US Accident Severity Prediction

PAML - Team 6

1st Henry Bao
Cornell Tech
Information Science
New York, USA
hb482@cornell.edu

2nd Stella Hong
Cornell Tech
Information Science
New York, USA
sh2577@cornell.edu

3rd Mingze Gao
Cornell Tech
Computer Science
New York, USA
mg2454@cornell.edu

4th Yunhao Hu
Cornell Tech
Information Science
New York, USA
yh2233@cornell.edu

***Abstract*—We're addressing the prediction of the severity of traffic accidents across the U.S., a crucial task for enhancing road safety and informing traffic management.**

This problem is important because accurate predictions can help reduce accidents through preventive measures.

Our approach involves using machine learning algorithms like Decision Trees, Random Forests, and Naive Bayes, incorporating road conditions, weather conditions, and temporal data to train models that predict traffic accident severity and embed this feature into a web application.

Prior work involved data analysis and visualization, whereas our approach not only includes data visualizations but also builds a real-time accident severity prediction feature by implementing machine learning methods such as Decision Trees, Random Forests, and Naive Bayes.

We're using a comprehensive Kaggle dataset called "US Accidents (2016 - 2023)" covering 49 states from 2016 to 2023, which includes road conditions, weather conditions, and other accident details in numerical and categorical data.

The front end has four pages: "Home," "Data Exploration," "Data Visualization," and "Model Prediction." The Home page gives an overview of the project. The Data Exploration page provides details about the dataset. The Data Visualization page offers interactive graphs such as line charts, bar charts, and maps. The Model Prediction page allows users to select from three methods and enter the location, time, and driving environment to predict accident severity. For the backend, we pretrained three models and saved them as PKL files. We used Streamlit to write the code for the front end and used the PKL files to connect the front end and backend models.

To evaluate our models, we'll use cross-validation, confusion matrices, and ROC-AUC curves, with metrics like accuracy, precision, recall, and F1 score.

Our expected findings will demonstrate the models' effectiveness in predicting accident severity.

This project can significantly impact the machine learning field by highlighting the value of real-time data integration and advanced algorithms, potentially leading to better safety measures and resource allocation.

I. Introduction

Motivation: We are tackling the problem of predicting traffic accident severity across the United States. This is important because understanding the factors leading to accident severity can help reduce the severity rate through informed interventions. We aim to develop a predictive model that forecasts accident severity under various conditions and locations. This application will be useful for city planners, traffic management authorities, and drivers by providing valuable insights and enhancing road safety measures.

Technical focus: We are using machine learning methods such as Decision Trees, Random Forests, and Naive Bayes. These methods are well-suited for handling large datasets and providing results with relatively high accuracy. Our approach is unique because it incorporates road condition data, weather data, and temporal features to predict accidents, which is not extensively covered in existing literature. This allows us to offer more accurate and comprehensive predictions compared to traditional models.

Prior Work: Prior work by Muhammad Bilal Khan "Geospatial Insights- US Car Accidents" focused on visualizing US traffic accident data and identifying hotspots using tools like Folium and Marker Cluster for interactive maps. This approach effectively highlighted accident-prone areas but was limited to data visualization. Our project goes beyond this by using machine learning algorithms like Decision Trees, Random Forests, and Naive Bayes to predict accident severity. We integrate road conditions, weather, and temporal data, offering real-time severity predictions through a web application. While prior work emphasized data visualization, our approach not only provided data visualization but also added predictive analytics, providing actionable insights for improving road safety and traffic management.

Machine Learning Pipeline: Our machine learning pipeline started with extensive data exploration to understand patterns and features. We then preprocessed the data by handling missing values, encoding categorical variables, normalizing numerical data, and engineering new features. Model training involved tuning hyperparameters and using cross-validation to avoid overfitting. We evaluated the models using metrics like precision, recall, and F1 score. Finally, the model was deployed in a web application that provides real-time accident severity predictions. We expected our models to effectively predict accident severity and offer actionable insights for traffic management and road safety.

Impact: The application can significantly enhance road safety and inform drivers about potential risks. However, there are ethical concerns regarding data privacy and the potential for biased predictions. We will address these by anonymizing data and ensuring the model does not produce biased results based on location or time. The application aims to enhance public safety and support better urban planning and traffic management practices while maintaining human oversight in decision-making processes.

II. Background

Numerous studies have addressed the prediction and analysis of traffic accidents, each employing various techniques and focusing on different aspects.

Muhammad Bilal Khan's work "Geospatial Insights- US Car Accidents" concentrated on visualizing US traffic accident data and identifying hotspots using tools like Folium and Marker Cluster. This approach was effective in highlighting accident-prone areas but limited to data visualization [1].

Another relevant study by Moosavi et al. (2019) "A Countrywide Traffic Accident Dataset" introduced a countrywide traffic accident dataset and used machine learning models to predict accident risk based on heterogeneous sparse data. Their focus was on integrating diverse data sources to enhance prediction accuracy [2].

Zhou (2020) "ML to Predict Accident Severity PA Mont" explored the use of machine learning to predict accident severity, employing basic statistical methods and straightforward machine learning models. Their research provided insights into the factors influencing accident severity but did not incorporate real-time data or advanced algorithms [3].

Our proposed work builds on these foundations by using machine learning algorithms such as Decision Trees, Random Forests, and Naive Bayes to predict accident severity. Unlike previous studies, our approach integrates road conditions, weather data, and temporal features to provide real-time predictions through a web application. This combination of predictive analytics and real-time data integration offers a more proactive solution for improving road safety.

Based on prior work, we anticipate that our approach will provide more accurate and comprehensive predictions compared to traditional models. By incorporating advanced algorithms and diverse data sources, we aim to offer actionable insights that can significantly enhance traffic management and road safety.

III. End-to-End ML Pipeline

A. Data Collection, Exploration, and Processing

Our team trained our machine learning models using a comprehensive car accident dataset from Kaggle, which spans 49 states in the USA and covers the period from February 2016 to March 2023[4]. This dataset was compiled using APIs that collect real-time traffic incident data from transportation departments, law enforcement, and traffic cameras. The key attributes include accident severity, geographical coordinates, weather conditions, and a textual description of each incident. This dataset is rich in both categorical and numerical data types, including timestamps that indicate the duration of each accident.

For data exploration, our initial goal was to gain an understanding of how various features correlated with accidents. First, we began by examining a series of graphs that displayed different aspects of the accident data. The first line graph presented the number of accidents over time, revealing a noticeable upward trend in the frequency of accidents year by year. This trend was clear and consistent until the data cut-off point in March 2023, which resulted in a drop at the end of the graph due to incomplete data for that year. The second scatter plot graph offered a spatial representation of accident locations, with all accidents marked in blue and accidents with the highest severity highlighted in red in contrast. This graph allowed us to find patterns in the severity and frequency of accidents across different geographic areas (e.g., the most severe cases of accidents are more likely to happen in dense metro areas). Finally, in the third bar chart graph, we analyzed accidents by the time of day. This graph showed significant variations in the number of accidents occurring at different hours. More specifically, there was a peak in accidents during the early morning and late afternoon hours, suggesting that these periods might be associated with higher traffic volumes or more hazardous driving conditions.

In preparation for applying machine learning algorithms, we did several thorough preprocessing steps. First, we addressed issues such as missing weather-related details and location coordinates, by imputing by average/zero or omitting them based on their impact on our analysis. Then, we performed feature encoding and converted categorical variables like weather conditions and wind direction into numerical formats through one-hot and integer encoding. Finally, we removed outliers to maintain model accuracy, reduced dimensionality to simplify the model, and engineered new features, such as months and years as they might reveal trends related to accident frequency and severity throughout different times of the year or week.

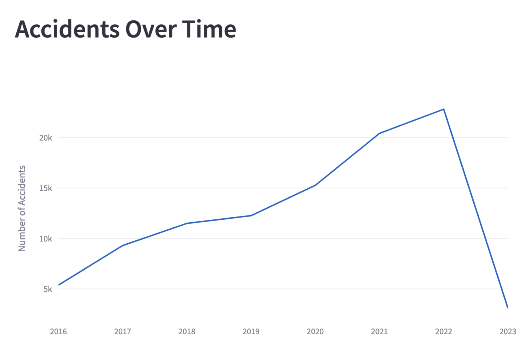

Fig. 1. A line graph that shows the accidents that occurred over time from February 2016 to March 2023

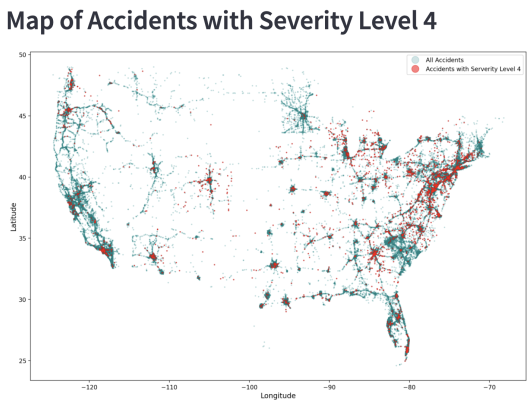

Fig. 2. A scatter plot showing all accidents happened marked in blue compared to the accidents with the highest severity marked in red

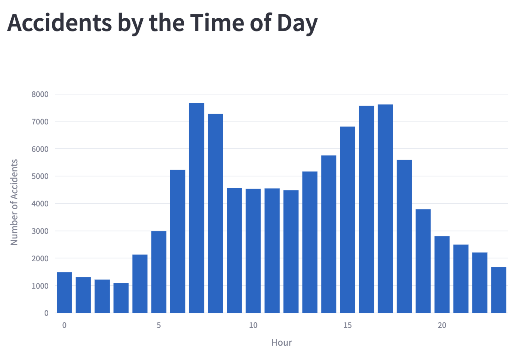

Fig. 3. A bar chart that shows the number of accidents based on the time of the day

B. Methods and Model Training

We used three machine learning techniques: Decision Trees, Random Forests, and Naive Bayes. Each of these methods is well-suited for different aspects of our problem of predicting accident severity. Decision Trees are excellent for their simplicity and ability to handle both numerical and categorical data. Random Forests, being an ensemble method, combine multiple decision trees to improve prediction accuracy and reduce overfitting. Naive Bayes is particularly effective for classification tasks and works well with categorical data, making it a good fit for our problem where we have several categorical inputs.

Our models were trained using a variety of inputs that included both categorical and numerical data. The categorical inputs included factors like location and weather conditions, while the numerical inputs included time and temperature. The output from our models was the predicted accident severity, which ranged from 1 to 4. We assessed the models by testing how accurately they could predict the accident severity. To train and validate our models, we used an 80-20 train-test split, with a random state of 43 to ensure reproducibility.

Each model had specific parameter settings to optimize its performance. For the Random Forest model, we set `n_estimators` to 100, `max_depth` to None, `min_samples_split` to 2, and `min_samples_leaf` to 1. This configuration allows the model to build a large number of trees and grow them fully, capturing complex patterns in the data. The Decision Tree model had a `max_depth` of 10, which provided a balance between model complexity and interoperability. For the Naive Bayes model, we used the default priors and a `var_smoothing` of 1×10^{-9}, which helps to stabilize the model's calculations.

Initially, our models encountered underfitting due to the high skewness in our dataset towards severity levels 2 and 3. We addressed this by underfitting the target feature, essentially rebalancing the dataset to provide a more even distribution of severity levels. This adjustment resulted in more balanced prediction results, reducing the likelihood of the model consistently predicting the majority classes and ignoring the minority classes.

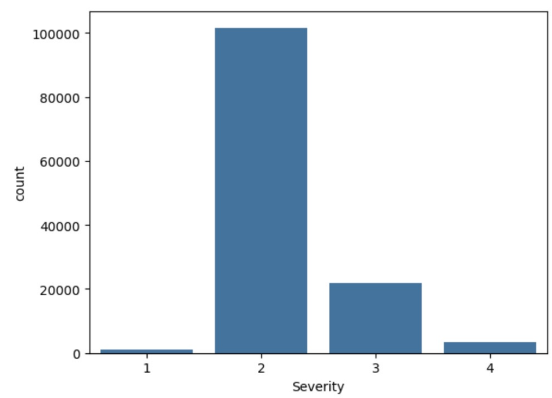

Fig. 4. Initial distribution of accident severity in the dataset, showing a high skewness towards severity 2 and 3.

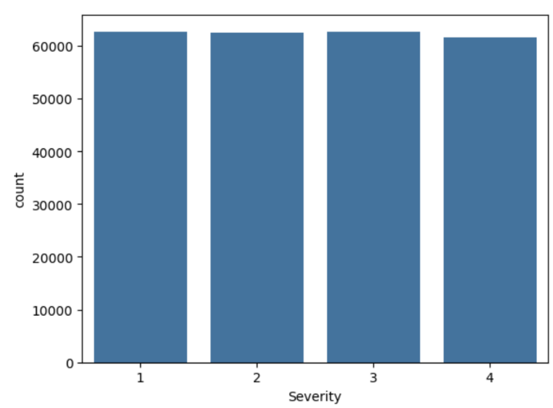

Fig. 5. Balanced distribution of accident severity in the dataset after undersampling the target feature.

C. Model Evaluation

For this project, three machine learning models were evaluated: Random Forest, Decision Tree, and Naive Bayes. Each model was subjected to a rigorous evaluation process to determine their performance and identify any errors. The models were trained and tested using a standard train/test split and further validated with 5-fold cross-validation to ensure robustness and reduce the risk of overfitting.

a) Evaluation Methods: To compare the models effectively, the following evaluation methods were employed:

- Confusion Matrix: Provides a detailed breakdown of true positives, false positives, true negatives, and false negatives.
- 5-Fold Cross Validation: Used to validate the models and ensure their robustness by averaging the results of five different subsets of the training data.
- Train/Test Split: The dataset was split into training (80%) and testing (20%) sets to evaluate the models on unseen data.

b) Metrics: The evaluation metrics used to assess the models were:

- Accuracy: The ratio of correctly predicted instances to the total instances.
- F1-Score: The harmonic mean of precision and recall, giving a balance between the two.

c) Hyperparameter settings:

- Random Forest: Number of trees (*n_estimators*) = 100, *max_depth* = None, min_samples_split = 2, min_samples_leaf = 1, random_state = 42
- Decision Tree: *max_depth* = 10, *criterion* = 'gini', *random_state* = 42
- Naive Bayes: *Priors* = None, *var_smoothing* = 1e-9

d) Overfitting and Underfitting Assessment: The chosen metrics helped in identifying any signs of overfitting or underfitting. Overfitting was indicated by a significant disparity between training and validation performance, while underfitting was indicated by poor performance on both training and validation data.

e) Commonality of Metrics: The metrics used in this evaluation are widely recognized and commonly employed in similar classification tasks. For example, confusion matrices, cross-validation, and train/test splits are standard practices for evaluating machine learning models (Agrawal).

f) Detailed Setup for Reproducibility:

1) Data Preparation: The dataset was split into training (80%) and testing (20%) sets using a random state of 43 to ensure reproducibility.
2) Cross-Validation: 5-fold cross-validation was performed on the training data to validate the model performance.
3) Model Training: Each model was trained using the specified hyperparameters and evaluated on the test set.
4) Evaluation: The confusion matrix, 5-fold cross-validation results, and train/test split accuracy were generated for each model.

g) Experimental Analysis: A detailed experimental analysis was conducted for each model:

- Random Forest: Showed the highest accuracy and F1-score, indicating strong performance in classifying the dataset.
- Decision Tree: Performed moderately well but showed slight signs of overfitting, as indicated by the variance in cross-validation scores.
- Naive Bayes: Had the lowest performance metrics but was computationally efficient and easy to implement.

D. Results

a) Random Forest:

- Accuracy: 0.69 (Train/Test Split)
- Classification Report:

Classification Report for Random Forest Model

	precision	recall	f1-score	support
1	0.81	0.91	0.86	6491
2	0.65	0.52	0.58	6658
3	0.62	0.67	0.65	6657
4	0.65	0.64	0.65	6374
accuracy			0.69	26180
macro avg	0.68	0.69	0.68	26180
weighted avg	0.68	0.69	0.68	26180

Fig. 6. Classification Report for Random Forest

- Confusion Matrix:

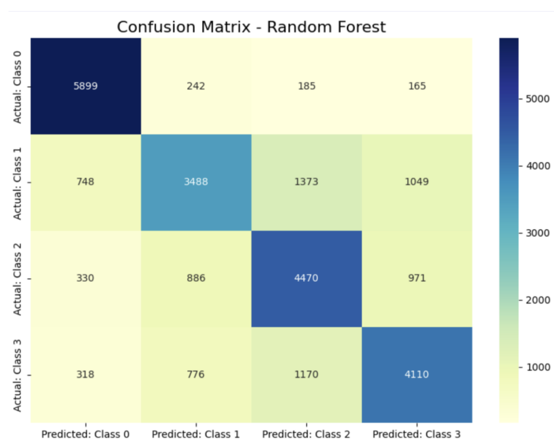

Fig. 7. Confusion Matrix for Random Forest

- 5-Fold Cross Validation:
  - Scores: [0.68235294, 0.68846448, 0.69009511, 0.68138584, 0.68723022]
  - Average Score: 0.69
- ROC-AUC Curve:

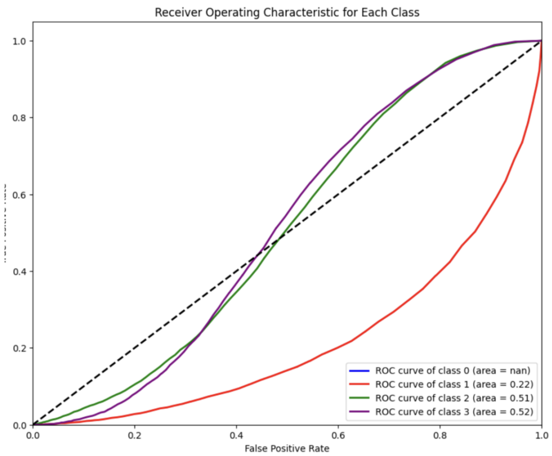

Fig. 8. ROC-AUC Curve for Random Forest

b) Decision Tree:

- Classification Report:

Classification Report for Decision Tree Model

	precision	recall	f1-score	support
1	0.55	0.98	0.70	6491
2	0.51	0.36	0.42	6658
3	0.53	0.59	0.56	6657
4	0.53	0.20	0.30	6374
accuracy			0.53	26180
macro avg	0.53	0.53	0.49	26180
weighted avg	0.53	0.53	0.49	26180

Fig. 9. Classification Report for Decision Tree

- Confusion Matrix:

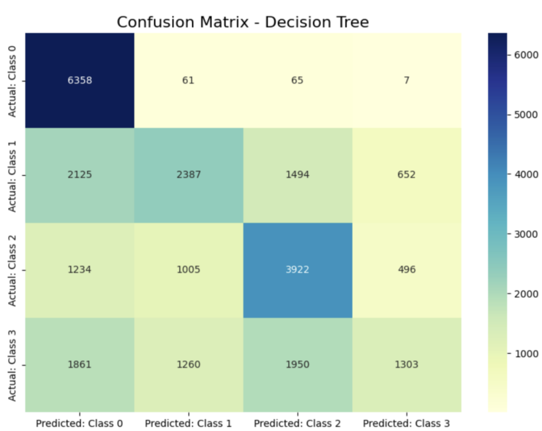

Fig. 10. Confusion Matrix for Decision tree

- 5-Fold Cross Validation:
  - Scores: [0.5302902979373567, 0.5389992360580595, 0.5366133160166545, 0.5297375759196302, 0.5353527636655334]
  - Average Score: 0.534
- ROC-AUC Curve:

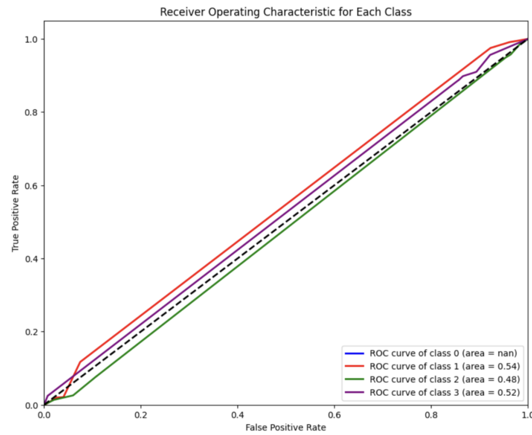

Fig. 11. ROC-AUC Curve for Decision Tree

Classification Report for Naive Bayes Model

	precision	recall	f1-score	support
1	0.41	0.81	0.55	6491
2	0.43	0.14	0.21	6658
3	0.45	0.54	0.49	6657
4	0.73	0.36	0.48	6374
accuracy			0.46	26180
macro avg	0.50	0.46	0.43	26180
weighted avg	0.50	0.46	0.43	26180

Fig. 12. Classification Report for Naive Bayes

c) Naive Bayes:

- Classification Report:
- Confusion Matrix:

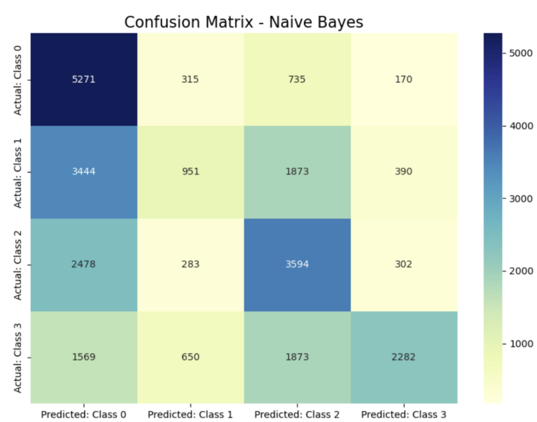

Fig. 13. Confusion Matrix for Naive Bayes

- 5-Fold Cross Validation:
  - Scores: [0.41037878, 0.41144152, 0.41106054, 0.41409665, 0.41098857]
  - Average Score: 0.41
- ROC-AUC Curve:

d) Analysis: The Random Forest model outperformed the other two models (Decision Tree and Naive Bayes) across both

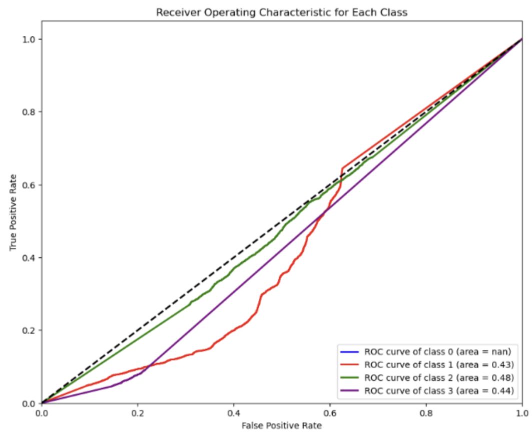

Fig. 14. ROC-AUC Curve for Naive Bayes

accuracy and F1-score metrics. The Decision Tree showed moderate performance but with some indications of overfitting. The Naive Bayes model had the lowest performance metrics but was computationally efficient. These results highlight the importance of selecting the right model for the task at hand, with Random Forest being the most suitable for this particular dataset.

E. Model Deployment

The deployed model will form the core of a web application designed to predict the severity of car accidents in real-time. The application will allow users to input accident-related parameters, which are then processed by the underlying models to predict the severity.

User inputs: Users can enter data such as weather conditions, time of day, road conditions, traffic information, etc.

User outputs: The application will provide a severity prediction ranging from minor to severe, rated on a scale of 1 to 4.

This tool will be invaluable for emergency responders and traffic management systems, enabling more efficient resource allocation based on predicted accident severity. Additionally, it can assist urban planners and policymakers in understanding accident trends and focusing on safety improvements where they are most needed.

The ethical considerations involve ensuring data privacy and security, particularly as users' location and potentially sensitive information are involved. Societally, the tool aims to enhance public safety and inform better urban planning and traffic management practices. However, care must be taken to avoid reliance solely on automated predictions in decision-making processes, maintaining human oversight to interpret and act on model outputs appropriately.

F. Front-End (Streamlit)

The front-end of our web application focuses on providing a user-friendly interface that allows users to input data, receive predictions, and view insights into traffic accident severity. The layout is structured to ensure ease of use while maintaining a professional appearance that caters to both casual users and professional stakeholders.

Home Page: This is the landing page where users are greeted and provided with a brief overview of the application's purpose and how to use it. The navigation menus are always on the left side of the page to help users navigate to other sections of the site. The home page also contains a quick data preview.

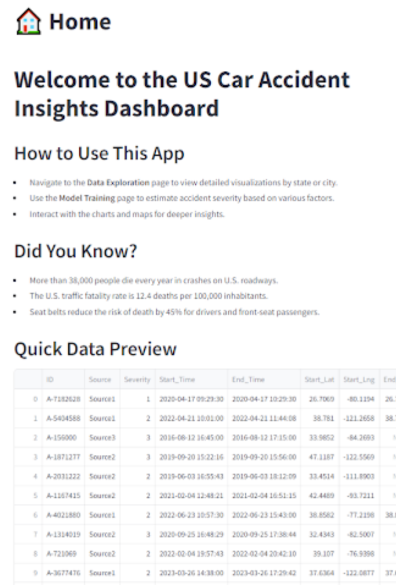

Fig. 15. Home Page

Data Exploration Page: This page provides detailed accident data by state or city, helping users understand patterns and trends. Users can filter the data by specific criteria, and the option to choose a year to see accident statistics in order to gain deeper insights into factors contributing to accidents in different regions. Features include:

- Data summary sample table

	Severity	Start_Time	Start_Lat	Start_Lng	End_Lat	End_Lng	Distance(mi)	Te
count	100,000	100000	100,000	100,000	56,012	56,012	100,000	
mean	2.2144	2020-06-03 18:44:54.075510016	36.2424	-94.6682	36.3236	-95.5989	0.5585	
min	1	2016-02-08 06:49:27	24.6027	-124.5357	24.6816	-124.5445	0	
25%	2	2018-11-22 12:08:08.750000128	33.4336	-117.2067	33.4847	-117.7184	0	
50%	2	2020-11-11 20:30:32	35.8657	-87.763	36.3843	-87.9651	0.029	
75%	2	2022-01-18 16:20:22.500000	40.1157	-80.3441	40.251	-80.2338	0.459	
max	4	2023-03-31 21:19:00	48.9958	-68.2269	48.9981	-68.2368	104.302	
std	0.4895	None	5.0609	17.3806	5.2644	18.0827	1.7234	

Fig. 16. Data Summary

- Data types table

	ID	Source	Severity	Start_Time	End_Time	Start_Lat	Start_Lng	End_Lat	End_Lng
Data Type	object	object	int64	datetime64[ns]	object	float64	float64	float64	float64

Fig. 17. Data Types

- Showing missing values

	ID	Source	Severity	Start_Time	End_Time	Start_Lat	Start_Lng	End_Lat
Null Value Count	0	0	0	0	0	0	0	43,988
Non-Null Value Count	100,000	100,000	100,000	100,000	100,000	100,000	100,000	56,012
Percentage	0	0	0	0	0	0	0	43.99

Fig. 18. Missing Values

- Total accidents in a specific year using a sliding bar, along with the average accident severity of that year

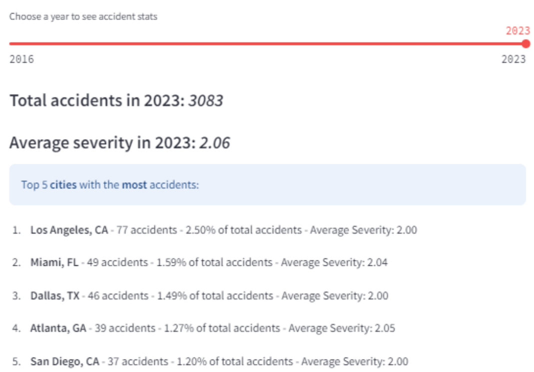

Fig. 19. Year Accidents by Using Sliding Bar

Data Visualization Page: This page offers various graphs to help users explore accidents by time and different conditions. The interface provides visual tools that enable users to identify trends and correlations, making it easier to understand the impact of different variables on accident occurrences. Features include:

- Line graph showing accidents over time

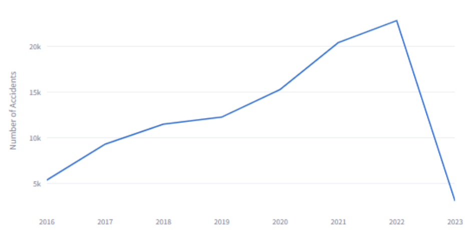

Fig. 20. Line Graph of Accidents Over Time

- Bar Chart showing accidents by the time of day

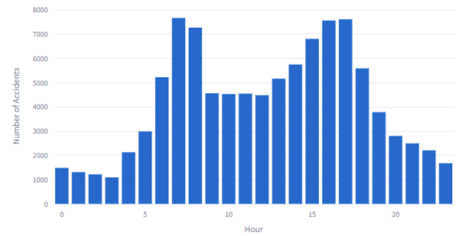

Fig. 21. Bar Chart of Accidents by the Time of a Day

- Bar Chart Top 10 weather conditions for accidents

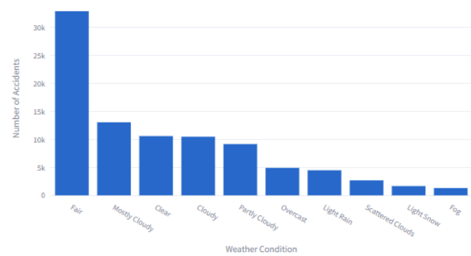

Fig. 22. Bar Chart of Top 10 Weather Conditions for Accidents

- Heat map for accidents by state

Fig. 23. Heat Map for Accidents by State

- Scatter plot of accidents with severity level 4

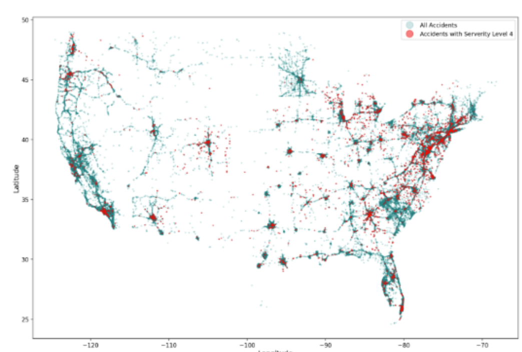

Fig. 24. Scatter plot of Accidents with Severity Level 4

Model Prediction Page: Users can make predictions about the severity of accidents based on their inputs. The intuitive interface ensures that even users with limited technical expertise can easily navigate the prediction process and obtain meaningful results. This page includes:

- Pre-trained models with three selection options, connected via pkl files from the backend

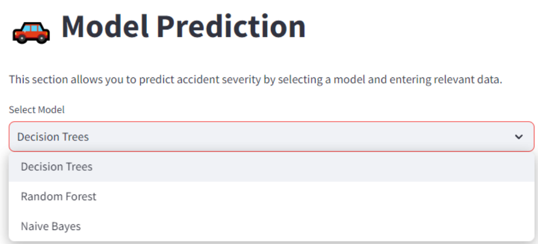

Fig. 25. Pre-trained Models

- Input fields for conditions such as weather, road type, or other features

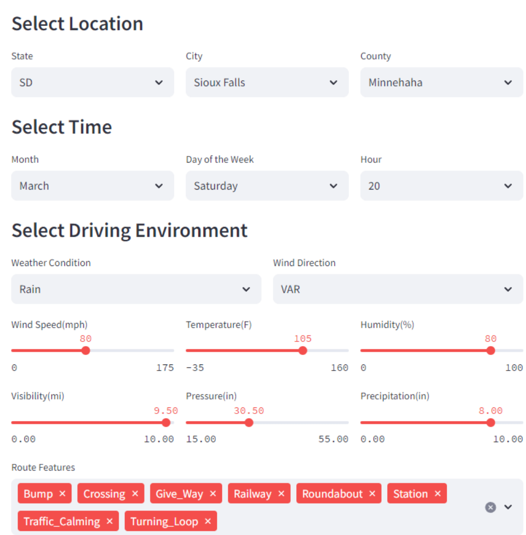

Fig. 26. User Input Fields

- A randomized button to automatically fill in inputs, making it easier to see how different scenarios affect the predictions

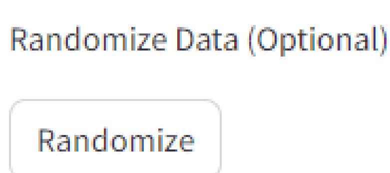

Fig. 27. Randomize Button

- Prediction results showing the severity level

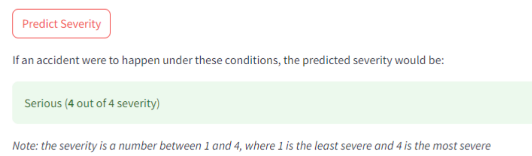

Fig. 28. Prediction Result

The connection between the front-end and back-end of our application is facilitated through APIs and pkl files. When a user selects a model from the three options on the prediction page, the information is transmitted to the back-end. It is then processed by our pre-trained models, which analyze the data and send back the prediction results based on the current inputs, ensuring a seamless flow of information and responses between the user interface and the backend.

The user interface of our web application features a left navigation bar for easy access to all pages. Users can interact with the application through various input widgets, including dropdown menus for selecting prediction models and feature categories like weather and road conditions, text inputs for entering location information, and a date selector for selecting dates for data analysis. Functional buttons such as 'Randomize', which inputs random data with different scenarios, and the 'Predict' button, which predicts the final results, facilitate user interaction. Additionally, the application includes interactive chart elements like bar charts, line graphs, and heatmaps for displaying trends and maps for visualizing accident hotspots. These elements are designed to enhance user engagement and provide insightful data visualizations.

IV. Conclusion

Overall, our project set out to develop a predictive model to analyze and forecast traffic accidents across the United States, employing advanced machine learning techniques such as Decision Trees, Random Forests, and Naive Bayes. This model integrates weather data and geographical insights, aiming to provide interpretable and actionable predictions for various stakeholders, including city planners and traffic authorities. Through rigorous data preparation, feature engineering, and model evaluation, we achieved a robust system capable of predicting accident severity with high accuracy, which was then integrated into a user-friendly web application for real-time use.

The deployment of this predictive tool carries significant implications for societal safety and resource management. For emergency responders and traffic management systems, it ensures more efficient resource allocation and quicker response times, potentially saving lives by mitigating the severity of accidents. Urban planners can leverage the insights derived from the model to enhance road safety measures in accident-prone areas, promoting a preventative approach to traffic management. Additionally, by adhering to ethical standards in data handling and model fairness, the project sets a precedent for responsible AI development, contributing to a safer and more informed society.

V. Team Member Contribution

A. Technical Components

- Henry Bao: Training Model
- Stella Hong: Front End (Streamlit)
- Mingze Gao: Front End & Data Processing
- Yunhao Hu: Model Evaluation

B. Writing Components

- Henry Bao: Data Collection, Exploration, and Processing, Methods and Model Training
- Stella Hong: Abstract, Introduction, Background, and Front-End(Streamlit)
- Mingze Gao: Model Deployment, Front-End (Streamlit), and Conclusion
- Yunhao Hu: Model Evaluation, Results

References

[1] Muhammad Bilal Khan. Geospatial analysis of us traffic accidents. Kaggle, 2021. https://www.kaggle.com/muhammadbilalkhan/geospatial-analysis-of-us-traffic-accidents.

[2] Sobhan Moosavi, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. A countrywide traffic accident dataset. *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2019. https://ar5iv.labs.arxiv.org/html/1906.05409.

[3] R. Zhou. Ml to predict accident severity. Kaggle, 2020. https://www.kaggle.com/code/phip2014/ml-to-predict-accident-severity-pa-mont.

[4] Sobhan Moosavi. Us accidents (2016 - 2023). Kaggle, 2023. https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents.