Predicting Accident Severity using Machine Learning

Khushii Wason

*Corresponding Author:f20210123@dubai.bits-pilani.ac.in

Abstract

TIn the present day world, road accidents are a major economic and societal challenge. They are responsible for a large amount of fatalities and cost hundred of billions of dollars annually. It is of vital importance that we mitigate the occurence of accidents, with focus on fatal accidents. Through this study, we use supervised machine learning techniques to predict the severity of an accident. Moreover, we will find the major factors that influence accidents. The study uses a data-driven approach to build a predictive model that is able to classify an accident from 1 to 4, where 1 indicates the least impact on traffic and 4 indicates a significant impact on traffic. Through data pre-processing, Exploratory Data Analysis, feature selection, and model refinement, we want to develop a highly accurate and reliable predictive model. Another goal of the study is to find the most influential factors on severity of road accidents on traffic. We will find the most important features for accident severity using various feature importance techniques. Finding these features will enable various stakeholders like the transport ministry, police, etc, to focus their resources and efforts on building suitable policies to address the most influential factors, thus allowing for a more targeted approach to prevent accidents. The tool developed as part of this study can have substantial implications for policymakers, as it helps in predicting the severity of road accidents. It can allow emergency services and law enforcement agencies to have a prompt response. Furthermore, the identification of influential factors provides evidence for decision-making and building programs for better intervention to prevent accidents. Moreover, it will allow for better resource allocation. In conclusion, this project can act as a pivotal step to enhance road safety by using the power of machine learning to predict accident severity and discerning the critical factors at play. The potential impact of this research extends far beyond the academic sphere, offering tangible benefits in terms of lives saved, injuries prevented, and economic resources conserved.

1. Introduction

Almost 1.19 million people die because of road accidents, while 20-50 million suffer non-fatal diseases, with some of them causing disabilities(WHO). Road accidents have become a massive challenge impacting public safety and economic stability. Currently, there is a need to find the most critical factors responsible for a road accident. To achieve that in this study, we use various machine learning techniques to find the most pivotal factors that influence accident severity.

Through the study we would train a model that predicts accident severity. The model is a potent tool as it can find the accident's severity, even without detailed information such as driver demographics or vehicle specifics. The capabilities of such models will allow the development of better accident response strategies, enabling prompt and targeted interventions in critical situations.

To achieve a model that generalises well, the study has used multiple pre-processing techniques. The data has been cleaned, missing values have been filled, and the data has been standardised to ensure that the outliers do not have an effect on the model. To see the presence of outliers we used box-plots.

One-hot Encoding, was used to convert categorical variables into numerical values. This was done so that models could use these cateogorical features to predict the severity. Following which, scatter plots were plotted along with correlation heatmap to find the potential relationships and correlations between different attributes.

Classification models, like Random Forests, decision trees, SVM, etc.,were applied in this study. These algorithms are designed to categorize data points into predefined classes, making them ideally suited for predicting accident severity. Through performance evaluation and comparison of various classification models, we will find the best framework for our problem statement, which balances accuracy, F1 scores, and AUC scores and prevents overfitting.

For Model Evaluation, Precision, recall, F1-score, and accuracy were to check how the models perform. These metrics will provide a good understanding of the model's ability to classify instances correctly, which is particularly crucial in accident severity prediction.

Finally, we used techniques like Gini Importance to find the most important features that cause road accidents. These techniques provides valuable insights into the hierarchy of feature importance, showing the attributes that exert the most substantial impact on accident severity prediction.

In conclustion, this project aims to develop predictive models for accident severity assessment using various techniques to provide reliable and credible results. We have created a model in this study that can generalise well to unseen data since the model has been trained using various data preprocessing techniques, classification models and feature importance methodologies.

2. Literature Survey

1. The author compares the prediction performance of three data mining techniques in Modeling Road Accident Severity with Comparisons of Logistic Regression, Decision Tree, and Random Forest. These techniques, namely Logistic Regression (LR), Classification and Regression Trees (CART), and Random Forest (RF), are assessed based on accuracy. The RF model, utilizing nine significant variables identified by LR, exhibits the highest accuracy (74.43) among the tested models. The study emphasizes the importance of identifying significant variables for improving prediction performance.

2. The International Journal of Scientific Research in Computer Science, Engineering, and Information Technology presents a research paper on "Analysis of Road Accident Severity Using Machine Learning Techniques." The study aims to use machine learning techniques to analyze road accidents and identify influencing factors. Utilizing data from the UK Department of Transport, the authors employ Decision Tree, Random Forest, and Artificial Neural Network (ANN) models. The Random Forest algorithm demonstrates the highest accuracy in predicting accident severity, leading to the conclusion that machine learning approaches are effective in analyzing road data.

3. The Impact of Variations in Urban Functional Zones on the Severity of Injury from Traffic Accidents explores variations in Beijing's urban functional zones and their influence on injury severity in traffic accidents. Analyzing 3982 accident data sets, the study identifies 17 variables associated with injury severity. The findings from binary logistic regression, categorization, and regression tree analysis highlight common and unique influencing elements across urban functional zones.

4. According to "Accident Severity Prediction Using Machine Learning," traffic accidents

pose significant global challenges, prompting the need for accurate severity prediction. The study employs supervised and unsupervised machine learning algorithms on traffic accident data but lacks detailed information on performance metrics or final model results.

5. "Data-Driven Urban Traffic Accident Analysis and Prediction Using Logit and Machine Learning-Based Pattern Recognition Models" investigates and predicts urban traffic accident severity in Rasht. Logistic regression and Artificial Neural Network (ANN) models are employed, with the ANN model outperforming in prediction accuracy (98.9). Critical variables such as operating speed, weather conditions, and vehicle quality are identified as impactful factors.

6. Focused on motorcycle injury severity assessment, this study compares various machine learning techniques, including random forest, support vector machine, and binary logistic regression. The random forest model emerges as the most effective, as evidenced by lower misclassification rates and a higher Area under the Curve (AUC) in 5-fold cross-validation.

7. "Applications of Machine Learning Methods in Traffic Crash Severity Modeling: Current Status and Future Directions" discusses the increasing application of machine learning methods, including decision trees, random forests, support vector machines, and artificial neural networks, in crash severity modeling. These methods prove advantageous in handling complex relationships and non-linear patterns, offering insights into traffic safety research.

8. The paper "Road Accident Prediction Model Using Machine Learning" focuses on creating a road accident prediction model using SVM Apriori, linear regression, and association rule mining. While mentioning the use of additional algorithms like RFCNNCNN and clustering, the paper lacks specific details on outcomes.

9. "A Case Study in Selangor of a Road Accident Prediction Model" aims to predict accidents using Multiple Linear Regression and Artificial Neural Network techniques. The comparison indicates higher predicted values and errors with Multiple Linear Regression, while the ANN model demonstrates lower predicted values, higher correlation, and lower errors.

10. "Machine Learning Techniques for Traffic Accident Analysis and Hotspot Prediction" uses accident data from Setubal, Portugal, employing supervised and unsupervised techniques. The C5.0 algorithm accurately identifies important variables indicating accident severity, with a suggestion that random forest may aid in predicting accident hotspots.

11. "Analysis of Accidents and Prediction of Severity" employs statistical methods and machine learning algorithms to analyze traffic congestion and accidents. Utilizing Kaggle's US Accidents dataset, the study employs algorithms such as Extreme Gradient Boosting, Random Forest Classifier, Decision Tree Classifier, and Logistic Regression, aiming to create predictive models for real-time decision-making.

12. "Analysis of Road Accidents" uses machine learning algorithms, including Support Vector Machine (SVM), K-Nearest Neighbour (KNN), and Random Forest, to analyze traffic accidents in the UK from 2005 to 2015. SVM performs well in predicting the "Sex of Casualty," while Random Forest excels in predicting "Rural or Urban" attributes.

13. "Machine learning techniques for forecasting the traffic accident severity" focuses on identifying factors influencing accident severity using classification algorithms. Support Vector Machine, K-Nearest Neighbors, and Multilayer Perceptron models outperform others, achieving the highest accuracy rates of 91, 92, and 94, respectively.

14. "Traffic Accident Injury and Severity Prediction Using Machine Learning Algorithms" explores supervised and unsupervised learning techniques for predicting injury severity. Unsupervised learning shows potential for accurate estimation, providing valuable insights for traffic departments and the public to reduce accident risks.

15. "Prediction of Fatalities in Vehicle Collisions in Canada" employs data mining techniques on Canada's National Collision Database. XGBoost and Lasso Regression models predict fatalities, with XGBoost achieving 83 accuracy. The study emphasizes the importance of factors like "used safety devices" and "collision configuration" in predicting fatalities.

3. Dataset

3.1 Description

We used the US Accidents dataset by Sobhan Moosavi.It consists of 7.7 million traffic accidents that took place in the United States, from ebruary 2016 to March 2023. Due to computational restrictions, we choose 40,000 samples, with 10,000 belonging to each of the 4 classes.

3.2 Preprocessing

In the study, we first looked at the features, and after careful analysis of the features. Some of the features like:

- County
- Zipcode
- Country
- Airport_Code
- Amenity
- Bump
- Crossing
- Give_Way
- Junction
- No_Exit
- Station
- Stop
- Railway
- Roundabout
- Traffic_Calming
- Civil_Twilight
- Nautical_Twilight
- Astronomical_Twilight
- Start_Lat
- Start_Lng
- Roundabout
- Traffic_Signal
- Turning_Loop
- ID
- End_Lat
- End_Lng
- Source
- Timezone
- Weather_Timestamp
- Description
- Street
- City

These columns were dropped because other columns consisted of the same information, or they had a large amount of missing values or irrelevant values. Some of the features also needed the usage of NLP techniques, which has not been covered till now.

Following the removal of columns, new columns were created to divide the data into year, day,month. The data was also divided into 4 seasons, Summer, Winter, Autumn, and Spring. The data was also divided into Morning, Afternoon, Evening and Night. Finally, the Data was also divided into Weekday and Weekend.

A new column for duration in minutes was also created by using the Start Time and End Time of an Accident.

Since multiple essential columns had missing values, Simple Imputer from Scikit Learn was used to fill numerical and categorical values. The numerical balues filling used Mean, while for categorical variables most frequent strategy was used.

Converted the Wind_Direction feature into 6 classes by combining similar categories, it initially consisted 18 categories.

Converted the Weather_Condition feature into 9 classes 'Clear, 'Cloud', 'Rain', 'Heavy Rain', 'Snow', 'Heavy Snow', 'Fog','Windy', and 'Others'.

Finally, we used One-Hot Encoding to convert categorical columns to numerical columns to ensure that model can also use these features.

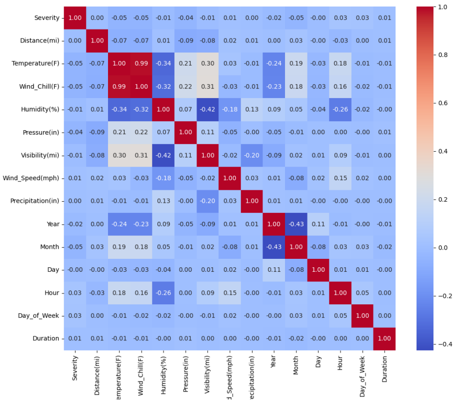


Figure 1: Heatmap of the Features

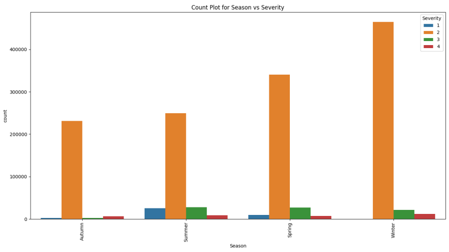


Figure 2: Count Plot-Season

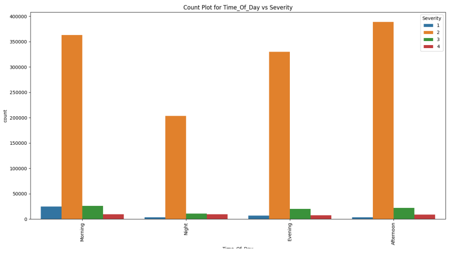


Figure 3: Count Plot-Time of Day

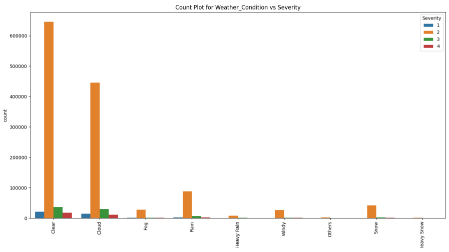


Figure 4: Count Plot-Weather Condition

4 Methodology

4.1 Classification

4.1.1 Classification Models

In the classification phase, various machine learning models were employed, each configured with specific hyperparameters using Randomized Search CV (RandomizedSearchCV) from the scikit-learn library to optimize their performance.

The Decision Tree model utilized the 'gini' criterion for information gain, employed the 'best' method for splitting nodes, balanced class weights, and set the random state to 0,min_samples_split as 4 and max_dept as 9.

For the Random Forest model, the 'entropy' criterion was selected, and the number of estimators was set to 140, min_samples_split as 4, min_samples_split as 9, max_samples as 0.7, max_features as 'auto',max_dept as 29, class weight as 'balanced_subsample' and bootstrap as True.

The Naive Bayes,K-Nearest Neighbors and Support Vector Machine (SVM) operated with their default hyperparameters, as they do not involve hyperparameter tuning because of their poor performance.

The Logistic Regression model, designed for multi-class classification, used the 'saga' solver, penaty as 'l1' and C as 456.

Gradient Boosting and AdaBoost, ensemble methods, employed hyperparameters with n_estimators as 300.

LGBM classifier another ensemble method was used with 900 n_esimatores.

A voting classifer with Hard voting, i.e. with majority voting was used. The models used for making the voting classifier were Random Forest and Gradient Boost.

A stacking classifier was also made with the base classifer being Gradient Boost. The stacking classifier comprised of Random Forest, Decision Trees, Adaboost and Gradient Boost

4.1.2 Data Splitting and Cross-Validation

The dataset was divided into training and testing sets, with an 80-20 split. To assess the models' robustness and generalizability, Stratified k-fold cross-validation with 10 folds was employed.

4.1.3 Evaluation Metrics

Classification models were evaluated using common metrics—accuracy, precision, recall, and F1 score to comprehensively assess their predictive

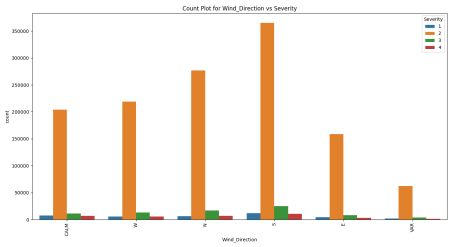


Figure 5: Count Plot-Wind Direction

capability, ability to identify positive instances, capture all positive instances, balance precision and recall, and discriminate between different classes. These metrics collectively provided insights into various aspects of model performance for a thorough assessment of suitability for the classification task.

These collectively aimed to mitigate class imbalance impact, providing a more reliable assessment of model performance across different classes.

Table 1: Classifier Performance

Classifier	Train Acc.	Test Acc.	Precision	Recall	F1 Score
Decision Tree	0.73	0.70	0.70	0.70	0.70
Random Forest	0.79	0.72	0.72	0.72	0.72
Logistic Regression	0.52	0.51	0.50	0.51	0.50
SVM	0.33	0.32	0.36	0.32	0.22
Naive Bayes	0.49	0.48	0.55	0.48	0.44
AdaBoost	0.66	0.64	0.65	0.65	0.64
Gradient Boost	0.78	0.74	0.74	0.74	0.74
KNN	0.71	0.58	0.57	0.58	0.57
LGBM	0.78	0.74	0.74	0.74	0.74
Voting Classifier	0.79	0.72	0.72	0.72	0.72
Stacking Classifier	0.78	0.74	0.74	0.74	0.74

The classifiers were evaluated on multiple performance metrics, including accuracy, precision, recall, and F1 score. Among the models, Decision Tree, Random Forest, AdaBoost, Gradient Boost, and Stacking Classifier demonstrated strong overall performance with high accuracy and balanced precision and recall. Logistic Regression, SVM, and Naive Bayes, however, exhibited lower accuracy and less balanced precision and recall. Notably, KNN showed a significant drop in accuracy on the test set, indicating potential overfitting or misalignment with the test data. The Voting Classifier displayed performance similar to individual models. In summary, the ensemble methods and certain individual models appear promising, while careful consideration is warranted for models with lower performance

4.2 Feature Importance

The bar plots that depict the performance of various models provide useful insights into the factors that influence accident severity. Distance and duration stand out prominently among these factors, indicating a strong correlation with accident severity. The inclusion of various weather-related factors, such as Pressure and Wind Chill, emphasizes their importance in determining accident outcomes. These findings suggest that geographical and temporal factors, as represented by distance and duration, are important in determining accident severity. Furthermore, as indicated by weather-related variables, the impact of weather conditions emerges as a critical determinant. This information can be used to guide future investigations and safety measures, emphasizing the importance of considering and addressing these influential factors in order to improve overall road safety and reduce the severity of accidents.

5. Conclusion

A thorough evaluation of classifiers using multiple performance metrics such as accuracy, precision, recall, and F1 score revealed distinct trends in their effectiveness. The Decision Tree, Random Forest, AdaBoost, Gradient Boost, and Stacking Classifier all demonstrated strong overall performance, with high accuracy and balanced precision and recall. Logistic Regression, SVM, and Naive Bayes, on the other hand, demonstrated lower accuracy and less balanced precision and recall. Notably, KNN's accuracy dropped significantly, indicating potential issues such as overfitting or misalignment with the test data. Individual models performed similarly to the Voting Classifier.

Simultaneously, bar plot analysis provided critical insights into the impactful factors influencing accident severity. Distance and duration emerged as significant contributors, indicating a strong correlation with accident severity. The inclusion of weather-related factors like Pressure and Wind Chill highlighted their importance in determining accident outcomes. These findings emphasized the importance of geographical and temporal factors, as well as weather conditions, in accident severity.

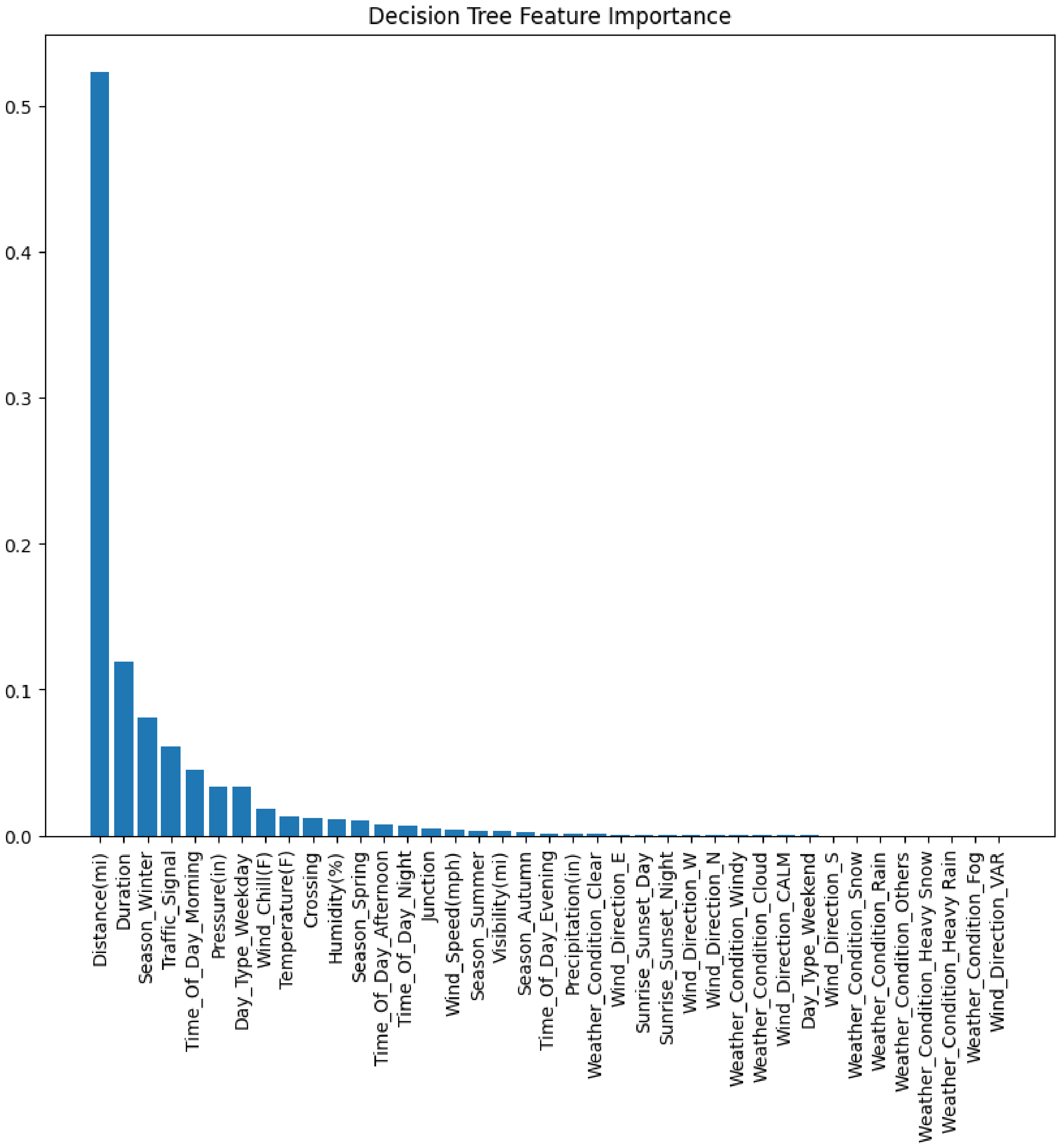


Figure 6: Decision Tree Feature Importance

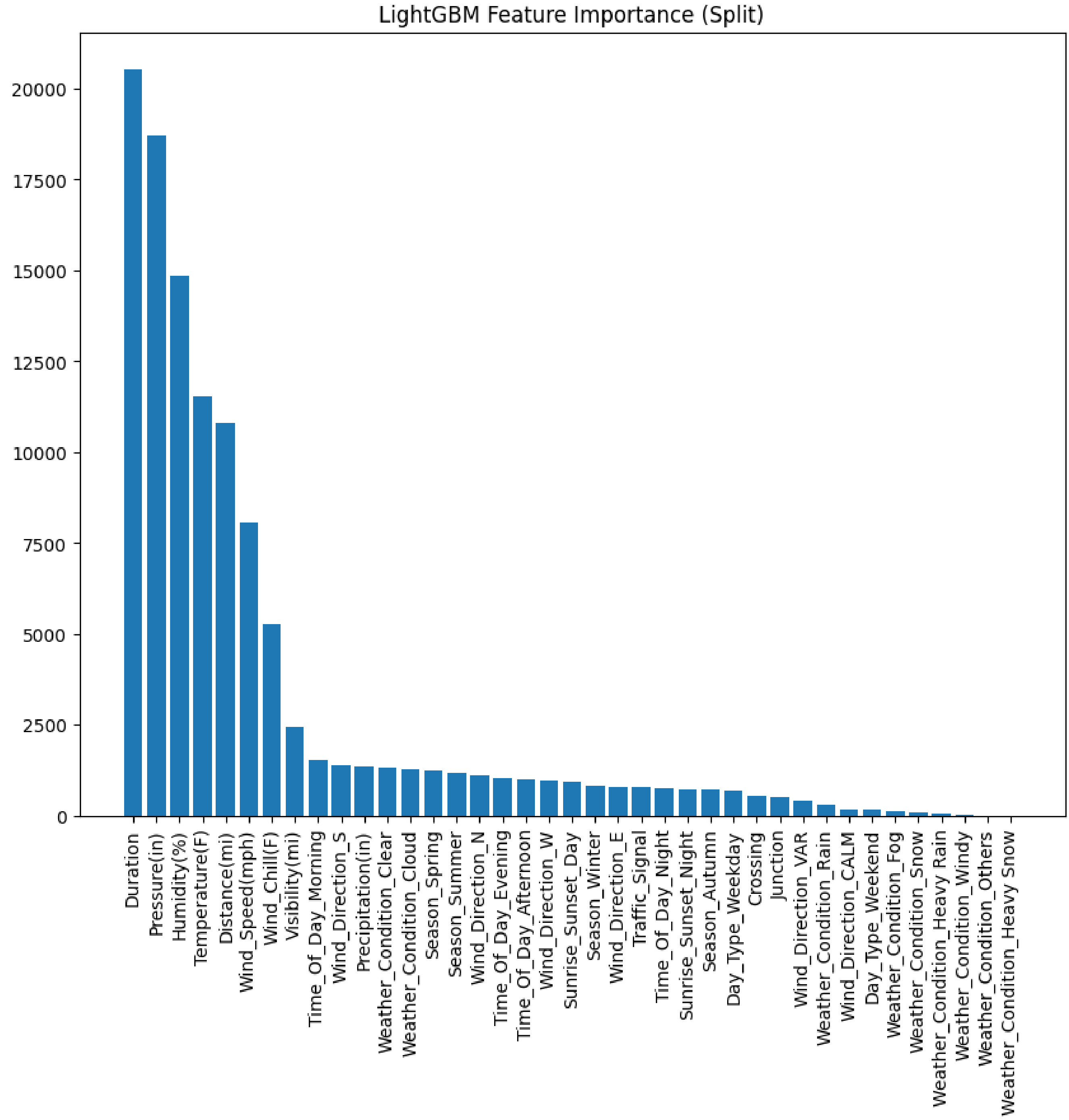


Figure 7: LGBM Feature Importance

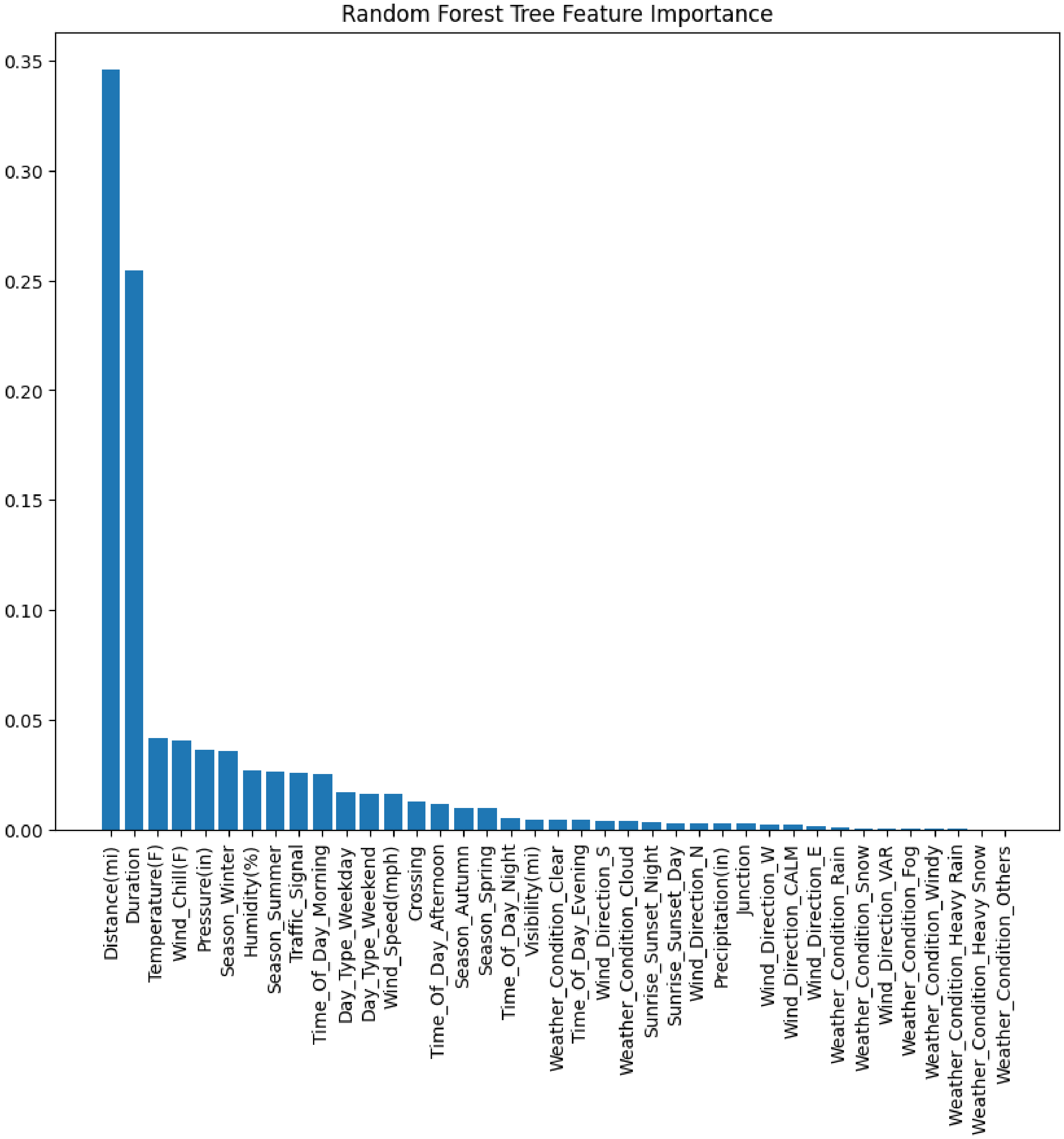


Figure 8: Random Forest Feature Importance

This collective understanding can be used to create a predictive model that improves road safety and reduces the severity of accidents. By incorporating the identified influential factors, the model can provide real-time predictions, allowing proactive measures to be taken to prevent or reduce the severity of accidents. This proactive approach not only reduces the overall number of accidents, but it also reduces associated costs and fatalities. The model can help authorities optimize resource allocation by allowing them to focus on specific areas and conditions that pose a higher risk.As a result, the predictive model is a valuable tool for decision-makers, policymakers, and stakeholders in their ongoing efforts to improve road safety and reduce the societal and economic costs of accidents.

References

1. Chen, M.-M.; Chen, M.-C. Modeling Road Accident Severity with Comparisons of Logistic Regression, Decision Tree and Random Forest. Information 2020, 11, 270. https://doi.org/10.3390/info11050270

2. R. Arunachalam, S. Peararulselvi, M. Saraswathi, & M. Saraswathi. (2023). Road accident severity prediction using machine learning. International Journal of Advanced Research in Science, Communication and Technology, 631–636. https://doi.org/10.48175/ijarsct-9629

3. Sun, Z., Wang, J., Chen, Y., & Lu, H. (2018). Influence factors on injury severity of traffic accidents and differences in urban functional zones: The empirical analysis of Beijing. International Journal of Environmental Research and Public Health, 15(12), 2722. https://doi.org/10.3390/ijerph15122722

4. Najafi Moghaddam Gilani, V., Hosseinian, S. M., Ghasedi, M., & Nikookar, M. (2021). Data-driven urban traffic accident analysis and prediction using logit and machine learning-based Pattern Recognition Models. Mathematical Problems in Engineering,

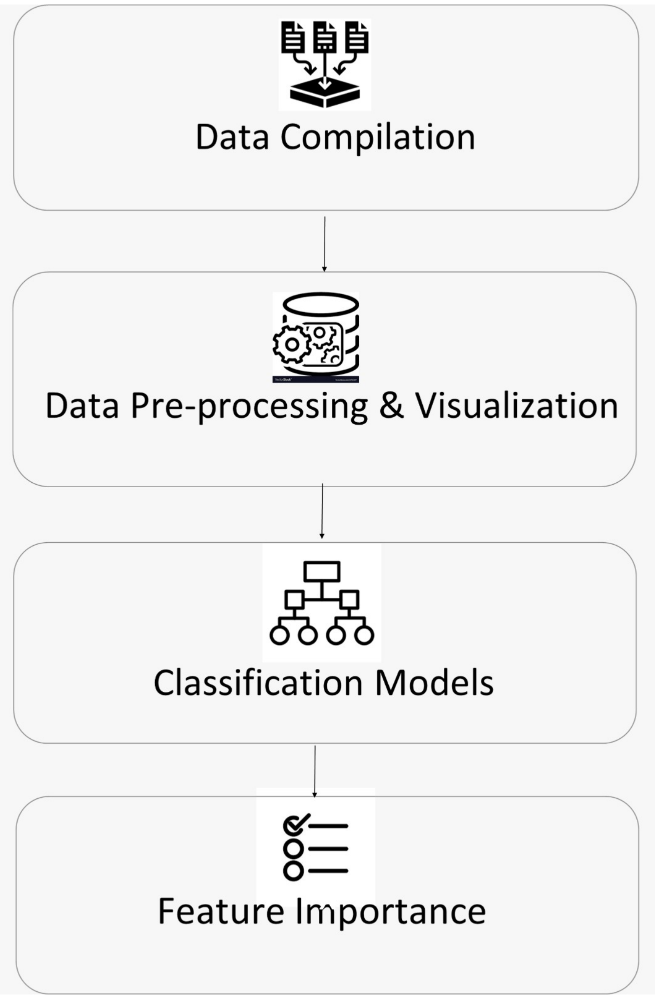
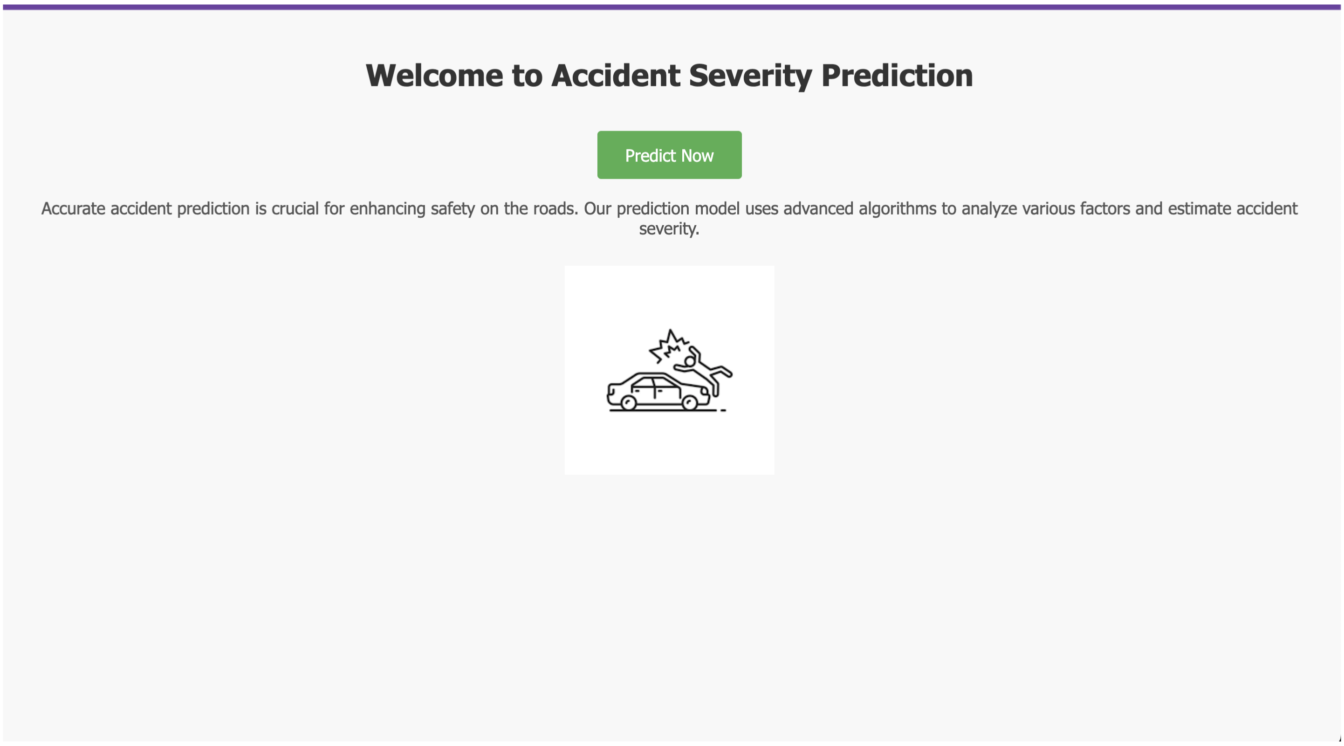


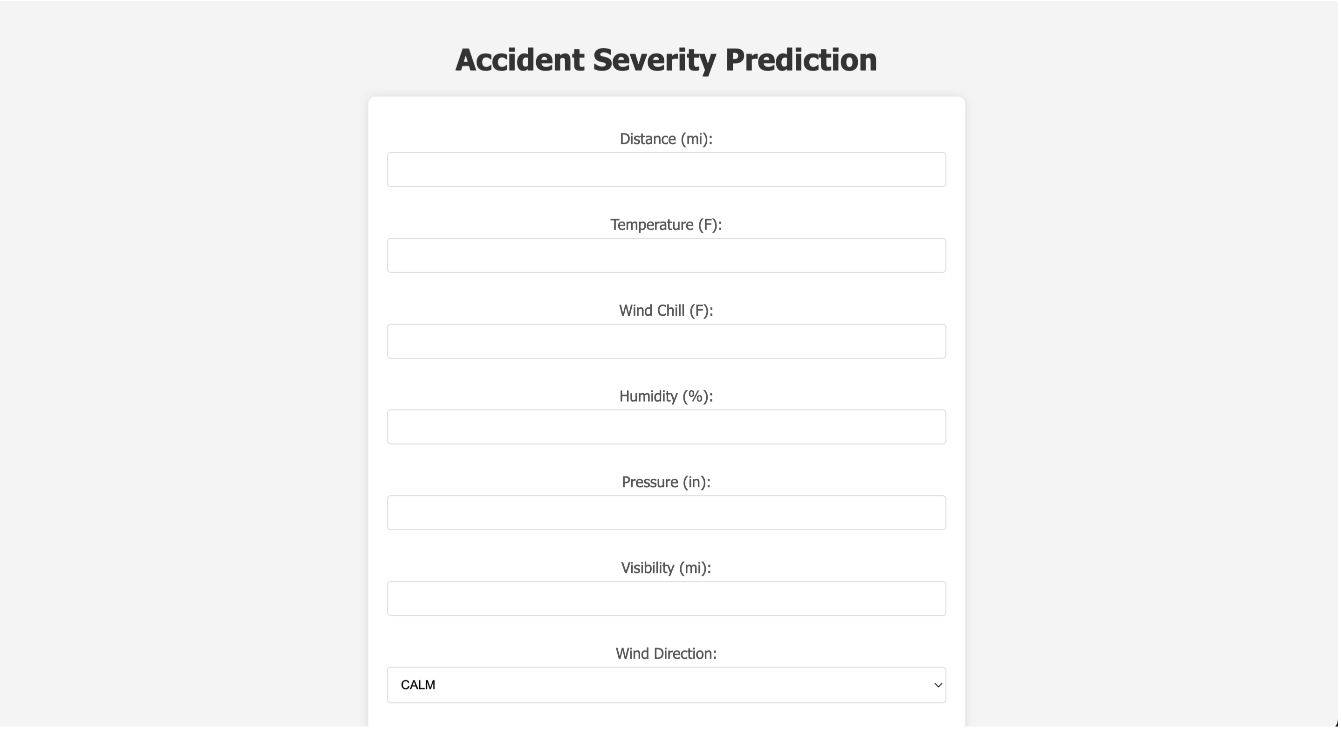
Figure 9: Architecture Diagram



Welcome to Accident Severity Prediction

Accurate accident prediction is crucial for enhancing safety on the roads. Our prediction model uses advanced algorithms to analyze various factors and estimate accident severity.

Figure 10: GUI Front



Accident Severity Prediction

Distance (mi):

Temperature (F):

Wind Chill (F):

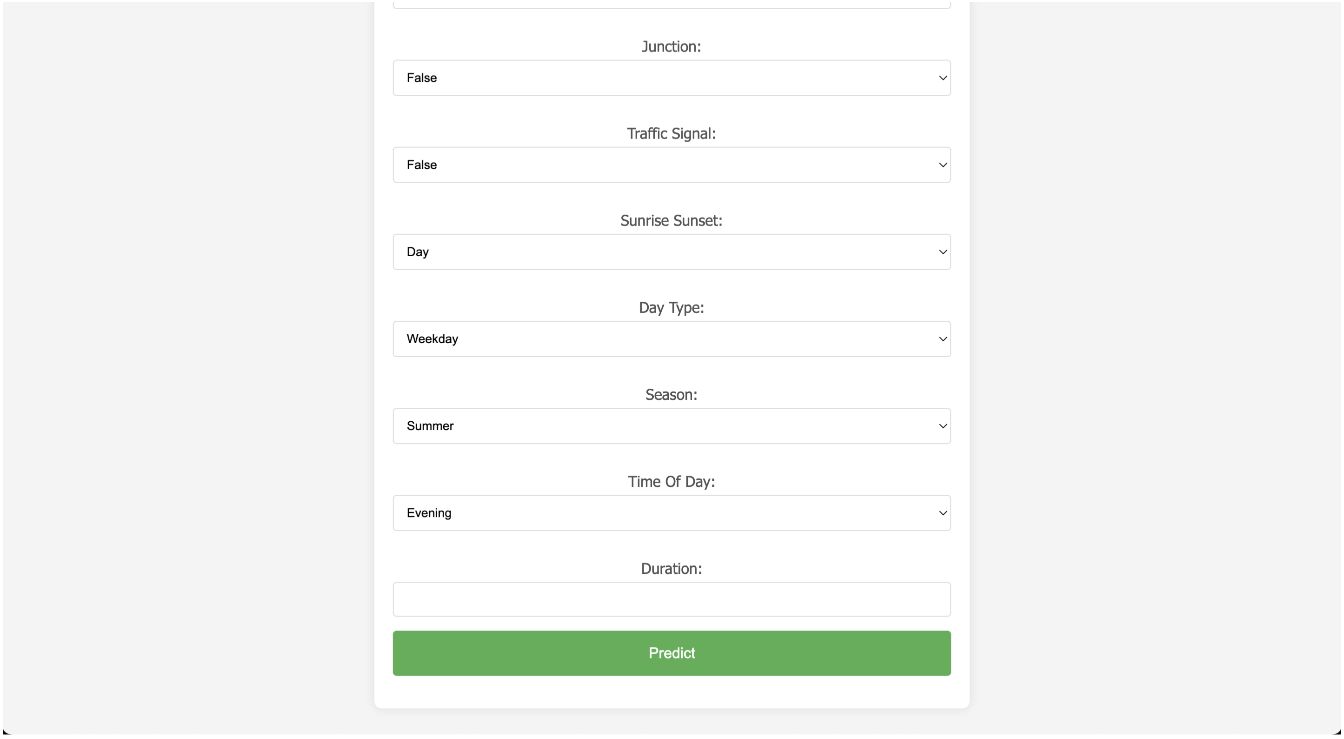
Humidity (%):

Pressure (in):

Visibility (mi):

Wind Direction: CALM

Figure 11: GUI Input 1


Junction:

False

Traffic Signal:

False

Sunrise Sunset:

Day

Day Type:

Weekday

Season:

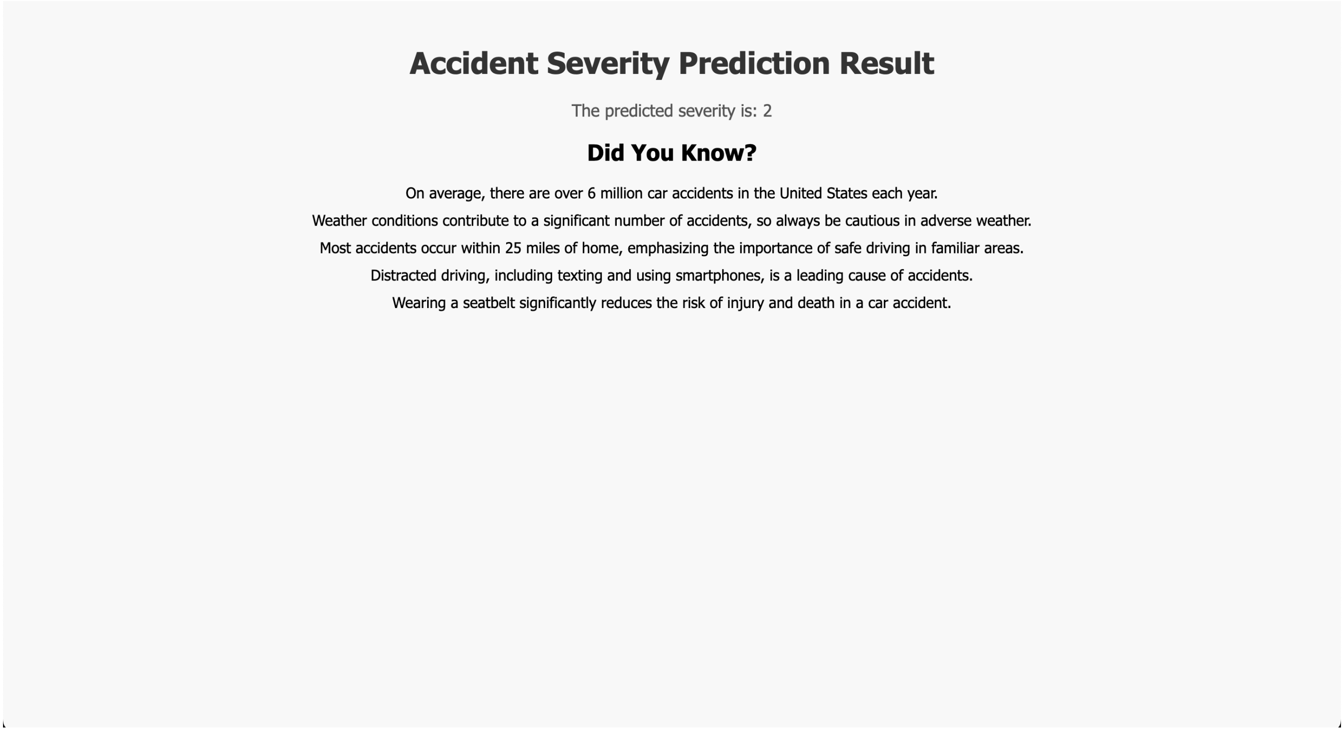
Summer

Time Of Day:

Evening

Duration:

Figure 12: GUI Input 2


Accident Severity Prediction Result

The predicted severity is: 2

Did You Know?

On average, there are over 6 million car accidents in the United States each year.

Weather conditions contribute to a significant number of accidents, so always be cautious in adverse weather.

Most accidents occur within 25 miles of home, emphasizing the importance of safe driving in familiar areas.

Distracted driving, including texting and using smartphones, is a leading cause of accidents.

Wearing a seatbelt significantly reduces the risk of injury and death in a car accident.

Figure 13: Prediction

2021, 1–11. https://doi.org/10.1155/2021/9974219

5. Nischitha M, Pavankumar B Jois, Sheethal C J, Shreya S, & Suman Jayakumar. (2022). Accident severity prediction using machine learning. International Journal of Advanced Research in Science, Communication and Technology, 858–862. https://doi.org/10.48175/ijarsct-5172

6. Rezapour, M., Farid, A., Nazneen, S., & Ksaibati, K. (2021). Using machine learning techniques for evaluation of motorcycle injury severity. IATSS Research, 45(3), 277–285. https://doi.org/10.1016/j.iatssr.2020.07.004

7. Wen, X., Xie, Y., Jiang, L., Pu, Z., & Ge, T. (2021). Applications of machine learning methods in traffic crash severity modelling: Current status and Future Directions. Transport Reviews, 41(6), 855–879. https://doi.org/10.1080/01441647.2021.1954108

8. Mane, S., Shelke, N., Patil, P., Bijale, S., & Shinde, Dr. J. (2023). Road accident prediction model using machine learning. International Journal for Research in Applied Science and Engineering Technology, 11(2), 1–4. https://doi.org/10.22214/ijraset.2023.48883

9. International Journal of Scientific Research in Science, E. and T. I. (2017, September 15). Study of road accident prediction model at Accident Blackspot area: A case study at selangor. Academia.edu. https://www.academia.edu/34575079/Study_of_Road_Accident_Prediction_Model_at_Accident_Blackspot_Area_A_Case_Study_at_Selangor

10. Santos, D., Saias, J., Quaresma, P., & Nogueira, V. B. (2021). Machine learning approaches to traffic accident analysis and Hotspot Prediction. Computers, 10(12), 157. https://doi.org/10.3390/computers10120157

11. Shetty, R. R., & Liu, H. (2021). Accidents Analysis and severity prediction using machine learning algorithms. Lecture Notes in Operations Research, 173–183. https://doi.org/10.1007/978-3-030-90275-9_15

12. Saini, A., Gauba, N., Chawla, H., & Ali, J. (2021). Road Accidents Analysis Using Comparative Study & Application of Machine Learning Algorithms. WSEAS TRANSACTIONS ON COMPUTER RESEARCH, 9, 78–86. https://doi.org/10.37394/232018.2021.9.9

13. Hala, H., Anass, C., Rajaa, B., Youssef, B., & Garza-Reyes, J. (2021). Machine learning techniques for forecasting the Traffic Accident Severity. 2021 International Conference on Digital Age & Technological Advances for Sustainable Development (ICDATA). https://doi.org/10.1109/icdata52997.2021.00018

14. Kashyap, N., Malali, H. R., S. E, K., G, R., & Sreenivas, T. H. (2020). Traffic accident injury and severity prediction using machine learning algorithms. Lecture Notes in Electrical Engineering, 1041–1048. https://doi.org/10.1007/978-981-15-7961-5_96

15. Babaoglu, L., & Babaoglu, C. (2021). Prediction of fatalities in vehicle collisions in Canada. Promet - Traffic&Transportation, 33(5), 661–669. https://doi.org/10.7307/ptt.v33i5.3782