National College of Ireland

Predicting Chicago Crash Severity Using Machine Learning Algorithms and Identifying Influential Factors

MSc Research Project
Data Analytics

Joseph Agoi
Student ID: x22121684

School of Computing
National College of Ireland

Supervisor: Dr. Anu Sahni

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Joseph Agoi
Student ID:	x22121684
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Dr. Anu Sahni
Submission Due Date:	14/12/2023
Project Title:	Predicting Chicago Crash Severity Using Machine Learning Algorithms and Identifying Influential Factors
Word Count:	7,035
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Joseph Agoi
December **Date:**	31st January 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predicting Chicago Crash Severity Using Machine Learning Algorithms and Identifying Influential Factors

Joseph Agoi
x22121684

Abstract

This study investigates the performance comparison of diverse machine learning algorithms in predicting accident severity in Chicago, drawing from crash records spanning 2021 - 2022. To achieve the research goal, Random Forest, Support Vector Machine, and Binary Logistic Regression machine learning algorithms were used to develop predictive models for accident severity. The Random Forest Classifier was used to assess the significance of each factor and sub-factor, evaluating their weights and contributions to crash severity. The evaluation of the models built incorporated accuracy, precision, recall, and F-1 score metrics. The findings of the research show Random Forest is the best-suited model for severe crash prediction with an accuracy score of 71.78%. Support Vector Machine had an accuracy score of 69.65% while the Binary Logistics Regression Model had an accuracy score of 69.52%. Spatial and temporal factors were more prevalent in severe crash incidents.

Keywords **- Machine Algorithms, Random Forest, Support Vector Machine, Binary Logistic Regression, Data Preprocessing**

1 Introduction

Road crash severity is a significant public safety issue, causing significant loss of life and injuries. According to the World Health Organization [1], annual road traffic accidents cause about 1.3 million deaths and 20 to 50 million injuries that leave permanent disabilities among survivors Liu et al. (2018). Understanding the factors influencing crash severity is crucial for developing targeted interventions, assessing the effectiveness of existing safety measures, and guiding the development of evidence-based policies.

Factors influencing crash severity vary and can include infrastructure design, traffic control measures, driver behavior, road condition, vehicle type, and weather condition Safari et al. (2020). It is imperative to investigate how various traffic-related factors correlate with the severity of road crashes.

This study investigates the performance comparison of various machine learning algorithms in predicting accident severity in Chicago. It also investigates the link between

[1] https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

elements related to traffic accidents and the severity of road accidents in Chicago. It uses an authentic dataset, ethically procured from Chicago's transport department, to develop predictive models using machine learning algorithms, identify significant factors, quantify their impact, and provide evidence-based recommendations for improving road safety. The findings provide insights into the factors contributing to crash severity.

The research questions stem from preliminary work done by Santos et al. (2021) and showed inconsistency in their initial tests. In their future work section, they stressed that more studies were required as well as the use of a big dataset to arrive at firm conclusions. Taking advantage of a more recent and comprehensive dataset, this paper attempts to answer these two questions.

RQ: *How do different machine learning algorithms compare in predicting the severity of road crashes in Chicago?*

SubRQ: *Which traffic-related factors and sub-factors have a significant impact on crash severity in Chicago?*

Objective	Details
Objective 1	Performance Comparison of Machine Learning Algorithms
Objective 2	Analysis of Influential Factors on Crash Severity

Table 1: Research Objectives

Table 1 above shows the summary of the key objectives set for this study.

The research makes a significant contribution by building models that predict the severity of crashes. The models can be used by city policymakers to make well-informed decisions about allocating resources, such as emergency response teams, law enforcement, and infrastructure improvements, to areas that are more susceptible to severe accidents.

Other notable contributions of the project encompass gaining insight into the intricate dynamics of traffic crashes. This was achieved by conducting spatial and temporal analysis, which unveiled valuable patterns and trends.

The rest of the technical report is structured as follows: **Chapter 2**: **Related Work**. This section will have a critical analysis of the recent related work done, identifying gaps. **Chapter 3**: **Research Methodology**. Introduces the modified research methodology used to complete the project. **Chapter 4**: **Design Specification**. Will capture the design process flow. **Chapter 5**: **Implementation**. Model building process **Chapter 6**: **Evaluation**. Provide details of the evaluation metrics and results of the two supervised machine learning algorithms used in the project. **Chapter 7**: **Conclusion and Future Work**. Conclude the evaluation outcome and identify the gaps to be explored for future projects.

2 Related Work: Road Crash Severity Factors

2.1 Introduction

This section looks closely at the recent studies on the various factors that impact the severity of road crashes. It also sheds light on the gaps and limitations found within the existing body of knowledge, which served as the impetus for conducting further research in this area.

2.2 Accident Characteristics

Understanding risk factors contributing to traffic crashes is crucial for planning and developing countermeasures for road safety, especially on rural roads Xie et al. (2020). Common risk factors include distracted driving, speeding, and poor road conditions. Implementing stricter laws, improving infrastructure, and addressing these factors can significantly reduce accidents and promote road safety.

Safari et al. (2020) conducted a comprehensive study to identify factors contributing to the severity of car crashes. While their research provided valuable insights, it was subject to certain limitations, including language restrictions and the scope of the literature papers they reviewed. However, their findings highlighted key factors such as age, sex, safety belt usage, alcohol and drug involvement, speed, weather and lighting conditions, time of day and week, vehicle type, road conditions, collision type, and crash location as significant contributors to crash severity. Given the limitations and the specific context of their study, there is a need to explore the applicability and relevance of these findings in a different geographical setting. This research aims to address this gap by focusing on the municipality of Chicago and utilizing a dataset specific to this region. The study will also seek to determine if the same factors identified by Safari et al. (2020) hold true when applied to the Chicago dataset.

A study to examine factors affecting accident severity in urban river-crossing tunnels in Shanghai by He et al. (2018), uses 12 factors, including vehicle type, tunnel length, speed limit, accident occurrence time, and weather. A binary logistic regression model was used to analyze the correlations and showed that traffic volume, number of vehicles involved, accident occurrence place, and type were the most significant influences on accident severity. While commendable, the results were confined to river-crossing tunnels and based on accident data for 2011 that had a volume of 12,000 records. This study builds on their findings by applying and expanding upon their methodology but with a significantly larger and more recent dataset. The dataset encompasses approximately 210,000 records, providing a much broader scope and allowing for a more extensive analysis of factors influencing accident severity. In addition to the binary logistics model, this research incorporates the application of Support Vector Machine and Random Forest models to explore the predictive capabilities of these machine learning techniques. Given these distinctions, this research seeks to assess whether similar results and conclusions can be drawn when examining factors influencing accident severity in the broader context of Chicago's traffic environment.

The following subsection covers the relation of road factors to crash severity.

2.3 Road Factor and Crash Severity

Road crashes are a major cause of preventable deaths worldwide, largely due to poor road design, inadequate signage, and lack of maintenance Trivedi and Shah (2022). Wu et al. (2022) found that road alignment impacts driving behavior, speed, and maneuverability. Studying alignment and crash severity can optimize road design, maintenance, and countermeasures, enhancing safety.

In a study conducted by Musa et al. (2020) on the impact of roadway conditions on accident severity on federal roads in Malaysia, they established that poor horizontal alignment significantly reduced the likelihood of serious accidents, suggesting local authorities should take proactive measures. The study analyzed 1067 cases from 2008 to 2015 using public work and police department databases. While the study is valuable and provides guidance for road safety measures in the context of federal roads in Malaysia, it is important to see if a similar conclusion can be drawn to a different geographical location. To ensure the relevance and applicability of findings in the current Chicago setting, it is necessary to conduct a study using more recent and localized data.

The subsection that follows discusses the relationship between vehicle factors and accident severity, focusing on speed, vehicle type, and technological advancements.

2.4 Impact of Vehicle Factors on Crash Severity

Speed and type of vehicle significantly impact crash severity, with higher speeds leading to more severe injuries and fatalities. Job and Brodie (2022) acknowledges the existence of multiple sources of evidence that address speed as a factor but notes variations in the estimates of the role of speed among the sources. This study aims to examine speed and speeding in Chicago crash severity.

Jehle et al. (2021) found that vehicle type significantly impacts head-on crash severity, highlighting a gap in safety ratings. Different vehicle types exhibit different accident severity factors Eboli et al. (2020). This study aims to explore the impact of vehicle type on crash severity in the Chicago context, considering crash dynamics and structural characteristics.

The advancement in vehicle safety technology is providing drivers with an additional layer of road protection. According to Alicioglu et al. (2022), the automotive industry is enhancing safety through intelligent technologies using sensor-based data, reducing accident severity and improving vehicle safety, making it a crucial factor for car buyers. Standard features like lane departure warning systems, automatic emergency braking, and adaptive cruise control help prevent human error-related accidents. As more advanced safety measures are introduced, accident severity rates are expected to decrease further.

The subsection below discusses studies on environmental factors and how they relate to severe accidents.

2.5 Studies on Environmental Factors

Environmental factors significantly influence car crash severity, including weather conditions, road design, visibility, and built environment. Understanding these relationships is crucial for improving road safety measures.

Several studies have examined the relationship between environmental factors and accident severity. Accident location is significantly associated with fatal accidents Eboli et al. (2020). Dezman et al. (2016) conducted research in Baltimore, Maryland, and found that socioeconomic indicators alone could not adequately explain the geographical distribution of crashes in the area. Guo et al. (2018) investigated cyclist safety in Baltimore and discovered positive relationships between crashes and factors such as household density, commercial area density, and signal density. Chen and Zhou (2016) explored the role of the built environment in pedestrian crashes in Seattle, uncovering connections between intersection density, land use mix, and the frequency and danger of pedestrian crashes. Bayiga-Zziwa et al. (2023), however, found that the presence of intersections had a lesser impact on pedestrian injury risk in Kampala, Uganda, but agreed with Chen and Zhou (2016) regarding the significance of land use combinations.

Weather conditions also play a vital role in accident frequency and severity. Das et al. (2021) highlighted that curves with wet pavement pose risks as drivers may not adequately adjust their driving behavior. Budzyński and Tubis (2019) emphasized differentiating between wet road surfaces with and without precipitation. Weather-related factors, including rain, snow, ice, and fog, have been shown to increase the likelihood and severity of accidents. Xi et al. (2019) revealed that rear-end collisions are 1.53 times more likely during bad weather conditions. Interestingly, a study conducted by George et al. (2017) to investigate road accident severity per vehicle type in Greece, established that good weather conditions are associated with increased accident severity.

Despite the existing research, there are still gaps and opportunities for further investigation. The previous studies focused on specific locations such as Baltimore and Seattle, and there is a need to examine the impact of environmental factors on accident severity within the Chicago context.

The subsequent subcategory delves into the correlation between human elements and the level of severity in crashes.

2.6 Studies on Human Factor in Relation to Crash Severity

Human factors like driver mistakes, tiredness, inattention, and impairment substantially heighten the likelihood of severe harm or death in traffic mishaps. Understanding the influence of these human factors on collision severity is vital for devising impactful measures and tactics to enhance road safety, curbing the tragic aftermath of car crashes.

Adanu et al. (2017) study highlights the role of human factors, such as driver error and fatigue, in car crashes. They identified factors like unemployment, seatbelt failure, older age, fatigue, and unlicensed drivers as contributing factors to serious injury crashes. In contrast, Eboli et al. (2020) found that young, inexperienced male drivers were more prone to errors and risky behaviors, contributing to a higher incidence of crashes among

this demographic. This contrasts with Adanu et al. (2017), which highlighted older age as a factor associated with serious injury crashes.

The review of the literature shows that several studies have been carried out on human factors linked to accident severity. By exploring the similarities and differences in findings across studies, this research can provide a more comprehensive understanding of the specific human factors that influence car crash severity in the Chicago area. The literature review on machine learning for crash severity prediction will be highlighted in the next subsection.

2.7 Machine Learning for Crash Severity Prediction

In the study conducted by Santos et al. (2021), various classification algorithms, including decision trees, random forests, logistic regression, and naive Bayes, were evaluated for their ability to classify accident severity and predict crashes. Although the results demonstrated excellent prediction accuracies for both algorithms, there was a notable absence of investigating the influence of different data properties on algorithm performance. Specifically, factors such as weather conditions or traffic control device conditions were not thoroughly examined.

Therefore, one significant contribution of this research is to bridge this gap by conducting a comprehensive investigation into how the random forest algorithm and support vector machine, which obtained the highest score in performance evaluation, would operate within diverse data contexts. In order to enhance crash severity prediction accuracy, an additional focus will be placed on incorporating weather data into the analysis. This expanded study aims to shed light on any potential improvements that can be made when considering these additional data elements.

Wu and Hsu (2021)'s study delved into the realm of machine learning techniques, focusing specifically on ensemble tree-based methods like random forest (RF) and gradient boosting regression tree (GBRT). The results illuminated the fact that these techniques outshine the use of a singular decision tree in terms of reliability and outcomes. This investigation proposes that ensemble methods possess enhanced accuracy due to their collective decision-making abilities, as they effectively minimize bias and variance within the models. While these findings are commendable, it is equally important to examine how another commonly used algorithm, support vector machine, performs against ensemble tree-based machine learning techniques—particularly considering its effectiveness in high-dimensional spaces and tolerance to outliers.

2.8 Conclusion

Various factors contributing to accident severity have been explored in the studies analyzed for this review. These factors encompass different elements such as human factors, vehicle factors, environmental factors, and road factors. Although considerable progress has been made in studying these severe accident-causing factors, there are still certain areas that require further development. These gaps offer opportunities for further research and investigation to gain a better understanding of the underlying causes of severe accidents in Chicago.

3 Research Methodology

3.1 Introduction

This study employs a structured research methodology, ensuring the integrity, accuracy, and reliability of its scientific inquiry through a comprehensive presentation of its sequential steps.

Outline in this section are details of systematic procedures for data collection, analysis, model building, and interpretation, highlighting their importance in facilitating robust and credible scientific investigations.

This overview explains how the chosen methodology aligns with research objectives, ensuring valid outcomes. It explores the implementation process, highlighting the importance of meticulous measures for rigor and validity, enhancing the credibility and robustness of research findings.

3.2 Modified Methodology Approach Used

This study introduced a novel scientific methodology to uncover traffic-related factors impacting the severity of road crashes in Chicago and established machine learning predictive models that effectively predict crash severity based on these factors.

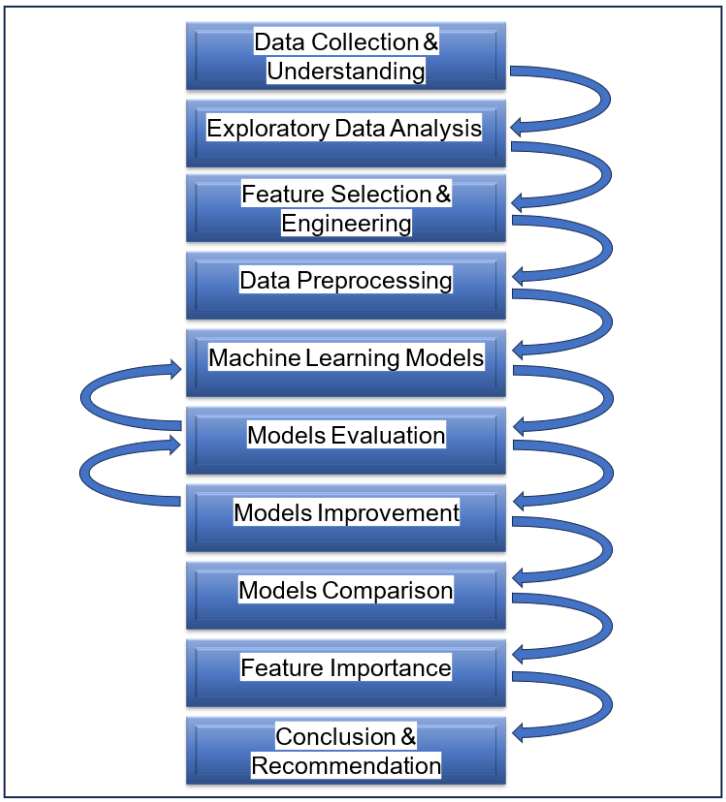

Figure 1: Modified Scientific Methodology used

Data Collection and Understanding: The first step in this research methodology

involved making an API request to [2]. This API call retrieved a dataset containing all the recorded observations from January 2021 to December 2022. The dataset encompasses various variables, such as Crash Date, Location, Posted Speed Limit, Traffic Control Device, Device Condition, Weather Condition, Lighting Condition, First Crash Type, Trafficway Type, Road Defect, and Crash Type.

Imported the dataset into a Python data frame and carried out a series of preliminary data explorations to gain an understanding of its underlying structure, formatting, and the various types of information it encompasses.

Using the panda's data frame info function to print information on the data frame to showcase all available features, the corresponding object types, and counts of non-null entries in each variable was done. Descriptive statistics was also printed to gain an understanding of the distribution of numeric features in the dataset. The shape function gives an overview of the data frame outlook in terms of rows and columns.

Exploratory Data Analysis (EDA): Detailed analysis and visualization of the data were conducted at this stage. Seasonalities in the crash dataset were checked at this step. A three-dimensional heat map was generated to show the correlation between accident crashes, the hour of the crash, and the day of the week.

Feature Selection and Feature Engineering: In a review of the risk factors affecting the crash severity, Safari et al. (2020) narrowed down the risk factor to 5 categories; Human, Environmental, Vehicle, Road, and Accident Characteristics. This study selected the feature variables based on these five categories. Features that fell outside the five categories were discarded.

Categorical variables were converted into numeric formats suitable for regression machine learning models. Location details used to identify specific areas or neighborhoods were also captured by the use of the latitude and longitude coordinates.

Data Preprocessing: During the data preprocessing stage, one of the crucial steps was to thoroughly clean and prepare the dataset to ensure its seamless utilization in subsequent stages. Null values in any of the remaining feature variables were dropped. The dropna function was used to achieve this.

Machine Learning Model: The dataset was separated into training and testing sets in order to create machine learning models. These models were then used to predict crash severity. Random Forest, Support Vector Machine, and binary logistic regression models were employed in the creation of these predictive models. The identified traffic-related features were used as the feature variables while the crash severity variable was set as the target variable.

Models Evaluation: At this step, the models' performances were evaluated using metrics such as accuracy, precision, recall, F1-score, confusion matrix, and classification report.

[2]https://data.cityofchicago.org/

Accuracy measures model performance by dividing correct predictions by total predictions. A score of 1 or near 1 indicates a perfect fit. It's not suitable for imbalanced datasets where one class dominates others.

Precision is the proportion of correctly predicted positive instances, indicating accuracy. High precision indicates low false-positive rates, crucial in high-cost scenarios.

Recall measures the model's ability to accurately predict positive classes, particularly in high-cost situations by comparing true positives to actual positives.

F1-score is a balanced metric that combines precision and recall, enhancing low values, useful in imbalanced classes or when both are crucial.

A confusion matrix summarizes a classification model's performance by displaying true positives, false positives, true negatives, and false negatives for each class, aiding in error visualization and comparison.

Fine-tuning and Models Improvement: Fine-tuned the models' hyperparameters using GridSearchCV from Scikit-Learn to find the best combination of parameters for the models, such as the number of trees, the maximum depth, and the criterion. 5-fold cross-validation was also applied to optimize the models' performances.

Comparison of Models Performance: Evaluation metrics were used to measure each model's predictive capabilities. The results were analyzed, and the best model was selected based on the highest score or lowest error, considering factors like complexity, interpretability, and scalability.

Feature Importance and Correlation Analysis: The correlation matrix and RandomForestClassifier were used to assess the impact of various characteristics on crash severity. These tools identified patterns and connections among features, revealing which ones have a stronger influence.

Recommendations and Conclusion: Based on the extracted insights, the project draws recommendations for traffic safety measures that can prevent or reduce the severity of road crashes. The recommendations emphasize the influential traffic-related factors.

3.3 Conclusion

The scientific principles were upheld in this study in the improvised framework used. To address the research question, the study involved a hybrid approach, incorporating elements from various methodologies and aligning with the general scientific process. The process included data collection, exploratory data analysis, feature selection, engineering, preprocessing, model implementation, training, evaluation, feature importance analysis, and model performance comparison. This systematic progression ensures scientific rigor and standards, demonstrating a thoughtful and scientifically sound approach to analyzing traffic-related factors and accident severity.

4 Design Specification

This section details the research project's design specification, including techniques, architecture, and requirements. It outlines data collection, analysis, and modeling using Python and Scikit-Learn for traffic-related factors and crash severity in Chicago. The section introduces the research design, data sources, analysis techniques, evaluation metrics, and design specifications.

4.1 Design Process Flow

The design process flow captures the graphical representation of the steps involved in the study, from the initial step of data collection to the final step of visualization of the output. It captures how various tools and techniques were employed at each critical step. The design process flow also illustrates the inputs, outputs, and dependencies of each step, as well as the expected outcomes and deliverables of the project. The following figure shows an overview of the design process flow.

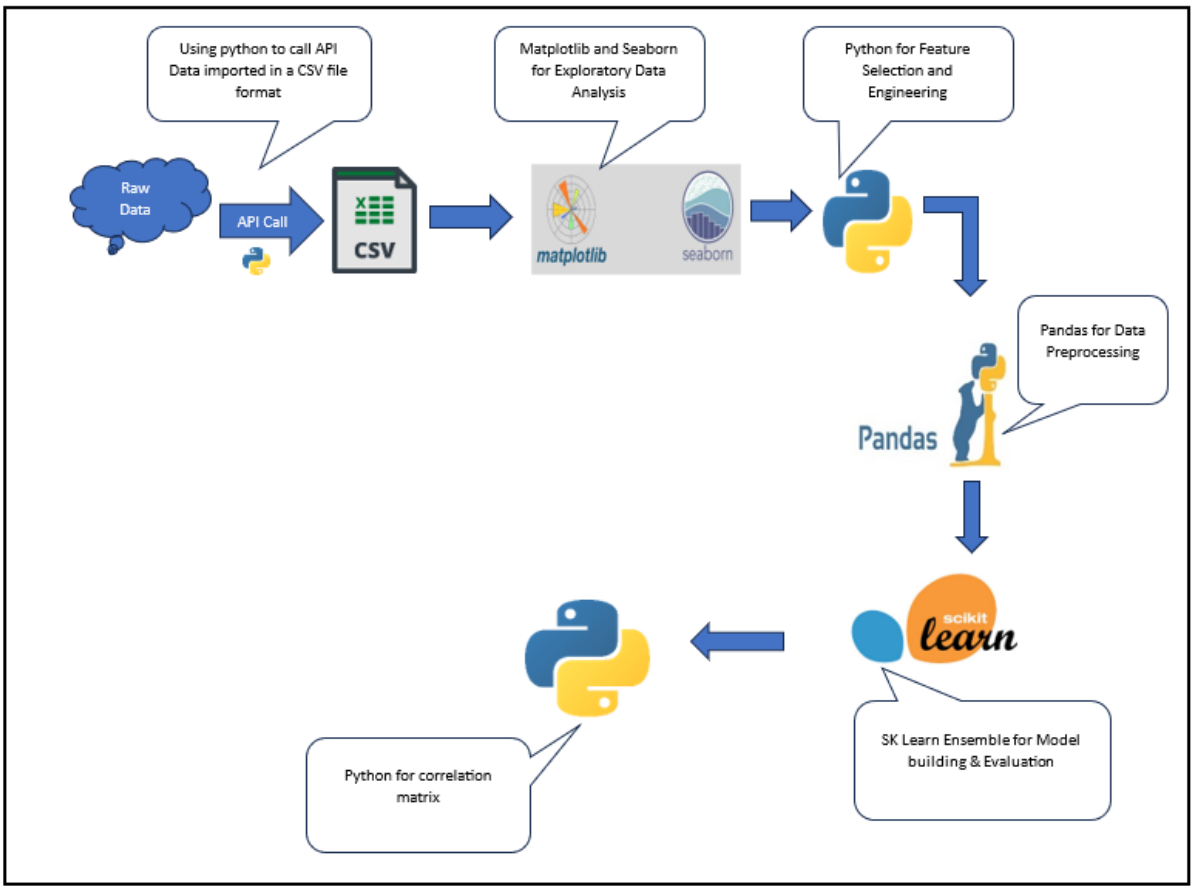

Figure 2: Design Process Flow

Figure 2 shows the process of extracting data from [3] to visualization of the output.

[3] https://data.cityofchicago.org/

4.2 Techniques Selection

The study used Random Forest, support vector machine(s), and logistic regression machine learning algorithms. The selection of these algorithms was done due to their abilities to handle classification tasks conveniently to predict the injury severities using historical patterns. A Random Forest is an ensemble of multiple randomized trees whose output combines to form the overall prediction. It's known for its high accuracy and robustness. Logistic regression is a linear classifier based on a logistic function to estimate the probability of a binary outcome and is applied in many classification problems. The effectiveness of a support vector machine derives from its ability to identify a hyperplane that separates two classes while keeping the corresponding separation distance as large as possible.

4.3 Framework Choice

To implement this project, Python-based machine learning technologies, like Scikit Learn, Matplotlib, Seaborn, and Pandas, were incorporated. Scikit-learn offers many tools related to machine learning problems including data preparation, model selection, model assessment, and model deployment. Visualization was achieved using Matplotlib and Seaborn libraries. Data manipulation and preprocessing such as reading, cleaning, merging, and transforming was facilitated by pandas. Microsoft Office suite including Excel, CSV was also used to store the historical data.

4.4 Requirements Identification

Access to comprehensive data on the Chicago crash from 2021-2022 is crucial for this project. The dataset was collected from the Chicago Data Portal [4], an online platform that provides open information on the city's data. The dataset has 217,149 recorded traffic crash incidents Including variables such as crash date, crash time, crash location, severity type, weather condition, lighting condition, road defect information, road alignment, and posted speed limits, amongst others. Cumulatively, the data set contains 49 variables with over 217,149 observations.

The project also encompasses an environment of computing that has Python and adequate memory and processing speeds to accommodate the models' training and assessments. The integrated development environment is Jupyter Notebook.

4.5 Integration Plan

A comprehensive pipeline consisting of chosen machine learning algorithms was set up. The models were applied on data that had undergone preprocessing stages such as feature scaling, encoding of categorical variables, and missing value detection. The modeling process involved data splitting, model training, hyperparameter tuning using GridSearchCV, and model evaluation using appropriate performance metrics. The data were split into training and testing sets, with a ratio of 85: 15. The training data was used for training models which in turn, were tested using the test set. Accuracy, Precision, Recall, and F1-Score measures were selected as performance metrics in this study.

[4]https://data.cityofchicago.org/

5 Implementation

5.1 Introduction

This section provides a detailed account of the implementation of the proposed solution for analyzing the crash severity data in Chicago and identifying how different traffic-related factors impact the severity level of traffic accidents. It also describes the models building process for the crash severity prediction algorithms. The implementation section focuses on the final stage of the solution, encompassing several key sub-sections: Tools and Languages Used, Transformed Data, Models Developed, and Models Improvement. These sub-sections will delve into the specific tools, techniques, and methodologies employed to transform the data, develop predictive models, and enhance their performance.

5.2 Data collection

The data collection process played a crucial role in obtaining the necessary crash severity data for our research. The process was initiated by making an API call to the online portal endpoint of the city of Chicago's Department of Transportation [5]. This API call gave access to the desired dataset, which encompassed a comprehensive record of traffic crash incidents that occurred between January 1, 2021, and December 31, 2022. Upon successful retrieval, the dataset was downloaded in CSV format and stored on a local PC for further analysis. The dataset contained a substantial amount of information, capturing a total of 217,149 recorded traffic crash incidents. Each incident was associated with 49 different features, providing a rich and diverse set of variables to explore.

The CSV file was imported into a Pandas data frame in Python for data manipulation and transformation. Initial exploratory analyses and data cleaning procedures were performed to ensure data integrity and reliability. This involved examining the dataset's structure. Python programming language and Microsoft CSV were used for API calls and dataset downloads. Python was also used for data manipulation and analysis specifically the Pandas library. Jupyter Notebook environment was used for code development, visualization, and exploratory analysis.

5.3 Transformed Data

In the Transformed Data section, the focus was on refining the initial dataset by selecting relevant variables and performing preprocessing and feature engineering tasks. Building upon the work done by Safari et al. (2020), the study narrowed down the 49 different features attributed to the crash incidents to 17 variables that align with the research objectives.

Several variables were discarded as they did not align with the study's goals. These included variables such as Chicago Police Department report number, Crash Record ID, Hit and Run, Report Type, and others that were not directly relevant to the research question. Variables with a high percentage of null values, such as Lane Count, Intersection Related, Not Right of Way, Dooring, and Crash Date Estimated, were also dropped

[5] https://data.cityofchicago.org/resource/85ca-t3if.csv

as they would not provide sufficient information for the models.

The selected variables for addressing the research question were Posted Speed Limit, Traffic Control Device, Device Condition, Weather Condition, Lighting Condition, First Crash Type, Trafficway Type, Alignment, Road Surface Condition, Road Defect, Crash Type, Most Severe Injury, Crash Hour, Crash Day of Week, Crash Month, latitude, and longitude. These variables were saved into a new data frame named new_df.

Data preprocessing and feature engineering tasks were then applied to the new_df data frame. Null values in latitude and longitude were dropped, ensuring data integrity. Binning of the Posted Speed Limit variable was performed to denote low, medium, and high speeds and a Day-Night Indicator was derived based on the Lighting Condition column values.

Categorical values in features such as Trafficway Type, Alignment, Road Surface Condition, Road Defect, Crash Type, Most Severe Injury, Traffic Control Device, Device Condition, Weather Condition, Lighting Condition, and First Crash Type were encoded for further analysis. The resulting transformed data frame was saved as clean_df, which was then ready for the subsequent model-building step.

The data preprocessing and feature engineering tasks were primarily conducted using Python's pandas library. The library provided the necessary functionalities and tools to manipulate, clean, and enhance the data, preparing it for the subsequent modeling phase.

5.4 Model Development

During the model development phase, three models were built to predict crash severity: Random Forest, Support Vector Machine (SVM), and Binary Logistic Regression. Each model underwent data preparation, splitting, initialization, training, prediction, evaluation, and results presentation. Here is a detailed breakdown for each model:

5.4.1 Random Forest

Data Preparation: In this step, the features (X) and the target variable (y) were defined based on the clean_df dataset. Selected columns with presumed predictive value for crash severity were assigned to features_cls, while the target variable column in numeric format was assigned to target_cls.

Data Splitting: The clean_df dataset was split into training and testing sets using the train_test_split function from scikit-learn's model_selection module. The data was divided into 85% for training and 15% for testing, allowing for the assessment of the model's performance on unseen data.

Model Initialization and Training: A Random Forest Classifier model was instantiated using the RandomForestClassifier class from the sklearn.ensemble module. Hyperparameters like the number of decision trees (n_estimators) were set to 100, ensuring a robust model. The random state was fixed to 42 for reproducibility. The model was then trained using the training data (X_train_cls, y_train_cls) with the fit method

(rf_classifier.fit).

5.4.2 Support Vector Machine

Data Preparation: Similar to Random Forest, the data is classified using features (X) and target variable (y), with features representing crash severity-related variables and target variable representing crash type.

Data Splitting: The cleaned_df dataset was then divided into training and testing sets using train_test_split from scikit-learn's model_selection, with the ratio being 85:15. Model evaluation involved using the testing set.

Model Initialization and Training: Initialized a Support Vector Machine Classifier (SVC) model with SVC from sklearn.svm. Proceeded to set the hyperparameter values such as C (the regularisation parameter), kernel (the type of kernel), gamma (the kernel coefficient), and random_state (for reproducibility). Finally, trained the model on the training data (X_train_cls, y_train_cls), using the method of svm classification (svm_classifier.fit).

5.4.3 Logistic Regression

Used the same process flow for data preprocessing and data splitting. Initialized and trained the logistic regression model (LogisticRegression) on the training data. Hyperparameter configurations done were: C (Inverse of regularization strength): for weaker regularization, allowing the model to fit the training data more closely, max_iter (Maximum Iterations): which sets the maximum number of iterations for the optimization algorithm to converge, and solver: which specifies the optimization algorithm used to fit the logistic regression model.

5.5 Models Improvement

To improve the performance of the machine learning models, the study employed GridSearchCV, a function provided by Scikit-Learn, to identify the optimal hyperparameter values for individual models. GridSearchCV systematically searches through different combinations of hyperparameter values to identify the optimal configurations. It performs a cross-validation procedure to estimate the model's performance for each combination of hyperparameters, allowing the researchers to compare and select the best set of values.

The identified optimal hyperparameter values were thereafter infused into the existing code of each model, replacing the default hyperparameter values. The modified codes were then rerun causing the model to be retrained with new hyperparameters that yielded better results. The new metric scores obtained after incorporating the optimized hyperparameters served as a measure of the enhanced capabilities of the models.

5.6 Identification of key Factors Influencing Crash Severity

This study had a second objective of investigating some determinants of crash severity in Chicago. A random forest classifier was used for its capabilities in capturing complex

relationships and patterns in the data. The analysis used a data set comprising of crash characteristics such as Crash hour, Posted speed limit, and weather conditions. To prevent bias, the dataset was cut into training and test sets with a ratio of 80-20, so that its evaluation could be objective.

Initializing and training the Random Forest Classifier with the provided features from the training set made the Classifier learn from them. After the model's training, the Random Forest Classifier was used to calculate the feature importances. The importance of each feature can be estimated by measuring such relative contributions it makes towards the intensity of a possible accident (crash severity). The feature importances were thus plotted into a graph that facilitated an easy interpretation of what factors mattered most in determining the seriousness of the crashes.

5.7 Identifying Most and Least Influential Sub-Factors for Severe Crashes

The process began by isolating the relevant columns (PRIM_CONTRIBUTORY_CAUSE and CRASH_TYPE) from the dataset to focus specifically on the primary contributory cause and crash type attributes. The PRIM_CONTRIBUTORY_CAUSE column, containing categorical values, was selected for feature encoding to transform categorical variables into a suitable format for machine learning analysis. The One-Hot Encoding technique (OneHotEncoder) was applied to convert categorical values into a numerical format for model training.

The dataset was the divided into training and testing sets using a 20% test size, with 80% of the data used for training the Random Forest Classifier.

A Random Forest Classifier, a robust ensemble learning method, was initialized with 100 decision trees (n_estimators=100) to capture relationships and patterns between the encoded features and crash types. The classifier was trained on the training dataset (X_train and y_train) to identify influential sub-factors impacting severe crashes.

After model training, feature importance was extracted using the feature_importances_ attribute of the Random Forest Classifier. The resulting importance scores were then sorted in descending order (ascending=False) to identify the top 10 most influential sub-factors contributing to severe crashes. This information was printed and visually represented in a horizontal bar plot using Matplotlib, aiding in easy interpretation and visualization of the most impactful causes.

Similarly, the process continued by analyzing the least influential sub-factors for severe crashes. Feature importance scores were sorted in ascending order (ascending=True), and the bottom 10 least influential causes were printed and visualized through another horizontal bar plot, highlighting the less impactful attributes.

6 Evaluation

In this section, we assess the accuracy of the three machine learning models (random forest, support vector machine, and binary logistic regression) in predicting accident severity in Chicago. The section also evaluates the impact of various factors and sub-factors in influencing crash severity in Chicago.

6.1 Experiment 1: Models before Hyperparameter Tuning

The first experiment involved the training and evaluation of the three models using the default hyperparameter settings. The results were the baseline score for the second experiment that was carried out after optimization was done using GridSearchCV and the best hyperparameters were identified.

Four measures (accuracy, F-1, precision, and recall) were applied to assess the abilities of the ML algorithms in the prediction of accident severity level. Accuracy denotes the number of predictions that are right out of total predicted cases, while precision denotes the number of correctly predicted positives out of all positive predictions. Further, recall denotes the proper prediction of positives out of actual positives and the F-1 score is the average of precision and recall.

The mathematical expressions for these evaluation measures are presented as Eqs. 1-4 below:

$$Accuracy = \frac{TP + TN}{TotalPredictions} \tag{1}$$

where TP is the number of true positives while TN is for true negatives

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

FP is the number of false positives

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

FN is the number of false negatives

$$F1Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{4}$$

In these formulas: True Positives (TP) are instances correctly predicted as positive. True Negatives (TN) are instances correctly predicted as negative. False Positives (FP) are instances incorrectly predicted as positive. False Negatives (FN) are instances incorrectly predicted as negative.

Table 2 below shows the results of the initial experiment.

Algorithm	Accuracy	Precision	Recall	F-1 Score
Random Forest	0.7105	0.6809	0.7105	0.6741
Support Vector Machine	0.6955	0.4837	0.6955	0.5705
Logistic Regression	0.6944	0.6995	0.9826	0.8172

Table 2: Experiment 1 Metrics Scores

Table 2 above shows that the Random Forest and Logistic Regression models showed better performance in predicting crash severity during Experiment 1, outperforming the Support Vector Machine model. The Random Forest model had high accuracy and recall, while Logistic Regression excelled in recall and F-1 Score. The Support Vector Machine model had lower precision and higher false positive rates, impacting overall performance. The models' balanced performance in multiple metrics makes them promising choices for predicting crash severity.

6.2 Experiment 2: Models after Hyperparameter Tuning

Experiment 2 focuses on improving predictive accuracy and model performance by implementing hyperparameter optimization using the GridSearchCV methodology. The process involves searching through a parameter grid to identify optimal hyperparameter configurations for each machine-learning model. These parameters are then integrated into existing models, and the resulting evaluation scores show enhanced performance metrics.

Table 3 shows the summary of the evaluation scores in Experiment 2.

Algorithm	Accuracy	Precision	Recall	F-1 Score
Random Forest	0.7178	0.6921	0.7178	0.6707
Support Vector Machine	0.6965	0.6703	0.6965	0.5758
Logistic Regression	0.6952	0.6994	0.9850	0.8180

Table 3: Experiment 2 Metrics Scores

As can be seen from Table 3, Random Forest outperformed the other two algorithms in accuracy, with a score of 71.78%, indicating its capability to predict crash severity accurately. The Random Forest model also had a high score of 69.21% in Precision metric and a Recall score of 71.78%, displaying balanced performance in correctly identifying severe crash incidents and minimizing false positives. Logistic Regression has a better precision (69.94%) and recall (98.50%) than the rest. This indicates its ability to detect most major crashes but at the expense of potentially increased false positives.

6.3 Identification of Risk Factors

To address the second research question, this study focused on identifying and analyzing key risk factors that influence the prediction of crash severity outcomes. The feature importance function within the sklearn.ensemble framework was applied to identify factors influencing crash severity, revealing pivotal determinants within the dataset. Table 4

shows the summary of the key factors that influence crash severity.

Feature	Importance
Latitude	0.226
Longitude	0.224
Crash Hour	0.136
Crash Month	0.107
Crash Day of the Week	0.078
Trafficway Type	0.060
Posted Speed Limit	0.037
Roadway Surface	0.028
Traffic Control Device	0.024
Weather Condition	0.021
Device Condition	0.020
Road Defect	0.019
Lighting Condition	0.012
Road Alignment	0.008

Table 4: Key Crash Severity Factors

As can be seen from Table 4 and Figure 3, geographical location details, latitude and longitude, have a considerable impact on accident severity thereby indicating the importance of spatial factors in the study of crash severity patterns. The temporal aspects include the Crash hour (0.136), Crash month (0.107), and Crash day of the week (0.078), all exhibit moderate importance scores. This shows several factors that suggest the importance of time-specific variables in affecting the severity of the accidents with special mention of different hours, months, or weekdays likely to be associated with serious crash cases. Crash Severity was moderately or weakly associated with roadway characteristics (trafficway type, posted speed limit, roadway surface, traffic control device, road defect, and roadway alignment) and environmental factors (weather conditions and lighting conditions).

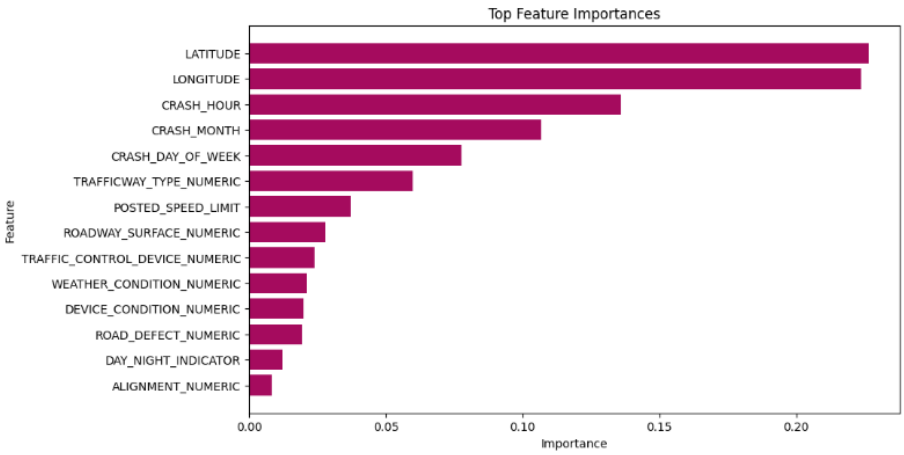

Figure 3: Key Crash Severity Factors Graph

6.4 Identification of Key Crash Severity Sub-Factors

The feature_importances_ function within the sklearn.ensemble framework was applied to compile the list of the significant sub-factors based on the documented values in the primary contributory variable. Through examining documented values within the primary contributory variable, this analysis aimed to identify and rank key sub-factors that significantly influence the severity of crashes in Chicago.

Sub-Factor	Importance
Disregarding Traffic Signals	0.189
Improper Braking	0.111
Failing to yield Right-of-Way	0.096
Unable to Determine	0.066
Under the Influence of Alcohol/Drugs (Use when arrest if effected)	0.064
Failing to reduce Speed to Avoid Crash	0.063
Following too Closely	0.062
Physical Condition of the Driver	0.055
Improper Overtaking/Passing	0.055
Equipment - Vehicle Condition	0.039

Table 5: Top 10 Crash Severity Sub-Factors

Table 5 and Figure 4, show that the most consequential sub-factors include disregarding traffic signals (18.9%), improper breaking (11.1%), and failures to yield right of way (9.6%). They confirm the importance of driver behavior in determining crash severity. Alcohol/Drug Influence (6.4%) and Physical state of a driver (5.5%) show how driver impairment and health condition contribute to the severity of accidents. Driving behavior factors such as failing to reduce speed, following too closely, improper overtaking, and passing contribute to increased crash severity at 6.3%, 6.2%, and 5.5% respectively. This indicates lack of caution and safe driving practices result in more severity in car crashes.

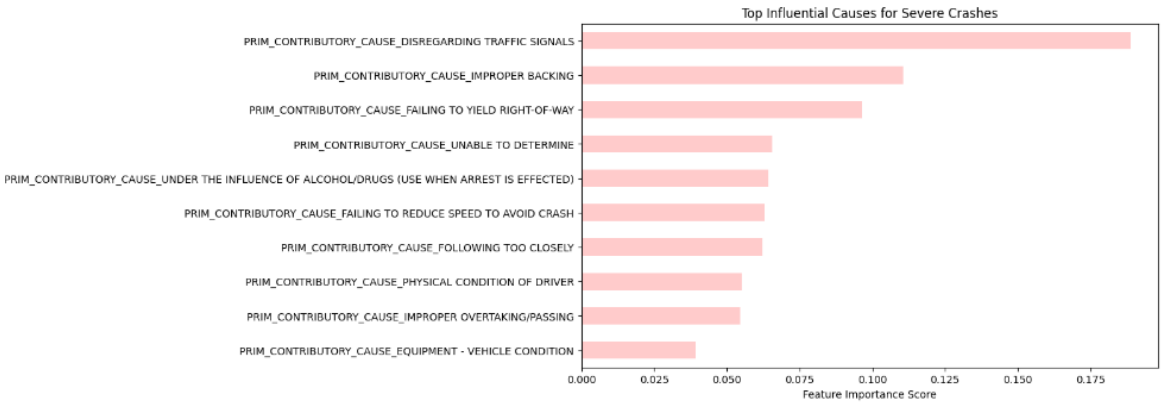

Figure 4: Key Crash Severity Sub-Factors Graph

6.5 Discussion

Experiment 1 evaluated three machine learning algorithms: Random Forest, Support Vector Machine (SVM), and Logistic Regression. Random Forest showed higher accuracy and recall but lower precision and F-1 Score. Logistic Regression excelled in precision, recall, and F-1 Score, accurately identifying severe crashes. SVM had a higher false positive rate.

Experiment 2 showed slight improvements in model performance post-hyperparameter optimization, with Random Forest showing slight increases in accuracy and recall, Logistic Regression maintaining high precision and recall, and SVM showing limited improvements, indicating model stability with minor performance alterations.

The analysis of feature importance revealed that Latitude and Longitude are the top determinants of crash severity, along with other factors like trafficway type and speed limit.

Sub-factors like disregarding traffic signals, improper braking, and failing to yield right-of-way, along with underage drinking and driving conditions, significantly contribute to crash severity outcomes.

The experiments highlighted the importance of hyperparameter tuning in improving marginal performance, while feature importance analysis revealed the multifaceted nature of crash severity, influenced by environmental and behavioral factors.

7 Conclusion and Future Work

This paper sought to examine the predictive analysis of various machine-learning algorithms and determine what factors influence crash severity in Chicago using a large and current data set. Three machine learning techniques (Random Forest, Support Vector Machine, and Logistic Regression) were employed for the prediction of crash severity and were evaluated using accuracy, precision, recall, and F1-score metrics. This study found that the Random Forest model performed better than the other two algorithms and had an accuracy score of 71.78%. Spatial and Temporal factors were prevalent in Crash severity incidents with human-related sub-factors, disregarding traffic signals and improper braking, leading the cause. The findings suggest that implementing more traffic enforcement campaigns is key to reducing road crashes in Chicago.

This study has limitations, including potential bias and noise in the dataset due to human and vehicle factors, and insufficient account for complex relationships among variables, affecting predictive accuracy and generalizability.

Future research should explore additional data sources like social media reports, Insurance Companies and Industry Reports, and the National Highway Traffic Safety Administration (NHTSA) amongst other sources to ensure inclusivity. Also, to be considered for future studies should be the Neural Networks and deep learning algorithms and evaluation metrics like the ROC curve and AUC score for more comprehensive and robust model performance.

References

Adanu, E. K., Jones, S. et al. (2017). Effects of human-centered factors on crash injury severities, *Journal of advanced transportation* **2017**.

Alicioglu, G., Sun, B. and Ho, S. S. (2022). An injury-severity-prediction-driven accident prevention system, *Sustainability* **14**(11): 6569.

Bayiga-Zziwa, E., Nsubuga, R. and Mutto, M. (2023). Factor analysis of community-ranked built environment factors contributing to pedestrian injury risk in kampala city, uganda, *Injury prevention* .

Budzyński, M. and Tubis, A. (2019). Assessing the effects of the road surface and weather conditions on road safety, *Journal of KONBiN* **49**(3): 323–349.

Chen, P. and Zhou, J. (2016). Effects of the built environment on automobile-involved pedestrian crash frequency and risk, *Journal of Transport & Health* **3**(4): 448–456.

Das, S., Geedipally, S. R. and Fitzpatrick, K. (2021). Inclusion of speed and weather measures in safety performance functions for rural roadways, *IATSS research* **45**(1): 60–69.

Dezman, Z., de Andrade, L., Vissoci, J. R., El-Gabri, D., Johnson, A., Hirshon, J. M. and Staton, C. A. (2016). Hotspots and causes of motor vehicle crashes in baltimore, maryland: A geospatial analysis of five years of police crash and census data, *Injury* **47**(11): 2450–2458.

Eboli, L., Forciniti, C. and Mazzulla, G. (2020). Factors influencing accident severity: an analysis by road accident type, *Transportation research procedia* **47**: 449–456.

George, Y., Athanasios, T. and George, P. (2017). Investigation of road accident severity per vehicle type, *Transportation research procedia* **25**: 2076–2083.

Guo, Y., Osama, A. and Sayed, T. (2018). A cross-comparison of different techniques for modeling macro-level cyclist crashes, *Accident Analysis & Prevention* **113**: 38–46.

He, W., Zhang, Z., Lu, L. and Wang, Z. (2018). Analysis on the influence factors of accident severity: evidence from urban river-crossing tunnels in shanghai of china, *Journal of Engineering Science and Technology Review* **11**(5): 100.

Jehle, D., Arslan, A., Doshi, C. and O'Brien, C. (2021). Car ratings take a back seat to vehicle type: outcomes of suv versus passenger car crashes, *HCA healthcare journal of medicine* **2**(4): 289.

Job, R. S. and Brodie, C. (2022). Road safety evidence review: Understanding the role of speeding and speed in serious crash trauma: A case study of new zealand, *Journal of road safety* **33**(1): 5–25.

Liu, G., Chen, S., Zeng, Z., Cui, H., Fang, Y., Gu, D., Yin, Z. and Wang, Z. (2018). Risk factors for extremely serious road accidents: Results from national road accident statistical annual report of china, *PLoS one* **13**(8): e0201587.

Musa, M. F., Hassan, S. A. and Mashros, N. (2020). The impact of roadway conditions towards accident severity on federal roads in malaysia, *PLoS one* **15**(7): e0235564.

Safari, M., Alizadeh, S. S., Bazargani, H. S., Aliashrafi, A., Maleki, A., Moshashaei, P. and Shakerkhatibi, M. (2020). A comprehensive review on risk factors affecting the crash severity, *Iranian journal of health, safety and environment* **6**(4): 1366–1376.

Santos, D., Saias, J., Quaresma, P. and Nogueira, V. B. (2021). Machine learning approaches to traffic accident analysis and hotspot prediction, *Computers* **10**(12): 157.

Trivedi, P. and Shah, J. (2022). Identification of road crash severity ranking by integrating the multi-criteria decision-making approach, *Journal of road safety* **33**(2): 33–44.

Wu, B., Zou, C., Li, Y., Fan, D., Zhu, S. et al. (2022). Impact of road environment on drivers' preference to merging location selection in freeway work zone merging areas, *Journal of advanced transportation* **2022**.

Wu, Y.-W. and Hsu, T.-P. (2021). Mid-term prediction of at-fault crash driver frequency using fusion deep learning with city-level traffic violation data, *Accident Analysis & Prevention* **150**: 105910.

Xi, J., Guo, H., Tian, J., Liu, L. and Sun, W. (2019). Analysis of influencing factors for rear-end collision on the freeway, *Advances in Mechanical Engineering* **11**(7): 1687814019865079.

Xie, S., Ji, X., Yang, W., Fang, R. and Hao, J. (2020). Exploring risk factors with crash severity on china two-lane rural roads using a random-parameter ordered probit model, *Journal of advanced transportation* **2020**: 1–14.