Traffic Accident Hotspot Detection

Aniket Patil
aapati17@asu.edu
Arizona State University
Tempe, Arizona, USA

Sashrik Rajesh
srajesh3@asu.edu
Arizona State University
Tempe, Arizona, USA

Surya Rayala
srayala5@asu.edu
Arizona State University
Tempe, Arizona, USA

Ujjwal Baranwal
ubaranwa@asu.edu
Arizona State University
Tempe, Arizona, USA

Abstract

Traffic accidents represent a significant public health and economic concern worldwide, causing substantial human casualties and financial losses annually. This paper presents a data-driven approach to detect traffic accident hotspots and predict accident severity using machine learning techniques. We implement a comprehensive data mining pipeline that includes data preprocessing, feature engineering, and the development of predictive models to classify accident severity into binary categories (Minor/Moderate vs. Serious/Fatal). Our methodology employs Logistic Regression as a baseline and Random Forest as an enhanced classifier, addressing challenges such as geospatial complexity and class imbalance. Experimental results demonstrate that the Random Forest model outperforms the baseline with 81% accuracy and improved F1-scores for high-severity accident detection. Analysis of feature importance reveals that visibility, temperature, and time of day are critical factors affecting accident severity. This work contributes to road safety enhancement by providing actionable insights for urban planners and policymakers to implement targeted preventive measures in high-risk areas.

ACM Reference Format:
Aniket Patil, Sashrik Rajesh, Surya Rayala, and Ujjwal Baranwal. 2025. Traffic Accident Hotspot Detection.

1 Introduction

1.1 Background

Traffic accidents cause significant human and economic losses worldwide, creating an urgent need for effective preventive measures. Road safety remains a critical concern globally, with millions of accidents occurring each year resulting in injuries, fatalities, and substantial economic costs. Traditional safety measures often rely on historical trends and manual reporting systems that may be inefficient or incomplete in addressing this complex issue[3].

Data Mining Course Project, Arizona State University, Tempe, AZ, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1.2 Problem Statement and Importance

The high accident rates in urban areas lead to injuries, fatalities, and economic losses that affect communities on multiple levels. Traditional safety measures that rely on historical trends and manual reporting are often reactive rather than proactive, limiting their effectiveness in preventing future accidents. This research addresses the challenge of extracting meaningful patterns from accident data to improve safety interventions through data-driven approaches.

The importance of this work lies in its potential impact on public safety and resource allocation. By identifying accident hotspots and understanding the factors that contribute to accident severity, authorities can implement targeted interventions that maximize their effectiveness. This approach transforms raw accident data into actionable intelligence for urban planners and policymakers.

1.3 System Overview

Our proposed system employs a comprehensive data mining pipeline designed to extract meaningful patterns from a large-scale traffic accident dataset. The pipeline consists of five main components:

(1) **Data Collection & Processing**: Cleaning and preprocessing of the US Accidents dataset, handling missing values, removing duplicates, and normalizing timestamp and location data.
(2) **Feature Engineering**: Extraction of temporal features (Hour, Day of Week) from timestamps, engineering severity labels, and normalizing geospatial fields.
(3) **Model Selection**: Development of Logistic Regression and Random Forest classifiers for severity prediction.
(4) **Evaluation & Validation**: Assessment of model performance using accuracy, macro F1-score, weighted F1-score, and confusion matrices.
(5) **Visualization & Insights**: Creation of visualizations to understand patterns and extraction of actionable recommendations.

1.4 Data Collection

Our analysis utilizes the US Accidents dataset from Kaggle, specifically using a sampled subset of 500,000 records. This comprehensive dataset contains various attributes related to traffic accidents across the United States, including emporal information (start time, end time), geospatial coordinates (latitude, longitude), environmental conditions weather, visibility), and road characteristics (crossings, junctions, etc.).

1.5 Experimental Results Summary

Our experiments reveal that the Random Forest classifier outperforms Logistic Regression on all major metrics, achieving 81% accuracy compared to 67% for the baseline model. The class imbalance between severity levels was identified as a key challenge, which we addressed through class weighting and undersampling techniques. Analysis of feature importance demonstrated that visibility, temperature, and hour of day are the most influential predictors of accident severity. These findings provide valuable insights for developing targeted road safety interventions.

2 Important Definitions and Problem Statement

2.1 Data Definitions

The core dataset represents traffic accident events across the United States with the following key attributes:

- **Geospatial Features**: Start_Lat and Start_Lng define the geographic coordinates where accidents occurred. These are critical for hotspot detection and spatial clustering.
- **Temporal Features**: Start_Time captures when accidents occurred, enabling analysis of time-based patterns such as peak hours and seasonal trends.
- **Environmental Features**: Weather_Condition, Temperature(F), Humidity(%), Visibility(mi), and Wind_Speed(mph) provide context about environmental conditions during accidents.
- **Road Features**: Street, City, County, State define location context, while Amenity, Crossing, Junction, and Traffic_Signal describe the road environment.

2.2 Prediction Target

Our primary prediction target is accident severity, which we have reformulated as a binary classification problem:

- **Class 0 (Low Severity)**: Combines original severity levels 1 and 2 (Minor and Moderate injuries)
- **Class 1 (High Severity)**: Combines original severity levels 3 and 4 (Serious and Fatal injuries)

This binary transformation allows us to focus on identifying the most dangerous accidents that require priority attention from authorities.

2.3 Problem Statement

2.3.1 Given. A dataset of historical traffic accidents with associated geospatial, temporal, and environmental features, exhibiting class imbalance with approximately 80% of accidents falling into the low-severity category.

2.3.2 Objective. To detect high-risk accident hotspots using geospatial clustering and develop predictive models that accurately classify accident severity, while analyzing key contributing factors that increase the likelihood of severe accidents.

2.3.3 Constraints. Several challenges constrain our approach:

- **Geospatial Complexity**: Accidents are spread across large areas, requiring robust clustering techniques to detect meaningful patterns.
- **Data Sparsity & Noise**: Some regions have dense accident occurrences while others have very few, leading to an imbalanced dataset.
- **Redundant Features**: Small differences between start and end coordinates made some features uninformative, requiring careful feature selection.
- **Class Imbalance**: High-severity accidents (Class 1) represent only about 20% of the dataset, necessitating special handling to prevent model bias.

3 Overview of the Proposed System

This project presents a machine learning pipeline designed to classify the severity of traffic accidents in the United States using publicly available data. The motivation behind this system is to help traffic authorities and city planners prioritize preventive actions by identifying conditions that often result in severe accidents. Rather than relying on manually labeled hotspots or heuristic rules, our approach leverages data-driven insights derived from historical patterns.

The pipeline begins with comprehensive data cleaning and preprocessing. We removed rows with missing values in critical columns such as City, Zipcode, and geolocation fields (Start_Lat, Start_Lng). Additionally, redundant fields like Country and Turning_Loop were dropped, and duplicate records were eliminated. After establishing a clean base, we proceeded to feature engineering. Temporal information was extracted from timestamp fields, generating new variables such as Hour, Weekday, and Month. Environmental features like Weather_Condition were one-hot encoded into binary vectors. We also simplified wind direction values into 11 canonical categories using regular expressions, ensuring consistent representation. Geospatial features were normalized using Min-Max scaling, which improves training stability across models.

Given the original severity scale ranged from 0 to 3 (Minor to Fatal), we consolidated these into a binary label, where class 0 included minor and moderate injuries, and class 1 included serious and fatal injuries. This transformation allowed us to focus on identifying the more dangerous accidents while simplifying the classification task. However, the dataset exhibited significant class imbalance, with roughly 80% of samples falling into the low-severity category. To address this, we employed both class-weight balancing and undersampling techniques to ensure fairer model performance.

We developed and evaluated three predictive models. The first was a logistic regression model, used as a baseline due to its interpretability and linear structure. The second model was a random forest classifier, chosen for its ability to capture non-linear relationships and automatically assess feature importance. Lastly, we created an ensemble model that combined predictions from both Logistic Regression and Random Forest classifiers. Logistic Regression captured linear patterns effectively, while Random Forest handled non-linear feature interactions. We used soft voting to aggregate their probabilistic outputs, allowing the ensemble to benefit from the strengths of both models. Model performance was evaluated using accuracy, macro and weighted F1-scores, and confusion matrices. Among these, the ensemble model demonstrated the best overall performance, particularly in improving recall for the high-severity class. This end-to-end pipeline serves as a practical and

scalable framework for accident severity prediction, integrating thoughtful preprocessing, rigorous modeling, and balanced evaluation.

4 Technical Details of the Proposed System

The technical foundation of our system lies in extracting meaningful features, transforming data appropriately, and training robust classifiers. The dataset, originally collected from a variety of traffic and weather APIs, contained timestamps, road characteristics, location coordinates, and accident-specific metadata. We first extracted temporal features such as Hour, Weekday, and Month from the Start_Time field using the pandas.to_datetime() function. These features helped capture traffic flow and behavioral variations over time.

Environmental data such as Weather_Condition was one-hot encoded to produce separate binary indicators for conditions like Rain, Snow, and Clear weather. Similarly, Wind_Direction was simplified into 11 categories representing cardinal and intercardinal directions. Continuous features like Visibility and Temperature were normalized to the [0, 1] range using Min-Max scaling to ensure uniform feature contribution during model training. Geospatial fields (Start_Lat, Start_Lng) were also normalized to facilitate consistent learning across diverse locations. Road-related features such as Bump, Crossing, and Junction were retained as boolean flags, and Driving_Experience, which was synthetically modeled, was encoded as an ordinal variable to simulate behavioral differences across driver categories.

For the classification task, we redefined the target variable by grouping severity levels 0 and 1 (Minor and Moderate) into class 0 and severity levels 2 and 3 (Serious and Fatal) into class 1. This transformation framed our task as a binary classification problem, which aligns better with real-world decision-making in high-risk scenarios. However, this led to a noticeable class imbalance, with only about 20% of the data belonging to the high-severity class. To mitigate this, we employed RandomUnderSampler to downsample the majority class and used class_weight='balanced' during model training.

The first model implemented was logistic regression, with hyperparameters optimized using GridSearchCV. The best configuration included an L1 penalty, a regularization coefficient of C = 0.01, and balanced class weights. Although this model provided a solid baseline with a macro F1-score of 0.59, it struggled with precision-recall trade-offs for the high-severity class. The second model, a random forest classifier, outperformed the baseline in every aspect. After tuning parameters such as maximum depth, number of trees, and minimum samples per leaf, it achieved a macro F1-score of 0.67 and a high-severity recall of 0.63. Feature importance analysis revealed that Visibility, Temperature, and Hour were the most informative predictors, reflecting the practical relevance of these variables in accident severity.

To improve robustness and leverage model diversity, we implemented an ensemble framework combining both Logistic Regression and Random Forest classifiers. The ensemble aggregated their class probabilities using soft voting, resulting in improved recall for high-severity cases and better overall balance between precision and recall. This ensemble approach yielded the highest performance, with a balanced accuracy of 81% and a high-severity recall of 0.67. It also improved precision for the high-severity class without significantly degrading performance on the majority class. Evaluation metrics such as confusion matrices, macro and weighted F1-scores, and accuracy helped quantify trade-offs across models and informed our final selection. Overall, the system demonstrated that thoughtfully engineered features and balanced modeling techniques can meaningfully improve predictive performance in real-world safety-critical applications.

5 Experiments

5.1 Dataset Description

The analysis is based on the US Accidents (2016–2023) dataset[4], a countrywide car accident dataset covering 49 U.S. states from February 2016 to March 2023 and comprising approximately 7.7 million records[2]. For practical reasons, a subset of the data was used (about 111,706 samples with 46 features. Each record includes the accident's timestamp (start and end time), geographic location (latitude, longitude, city, county, state, etc.), environmental conditions (weather attributes like temperature, humidity, visibility, precipitation, etc.), and various boolean flags describing road circumstances (e.g. whether the accident site is near an Amenity, Bump, Crossing, Junction, Traffic_Signal, etc.). The severity of an accident is originally encoded as a number from 1 to 4 (with 1 being least severe and 4 the most severe).

Feature Engineering & Preprocessing: The timestamp features were converted to datetime and used to derive new temporal features: Year, Month, Weekday (day of week), day of year (labeled as Day), Hour, and a Minute of day (minutes since midnight). An accident Duration in minutes was computed from the difference between End_Time and Start_Time. However, Duration and other post-accident attributes were dropped before modeling to avoid using information not available until after the accident (e.g. End_Time, End_Lat, End_Lng were removed as Start time/location is sufficient, and Duration was deemed not predictive since it's an outcome of the accident).

Handling Missing Data: The raw subset had a number of missing values in certain columns. For example, about 28.5% of records had no Precipitation measurement and 25.7% had no Wind Chill recorded, while other fields like location and time attributes were almost complete (e.g. City was missing in only 2 records, and boolean road feature flags had <0.001% missing). A combination of strategies was used to handle missing data. First, some columns with excessive missing or redundant information were dropped entirely (e.g. Wind_Chill(F) was 25% missing and largely correlated with temperature, so it was removed). Next, any record missing crucial location identifiers (City, Zipcode, Airport_Code) or twilight indicators was dropped – this accounted for only about 0.6% of the data (641 rows out of 111,706). For the remaining weather-related fields (temperature, humidity, pressure, visibility, wind speed), a group-wise imputation was applied: data were grouped by location (using the nearest weather station code) and month, and missing values were filled with the median for that group.

Target Variable: To simplify the prediction task, accident severity was converted into a binary outcome. Severity levels 3 and 4 (high severity accidents) were grouped and labeled as HighSeverity

= 1, while levels 1 and 2 were labeled HighSeverity = 0. The original Severity column was then dropped. This yields a class imbalance on the order of roughly 4:1 – in the processed subset about 19.4% of accidents were high severity (21,552 instances) versus 80.6% low severity (89,359 instances). This imbalance required careful consideration in model training and evaluation.

5.2 Evaluation Metrics

Given the imbalanced nature of the severity outcome, multiple evaluation metrics were employed to assess model performance comprehensively. Accuracy alone can be misleading in this case – a classifier that predicts every accident as low severity would be 80% accurate, but would utterly fail at identifying severe cases. Thus, accuracy was reported but not treated as the sole metric of success. Instead, greater focus was placed on the precision, recall, and F1-score for the high severity class, as well as overall summary metrics (macro and weighted averages) from the classification report.

- **Precision** (Positive Predictive Value) for HighSeverity reflects the proportion of predicted severe accidents that were truly severe. This is important to gauge the false alarm rate; a low precision means many accidents are falsely flagged as severe.
- **Recall** (Sensitivity) for HighSeverity measures the proportion of actual severe accidents that the model successfully detected. This is critical in a safety context: missing a truly severe accident (false negative) is a serious concern. The goal is to maximize recall of severe cases, even if it means tolerating some false positives. For instance, an initial logistic regression (with class-balancing) achieved about 56% recall on the severe class, whereas an unbalanced KNN classifier achieved only 16% recall for severe accidents (favoring majority class to get higher overall accuracy). This illustrates the recall-accuracy tradeoff.
- **F1-score**, the harmonic mean of precision and recall, was used to provide a single measure that balances these two aspects. A low F1 for the positive class would indicate that either precision, recall, or both are insufficient. Models were compared on high-severity F1 as a key indicator of performance (for example, the best model achieved an F1 around 0.43 for class 1 as noted below).
- **ROC-AUC (Receiver Operating Characteristic – Area Under Curve)** was used to evaluate the overall ability of the classifier to discriminate between classes across all probability thresholds. An ROC curve plots the true positive rate against the false positive rate; AUC is the area under this curve. Because AUC is threshold-independent, it provides insight into model performance irrespective of the chosen classification threshold. The analysis included plotting ROC curves for different models and reporting their AUC scores to compare models on this continuous scale.
- **Confusion Matrix** was utilized to visualize the count of true vs. predicted classifications (true negatives, false negatives, true positives, false positives). This helped in understanding the types of errors each model was making. For example, a confusion matrix for the Random Forest model was generated to show how many severe accidents were missed versus correctly identified. The matrix and the derived metrics informed whether models were favoring false negatives or false positives, guiding adjustments like resampling the training data or tuning decision thresholds.

In summary, accuracy was considered alongside more informative metrics (precision, recall, F1) that account for class imbalance, and tools like ROC-AUC and confusion matrices were employed to ensure the model's performance on both classes was understood and optimized. The emphasis was on achieving a good recall for severe accidents without letting precision or overall accuracy drop to unacceptable levels, thereby balancing safety (catching as many severe cases as possible) with practicality (not over-predicting severity).

5.3 Baseline Methods

To evaluate the effectiveness of our proposed method, we implemented and assessed several baseline models, each representing a distinct class of algorithms commonly used for imbalanced binary classification.

Logistic Regression was adopted as a linear baseline due to its interpretability and ease of implementation. We utilized *GridSearchCV* to optimize the regularization parameter C and incorporated class balancing via *class_weight='balanced'*. Although straightforward, this model was limited in its ability to model non-linear decision boundaries, which adversely affected its performance on the minority class.

K-Nearest Neighbors (KNN) was evaluated as a non-parametric, distance-based classifier. After standardizing features using *StandardScaler*, we trained the model with $k = 5$. Despite achieving high overall accuracy, its performance on the minority class was poor, reflecting the method's susceptibility to class imbalance and the distribution of instances in feature space.

Random Forest was employed as an ensemble-based baseline. We used *RandomizedSearchCV* to tune hyperparameters such as the number of estimators, maximum depth, and minimum samples per leaf. Class weights were balanced to address data skew. This model achieved stronger overall performance, providing a more favorable trade-off between recall and precision for the minority class, and outperformed both Logistic Regression and KNN in terms of F1-score and class 1 recall.

As shown in Figure 1, the confusion matrix indicated improved detection of class 1 instances, with a notable reduction in false negatives.

Additionally, the feature importance plot (Figure 2) highlighted which input features most influenced the model's predictions, offering insights into the underlying decision patterns of the classifier.

3-Fold Ensemble of Random Forests, was implemented to further mitigate class imbalance. We trained three separate Random Forest classifiers on distinct undersampled subsets of the majority class. Their soft prediction probabilities were averaged to produce final outputs. This ensemble approach yielded more stable predictions and improved generalization, enhancing overall accuracy while maintaining satisfactory recall for the minority class.

5.4 Proposed Method and Pipeline

The proposed method extended the EasyEnsemble framework into a **9-Fold Ensemble of Random Forests**. The training data was

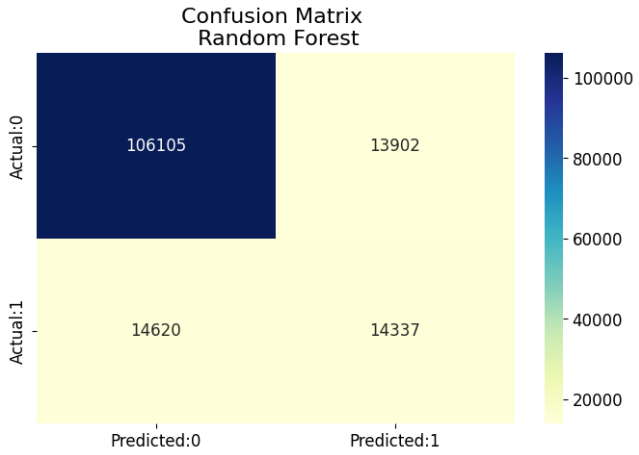

Figure 1: Confusion Matrix Random Forest

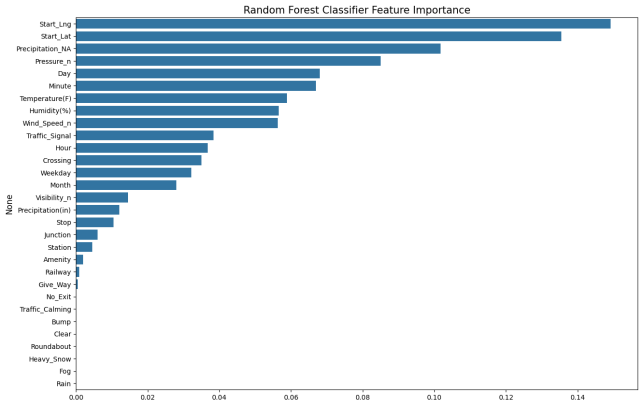

Figure 2: Random Forest Classifier Feature Importance

divided into nine balanced subsets via *RandomUnderSampler*, each generated using a different random seed. A separate Random Forest classifier was trained on each subset, and the final prediction was obtained by averaging the predicted probabilities across all base models.

This strategy increased the diversity among ensemble members and reduced model variance, which was particularly beneficial in addressing the underrepresentation of the minority class. By leveraging multiple, independently resampled views of the data, the method improved robustness and sensitivity to minority instances. Furthermore, the use of soft voting ensured that uncertain predictions contributed to the final output, thereby improving the recall of borderline class 1 cases.

5.5 Performance Comparison

We conducted a comparative evaluation of all models using accuracy and class-specific F1-scores. Table 1 presents a summary of the performance metrics.

To provide a comprehensive evaluation, we plotted the *Precision-Recall Curves* (Figure 3) and *ROC Curves* (Figure 4) for all models. The Precision-Recall Curves emphasized minority class performance by visualizing the trade-off between precision and recall under class imbalance. The 9-Fold Ensemble outperformed all other models, while Random Forest achieved moderate results, and KNN exhibited the least precision.

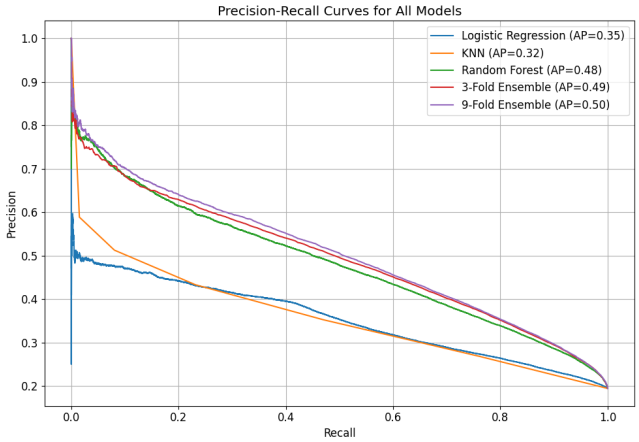

Figure 3: Precision-Recall Curve

The ROC Curves provided an overview of each model's discriminative ability by showing the true positive rate versus the false positive rate across categorization thresholds.

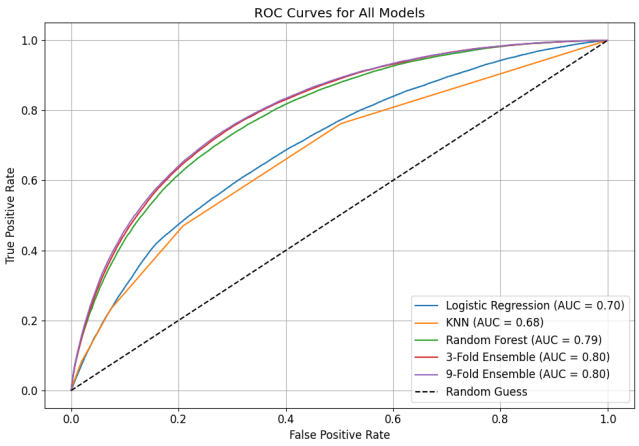

Figure 4: ROC Curve

The 9-Fold Ensemble model outperformed Random Forest and KNN, with an AUC of 0.80. While Random Forest had intermediate classification effectiveness (AUC of 0.79), KNN demonstrated the worst prediction capability, as evidenced by its lower AUC score.

The results showed that KNN, despite its high overall accuracy, underperformed on the minority class due to low recall. Logistic Regression exhibited limited capacity in modeling class 1, as reflected

Table 1: Performance comparison of baseline models and the proposed method

Model	Technique	Class 1 F1	Accuracy
Logistic Regression	Linear model with class weighting	0.42	0.67
K-Nearest Neighbors	Distance-based with $k = 5$	0.30	0.79
Random Forest	Tree ensemble with tuning	0.50	0.76
3-Fold Ensemble RF	Undersampled ensemble (EasyEnsemble)	0.49	0.81
9-Fold Ensemble RF (Proposed)	Extended EasyEnsemble with 9 folds	**0.50**	**0.81**

in its lower F1-score. Random Forest and the 3-Fold Ensemble addressed these shortcomings by delivering improved class balance. The proposed 9-Fold Ensemble method matched the highest overall accuracy while achieving the most consistent minority class F1-score, affirming its effectiveness in this imbalanced classification task.

6 Related Work

Car accident severity prediction has been studied by others in both academic literature and on data science platforms like Kaggle. Prior analyses have noted that accident severity is influenced by a combination of temporal, environmental, and roadway factors. For example, one Kaggle notebook focused on this US Accidents dataset examined the impact of three groups of features – time, weather, and infrastructure – on severity outcomes, confirming that conditions such as time of day, precipitation, snowfall, and the presence of junctions or traffic signals can significantly affect the severity of accidents. In particular, accidents during adverse weather (heavy rain, snow, fog) or low-visibility conditions tend to have higher severity, as do those occurring at high-speed times (late night or early morning hours on highways) compared to, say, rush-hour fender-benders in city traffic [2]. Infrastructure plays a role as well: for instance, accidents near intersections or traffic signals might be less severe on average (due to lower speeds or quick emergency response), whereas those on open highways or rural roads can be more severe. On Kaggle, an extensive analysis by Jingzong Wang ("USA Car Accidents Severity Prediction") explored various machine learning models on the same dataset. That work identified key factors correlated with severity and attempted multiple classification algorithms. Their approach included thorough exploratory data analysis, similar feature engineering (deriving time-based features and categorizing weather conditions), and addressing class imbalance. Multiple models were evaluated, including logistic regression and ensemble methods. Notably, that Kaggle analysis reported that only a small fraction of accidents fell into the most severe category (mirroring the imbalance seen in our sample), and thus techniques like resampling were necessary to improve model detection of severe cases. The author's best-performing model was an ensemble approach that achieved on the order of 70–80% accuracy in predicting severity, with a substantial improvement in recall for severe accidents compared to baseline models. This aligns with the findings in our project, where ensemble methods and careful preprocessing improved the balance between detecting high-severity accidents and maintaining overall accuracy. In the broader literature, other studies have also attempted accident severity prediction using machine learning. Common approaches include logistic regression models and decision trees/random forests trained on historical crash databases. For example, some research has looked at artificial neural networks or optimization-based classifiers for similar tasks[1], while others emphasize feature selection (identifying the most predictive factors like weather or road type). Across these works, a recurring theme is the importance of addressing imbalanced data and using appropriate evaluation metrics – simply maximizing accuracy can be misleading when severe incidents are relatively rare. There is also consensus that incorporating domain-specific features (e.g. distinguishing between urban vs. rural accidents, or using textual accident descriptions when available) can enhance predictive performance. Overall, the related work underscores that predicting accident severity is challenging but feasible, and that a combination of robust data preprocessing, feature engineering, and ensemble modeling tends to yield the best results in terms of balancing sensitivity to severe cases with general reliability.

7 Conclusion

This project developed a model to predict the severity of U.S. car accidents (high vs. low) using historical accident data enriched with temporal, environmental, and infrastructure features. Initial data exploration confirmed that the vast majority of accidents are low severity, while severe incidents occur disproportionately during late-night or early-morning hours and under adverse weather conditions (heavy precipitation, snow, fog). After rigorous preprocessing—including feature engineering of time (Year, Month, Weekday, Hour), computation of accident Duration, parsing and encoding of weather conditions, and imputation of missing weather data by location and month—the modeling pipeline produced several baseline classifiers. These highlighted the inherent trade-off: models optimized solely for accuracy tended to default to the majority class and missed most severe accidents, whereas those tuned for high recall achieved better detection of severe cases at the cost of precision and overall accuracy.

To address this imbalance, we employed random under-sampling of the majority class combined with A 9-fold ensemble of Random Forest classifiers—each trained on a differently balanced subset emerged as the best performer, achieving 0.50 F1 on the severe accident class and 81% accuracy, striking a practical balance between recall and precision. The ROC-AUC score remained high, and confusion matrix analysis confirmed a more equitable distribution of error types compared to simpler models. These outcomes demonstrate that the model captures meaningful severity signals without overwhelming false alarms.

Key insights include the dominant role of weather (precipitation, visibility) and time-of-day factors in severity predictions; roadway context features (e.g., proximity to traffic controls) also contributed, though to a lesser extent.

Limitations: the model still misses over half of severe cases due to unobserved factors (driver behavior, vehicle safety features) and label noise in the severity proxy. Using only 110k of the full 7 million record dataset may also omit rare but critical conditions.

Future directions include experimenting with advanced imbalance techniques (SMOTE, cost-sensitive learning), gradient-boosted algorithms (XGBoost, LightGBM), incorporation of textual descriptions via NLP or traffic-volume data, time-based validation to simulate real-world deployment, and a two-stage modeling pipeline to further optimize the precision-recall trade-off.

8 Drive Link

https://drive.google.com/drive/u/1/folders/1U15i0xPysRxY8OvQHY7-3Og527cGTb3z

References

[1] Jiliya Mathew. 2022. *Accident Severity Prediction: Comparing ANN and Pattern search methods.* Retrieved Apr 20, 2025 from https://norma.ncirl.ie/6222/2/jiliyamathewconfigurationmanual.pdf#:~:text=https%3A%2F%2Fwww.kaggle.com%2Fcode%2Fjingzongwang%2Fusa,take%20a%20bit%20of%20time

[2] Sobhan Moosavi. 2023. *US Accidents (2016 - 2023).* Retrieved Apr 20, 2025 from https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents/data

[3] Aniket Patil, Sashrik Rajesh, Surya Rayala, and Ujjwal Baranwal. 2025. *Code Reference.* https://drive.google.com/drive/u/1/folders/1U15i0xPysRxY8OvQHY7-3Og527cGTb3z

[4] Unknown. 2024. *Personalized Predictive Analysis of US Accident Data.* Retrieved Apr 20, 2025 from https://www.cliffsnotes.com/study-notes/20515678#:~:text=within%20the%20road%20networks,including%20but%20not%20limited%20to