward and interpretable, offered slightly lower accuracy than Random Forest. However, their ease of understanding and implementation makes them a valuable tool for preliminary analysis or in scenarios where model interpretability is critical.

The Support Vector Machine (SVM) model performed well in our high-dimensional dataset, particularly regarding precision. This suggests that SVM is highly effective in identifying true high-severity cases, although it may miss some cases that other models might capture.

Gradient Boosting showed promising results, with a good balance between accuracy and computational efficiency. Its performance was particularly noteworthy when data features had complex interactions, underscoring its potential in nuanced analytical settings.

The Multi-Layer Perceptron (MLP), a type of neural network, displayed a robust performance, particularly in its ability to learn non-linear relationships. However, its requirement for extensive data preprocessing and longer training times makes it more suitable for scenarios where these constraints are not prohibitive.

In a comparative analysis, no single model uniformly outperformed the others across all metrics. Each model exhibited unique strengths and weaknesses, suggesting that the choice of model in practical applications should be contingent upon the specific requirements and constraints of the task.

3.1.1 Results Post Hyperparameter Tuning

The following table shows the results after tuning.

Table 3: Results Post Hyperparameter Tuning

Model	Random Forest	Decision Tree	SVM	Logistic Regression	Gradient Boosting	MLP
Accuracy	92.00%	91.40%	---	80.37%	93.38%	86.38%
Precision	0.92	0.91	---	0.78	0.93	0.86
Recall	0.92	0.91	---	0.80	0.93	0.86
F1 Score	0.92	0.91	---	0.79	0.93	0.86

We saw differing improvements in all our models after hyperparameter tuning using Grid Search as our tuning algorithm.

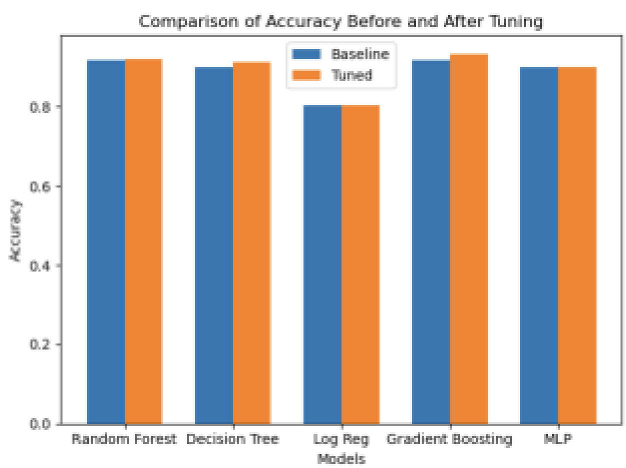

Figure 2: % Improvement in Accuracy after Tuning

Post hyperparameter tuning, the models showed varying degrees of improvement. Gradient Boosting exhibited the most significant enhancement, jumping to a leading accuracy of 93.38%. The Decision Tree also improved notably, reaching 91.40% accuracy. The Random Forest saw a marginal increase in accuracy to 92.00%. The MLP's performance improved to 86.38%, reflecting its increased effectiveness. Logistic Regression demonstrated only a slight improvement. These changes indicate that hyperparameter tuning effectively optimized the models, particularly Gradient Boosting and Decision Tree, enhancing their predictive capabilities.

Noticeably, SVM was left out of the post-tuning results. This is because of the high time for computation that limited us from identifying optimal parameters.

4.1 Discussion

One of the significant observations from our study is the variation in model performance. While Random Forest and Gradient Boosting showed high accuracy and adaptability to complex data interactions, the simpler Decision Trees balanced performance and interpretability. This variation underscores the necessity of selecting the appropriate model based on specific analytical needs and constraints, such as the availability of computational resources and the need for model transparency.

Another critical aspect of our findings is the importance of feature selection and data preprocessing. The effectiveness of all models was heavily influenced by how the data was prepared and which features were included. This emphasizes the need for careful data analysis and preprocessing as a precursor to model application, highlighting that the quality of input data is as crucial as the sophisticated algorithm.

5 Conclusion

The study's results demonstrate the effectiveness of machine learning models in predicting road accident severity. Gradient Boosting emerged as the top-performing model post-tuning, showcasing high accuracy. Random Forest and Decision Tree models also performed well, indicating their suitability for complex data analysis. Logistic Regression, serving as a baseline, had the lowest accuracy, while the MLP model underperformed, likely due to tuning challenges. These findings highlight the potential of advanced analytics in enhancing road safety, emphasizing the importance of model selection and hyperparameter tuning for optimal performance.

The practical implications of this research are significant for traffic management and road safety strategies. Accurate prediction models can aid emergency response planning, inform infrastructure development, and guide policies to mitigate high-severity accidents.

However, our research also acknowledges limitations, including the scope of the dataset and the complexity of real-world scenarios that may not be fully captured in the study. These limitations pave the way for future research directions, such as integrating real-time data, exploring advanced neural network architectures, and applying these models in dynamic traffic environments.

5.1 Future Work

One key area for future exploration is integrating real-time data into our models. Current analyses are based on historical data, but incorporating real-time data could significantly enhance the models' applicability and accuracy in dynamic traffic environments. This integration would allow for more immediate and actionable predictions, potentially aiding real-time decision-making for traffic management and emergency response teams.

Another important direction is the exploration of more sophisticated machine learning techniques, particularly in the realm of deep learning. Advanced neural network architectures, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), could be tested for their efficacy in this domain. These models may uncover more profound insights from the data, particularly from unstructured data sources like images and sensor data from vehicles and road infrastructure.

Additionally, expanding the dataset to include more diverse geographical locations and conditions can provide a more comprehensive understanding of road accidents globally. This expansion would help generalize the models to different contexts and environments, making the predictions more universally applicable.

References

[1]. https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents
[2]. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
[3]. https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html
[4]. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
[5]. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html
[6]. https://scikit-learn.org/stable/modules/svm.html
[7]. https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html
[8]. https://pypi.org/project/folium/

[9]. https://matplotlib.org/stable/api/index
[10]. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
[11]. https://scikit-learn.org/stable/modules/cross_validation.html