

**CAR CRASH REPORT ANALYSIS**

CNST 6308 / Data Analysis in Construction Management



Submitted By:

**Group – 40**

**Under the guidance of Dr. Lu Gao**

College of Technology University of Houston – Main Campus

# ABSTRACT:

Vehicles are one of the most important and as well as dangerous means of transportation, yet they pose significant risks. Vehicle accidents result in a significant number of deaths and have widespread consequences. In the United States alone, annual costs due to automobile accidents amount to $277 billion, which does not include the psychological toll on those involved and their families. This project aims to predict the severity of injury type in crash information by examining data from Missouri state. The analysis focuses on injury trends and how they relate to the use of safety devices in vehicles. With the projected increase in population over the years, the number of automobiles on the road is expected to rise, leading to more accidents. However, the growing usage of electronic vehicles will undoubtedly affect these trends. The target variable in this study is Injury Type, while the feature variables are safety devices, driver age, gender, county, location, damage, and disposition. To predict injury type accurately, a model must take all these factors into account. This project employs various machine learning models for analysis, including Naive Bayes classification, Decision Tree, Random Forest, K-Nearest Neighbors, and Multi-Layered Perceptron. The best model is chosen to predict the injury type for future events. To evaluate the model's performance, we used the accuracy score as a metric. The accuracy score reflects the ratio of correct predictions to the total number of predictions, with the highest accuracy value showing the best approach developed. Machine learning models enable the identification of valuable patterns through the processing of multiple classification parameters. These models are efficient in handling large amounts of data and can be scaled to cater to a range of research requirements. Additionally, artificial neural networks are an effective computational tool in the field of computer science as they mimic the human brain's central nervous system. With these networks, machines can learn to identify and analyze data in a way like humans, making them a valuable asset in data analysis.

1. **INTRODUCTION**:

In the United States, vehicle crashes are responsible for causing the highest number of fatalities and injuries, with approximately 2.5 million individuals being injured. Vehicle crashes have a devastating effect on not only the one involved in the accident, but it will also affect their families. One of the most significant global issues is the occurrence of road accidents, which can have far-reaching effects on a country's society, economy, and development. Traffic conflicts between vehicles are often the cause of road accidents, resulting in traffic delays and congestion. Driving behavior is influenced by numerous factors such as human error, traffic conditions, road conditions, vehicle conditions, and the surrounding environment, which can all contribute to the incidence of road accidents. The World Health Organization estimates that 1.35 million people die each year from road traffic accidents, and this number is projected to increase if effective measures are not implemented. Road safety campaigns and interventions have been introduced to reduce the number of road accidents, such as promoting the use of seat belts, enforcing speed limits, and improving road infrastructure. However, more efforts are needed to address this problem effectively.

The application of Machine Learning algorithms to road safety has become increasingly popular due to its potential in predicting and analyzing road accidents. These models are capable of processing large amounts of data related to road infrastructure, weather conditions, and driving behavior to identify influential factors that could lead to road accidents. By using Machine Learning models, researchers can predict the probability of road accidents and suggest effective interventions to prevent them. Therefore, the use of Machine Learning models in road safety is essential to reduce the number of road accidents and improve road safety. The impact of road accidents is significant, not only for the people involved but also for their families and society, making it a global issue that affects millions of people worldwide. The use of Machine Learning models presents a hopeful solution in addressing the issue of road accidents. These models can provide valuable insights into the significant factors that contribute to these accidents. As Machine Learning algorithms continue to evolve and become more sophisticated, interventions aimed at improving road safety are expected to become more effective, leading to a decrease in the number of road accidents. Therefore, ongoing research in the field of Machine Learning is crucial in developing more advanced models that can provide insights to policymakers and stakeholders, resulting in improved road safety.

The ability of Machine Learning algorithms to process numerous classification parameters and extract significant patterns from vast amounts of data makes them highly efficient and scalable. Computer scientists and related professionals use artificial neural networks, which simulate the central nervous system of animals, especially the brain, to enable machines to learn and identify information much like humans. Vehicle crashes result from a combination of factors, including driver errors, speeding, reckless driving, and weather conditions. This research project focused on analyzing accident data from Missouri state, which was obtained from a website using incident numbers. The researchers employed diverse methodological strategies to determine the factors that contribute to the severity of crashes. These strategies included traditional statistical models and modern data mining techniques. Researchers used different methods to find out what factors make car crashes worse, like statistical models and data mining techniques. They compared traditional models with newer models to see which worked better. This work made progress in the way people study car crashes.

A screenshot of a computer

Description automatically generated with medium confidence

Figure displaying the [website](mailto:https://www.mshp.dps.missouri.gov/HP68/SearchAction?searchTroop=%7btroop%7d) to extract the data.

**Dataset Extracted Link:**

<https://docs.google.com/spreadsheets/d/1vkmlBrxHKqS9enivw7VbZEFg_FBy7kxR/edit?usp=sharing&ouid=115159972945781347615&rtpof=true&sd=true>

# LITERATURE REVIEW:

* 1. **From "Motor vehicle fatalities during Memorial Day weekends, 1981–2016”** [(Tang, et al.,1981)](#tang)

This research paper was published by Tang et al. in BMC Research Notes in 2020. This paper has studied the risk of motor vehicle fatalities during Memorial Day. The papers also identified the numerous factors associated with the accidents. They compared the Memorial Day weekends to accidents that occurred during the non-holiday periods. The paper considered the days, weeks, years, age, sex, and role of a driver, passenger, or pedestrian. 35% of the accidents have occurred during holiday weekends compared to non-holiday. Memorial Monday recorded the highest number of accidents during the holidays.

* 1. **From "Paving the Way for Autonomous and Connected Vehicle Technologies in the Motor Carrier Industry."** [(Bernard, et al.,2018)](#pavingway)

The report was published in May 2018 and sponsored by the Midwest Transportation Center, which is a University Transportation Center (UTC) sponsored by the U.S. Department of Transportation Office of the Assistant Secretary for Research and Technology. The main aim of the report is potential safety issues and the need for infrastructure needed for the mass adoption of autonomous vehicles and CV (Connected Technologies in motorcycles. The paper has shown the benefits of these technologies like driver shortages, and mitigating driving errors.

* 1. **From "Investigation of Gender Differences in Large-Truck Crash Injury Severity in Missouri."** [(Bernard, et al.,2018)](#investigation),

The report was prepared by Jill M. Bernard Bracy and Christopher Mandy. The paper's main aim is to focus on the gender differences in the severity of the crash for male and female drivers.

The dataset has been taken from the Missouri State Highway Patrol STARS database and has the following variables road conditions, weather conditions, contributing circumstances, and crash injury severity. CHAID decision tree models have been applied to the crash data. The authors propose that customized training programs that address these specific factors can enhance driver safety and produce safer truck drivers.

* 1. **“An Application of Data Analytics to Outcomes of Missouri Motor Vehicle Crashes.”** [(Bernard et al.,2015)](#application),

The combination of existing literature, the CHAID decision tree model, and Missouri’s traffic regulations will get us better-driving policies. This paper puts a great emphasis on driving conditions when changing infrastructure or driving models.

# DATA ANALYSIS:

# The data that we use is pulled from a [website](mailto:https:https://www.mshp.dps.missouri.gov/HP68/SearchAction?searchTroop=%7btroop%7d) with crash reports in Missouri state. We web scrap this information using python libraries “Openpyxl” and “xlsxwriter” to read write the excel files. Following that, perform analysis on different variables produced in dataset. To read the data we use package “pandas” which is widely used for numerical computing in data science and data analysis workflows. To briefly describe this dataset, it holds the different parties (people and vehicles) involved in the accident and police personnel records this incident using a unique reference number and notes the severity of injury type and briefly describes it in miscellaneous information cell.

# To perform analysis, first we must understand the dataset in detail and design critical tasks to reach meaningful information that is useful for real-time analysis and modelling the data which is useful in similar types of crash occurrences. We divided the tasks into two categories of visual representation and lexical analysis to be performed on this dataset. After using pandas to read the data, it is converted into structed data. Next, modules like “*matplotlib*” and “*nltk* (Natural Language Processing Toolkit)” are involved to use this data for exploratory data visualization and generating textual summaries for each token in text.

# As our primary focus is to analyze and predict severity of injury, we classified this data into different visual representations for enhanced analysis. We used stack bar, heatmap and pie-chart to represent these injury types, as you can see in the below figures. Figure 1(a) illustrates the distribution of various injury types. Upon analysis, it becomes apparent that a considerable proportion of drivers experience minor injuries, while the occurrence of fatal injuries is comparatively low, with a percentage of only 6.5. Approximately 28.5% of drivers suffer from moderate injuries. Figure 1(b) has observations made upon analyzing the stacked bar chart depicting the timings of several types of injuries. It is observed that the occurrence of minor injuries was low after midnight. On the other hand, a considerable number of moderate injuries were recorded after 3 PM, which gradually decreased as the day progressed and approached 9 PM. Additionally, it is important to highlight that the highest incidence of minor injuries is observed during the 3 PM time.

A picture containing text, screenshot, font, diagram

Description automatically generated A picture containing text, diagram, screenshot, font

Description automatically generated

Figure 1(a) Injury Distribution Figure 1(b) Stacked bar: Injury Type vs Time

*A picture containing text, screenshot, rectangle, square

Description automatically generated*

In the same way, we represent same data using heatmap, Figure 1(c) to easily view which hour in the day has densely populated number of incidents for each injury type. The point of analyzing this data is to draw certain results using time, and help people take certain actions at a particular period to drive safely by evading these incidents. We use library “*seaborn*” for the process of creating complex visualizations by high-level interface for creating informative and attractive statistical graphics.

Figure 1(c) Heatmap: Injury Type vs Time

Following these observations, we use co-ordinates data in dataset to map each other and assess the data for counties and places where the crashes happen. Key take-away from this observation is to know the location where the crashes happen often, so we can implement necessary measures to mitigate these incidents. We use heatmap to pinpoint the exact location of incident using co-ordinates recorded in the crash incident. Figure 2(a) depicts the incidents that took place in a particular area.

A picture containing map, text, atlas

Description automatically generated

Figure 2(a) Heatmap: Place of Incident using Latitude and Longitude in dataset.

Note: please open the html file in browser mode to view the results

To list the top counties statistically, we map the co-ordinates and the counties according to the location using the query statement in python to visualize this data. You can find the results of this analysis in Figure 2(b) displaying the top 20 counties. To analyze the aftermath of these crash information, we study driver insurance companies, visualize the data for “SERIOUS” and” FATAL” injury type accidents as they involve more amount of claim for both the parties involved. This information is displayed using Figure 3(a) resulting the top insurance companies used in United States of America – Missouri state.

A screenshot of a computer screen

Description automatically generated with low confidence A picture containing text, screenshot, diagram, plot

Description automatically generated Figure 2(b) List of top 20 counties vs Incident CountFigure 2(c) Heatmap: Place of Incident using Latitude and Longitude

# PRE-PROCESSING:

# To start with modelling, we first must observe the text in “Misc. Information” column in dataset and pre-process it before feeding it to the system. We must analyze this text by counting the words in “Misc. information” and display the actual length of this information text. This visualization report is displayed in Figure 4(a) by implementing modules like “*wordcloud*” and “*re*” which are used for functions to perform operations such as searching for patterns, matching patterns, replacing patterns in strings, and exploring word frequency.

# A picture containing text, screenshot, diagram, plot Description automatically generated

# Figure 3(a) Length of Misc. Information text and its word frequency

# Specifically, the goal is to identify the sequence of events during the vehicle crash process and figure out the vehicle actions that caused the crash. It involves the below sequence of events as mentioned.

* + - Segmentation: Sentence segmentation is the process of dividing a text into its component sentences. There are two methods: *rule-based* and *classification-based* approaches. Rule-based method uses a predefined set of rules while classification-based method uses machine learning algorithms to segment the sentences.

# Tokenization: Tokenization is the process of breaking up a sentence into individual tokens such as words, numbers, or punctuation marks.

# Part-of-speech (POS) tagging: Part-of-speech (POS) tagging is the procedure of assigning labels to words based on their grammatical function, using not only the word's definition but also its surrounding context. This context incorporates the word's connections with neighboring and associated words within a phrase, sentence, or paragraph.

# Chunking: It divides a sentence into nonoverlapping segments like noun, verb, and prepositional phrases (NP, VP, and PP respectively).

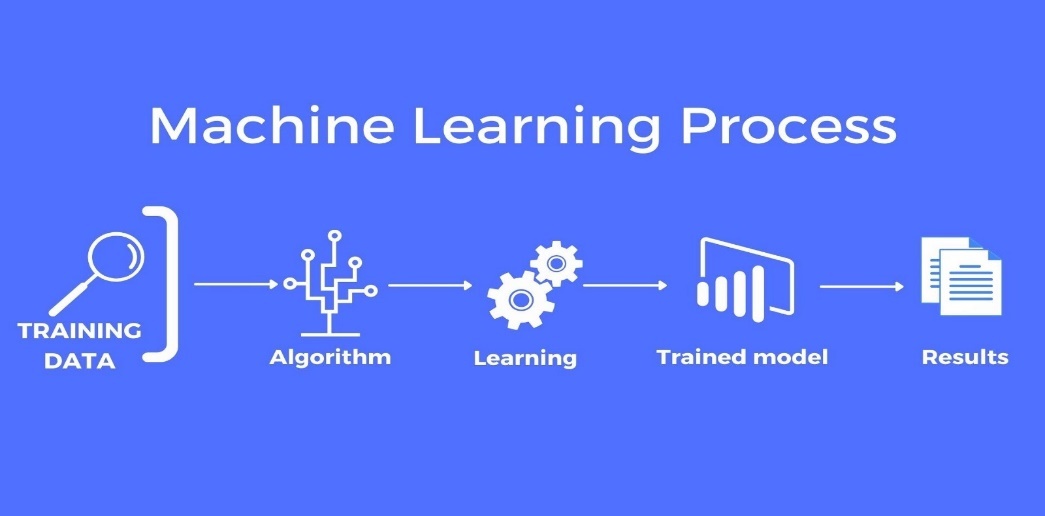
# By implementing these steps into programming, we get the below action sequences in crash information illustrated in figure 4(b). Figure 4(a) shows the code snippet to extract these sequences.

# A picture containing text, screenshot, font, software Description automatically generated A screenshot of a computer program Description automatically generated with medium confidence

# Figure 3(b) code snippet to extract action sequence Figure 3(c) Results of action sequence – Count.

Here, the function takes in a text string as input and performs a series of preprocessing steps on it. Firstly, it converts the text to lowercase, removes any punctuation marks and numbers, and then tokenizes the text. Next, it eliminates "*stopwords*" like "the," "and" etc. Finally, the remaining words are lemmatized, resulting in a preprocessed text string. The purpose of this function is to cleanse and normalize the input text, making it suitable for analysis or other natural language processing tasks.

1. **MODELLING**



* 1. **MACHINE LEARNING:**

The field of machine learning is a subset of artificial intelligence that enables machines to enhance their performance in each task without requiring specific programming instructions. It entails recognizing relationships and patterns in data for the purpose of making decisions or predictions.

There are three primary categories of machine learning:

* + - Supervised learning.
    - Unsupervised learning, and
    - Reinforcement learning.

Supervised learning involves training machine learning models on labeled data, where the desired outcome is known. Unsupervised learning, on the other hand, identifies patterns in unlabeled data without prior knowledge of the outcome. Reinforcement learning is based on the concept of trial and error, where an agent interacts with an environment to learn optimal decision-making. Machine learning finds application in various domains such as healthcare, finance, transportation, and entertainment.

Machine learning models are often preferred over statistical models because they can handle more complex relationships within data and can continuously improve with more data. Statistical models, on the other hand, are better suited for simpler relationships and may require more manual intervention for improvement. Additionally, machine learning models can handle larger datasets and can handle data with more missing or incomplete values. To analyze the performances with machine learning, we need to obtain values like features and targets which are essentials to initialize the model to divide train and test data.

Features refer to the input variables used in a machine learning model to make predictions or classifications. These can be numerical or categorical data and are typically organized in a matrix where each row represents an observation, and each column represents a feature. Feature selection is a crucial step in model development, as it directly impacts the model's performance and interpretability. The target variable is the output variable or the variable that we are trying to predict or classify. It is also known as the response variable. In supervised learning, the target variable is typically labeled, and the goal is to develop a model that can accurately predict or classify new observations. In unsupervised learning, there is no labeled target variable, and the goal is to discover patterns or structure in the data.

In machine learning, a common technique used to evaluate a model's performance is called train-test split. This method involves dividing the dataset into two subsets: the training set, which is used to train the model, and the testing set, which is used to assess the model's performance. During training, the model parameters are adjusted to fit the data in the training set, while the testing set is used to measure how well the model can generalize to new, unseen data. The split ratio can be adjusted based on the dataset, but it is typical to use 80% of the data for training and 20% for testing. A random split is essential to avoid any potential biases in the data.

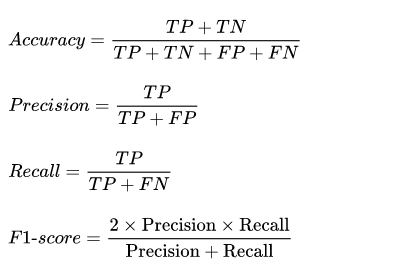
For modelling, we implement the classification machine learning models over regression because textual data classification models are better suited to predicting categorical outcomes. Textual data often involves classifying text into categories, such as positive or negative sentiment, spam or non-spam, or topic classification. Regression models, on the other hand, are designed to predict continuous numerical outcomes, which may not be well-suited to the discrete nature of textual data. For classification, there are certain model evaluation metrics followed to illustrate the model. This is explained in the model evaluation metrics section.

1. **MODEL EVALUATION METRICS:**

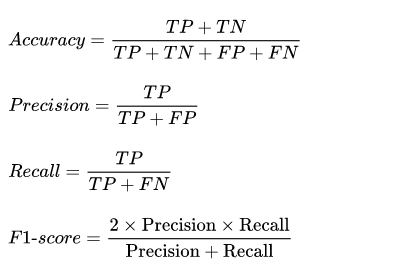
Machine learning models can be evaluated using metrics that measure their performance and effectiveness. These metrics provide a standardized way to assess how well a model is doing in terms of various measures, such as accuracy, precision, recall, F1 score. By using these metrics, researchers and practitioners can determine the strengths and weaknesses of a model and make improvements as needed.

**Classification Models Metrics:**

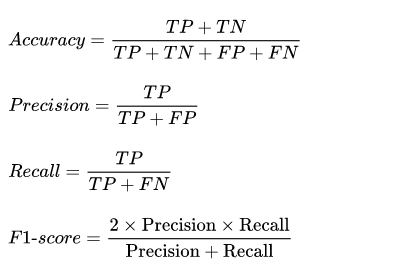
* Confusion Matrix: A confusion matrix is a table that displays the performance of a classification model by showing the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for a given set of predictions.
* Accuracy: Accuracy is a metric that calculates the percentage of correctly predicted instances from the total instances. It can be computed by dividing the sum of true positives and true negatives by the total number of instances. True positives are the correctly predicted positive instances, true negatives are the correctly predicted negative instances, false positives are the incorrectly predicted positive instances, and false negatives are the incorrectly predicted negative instances.



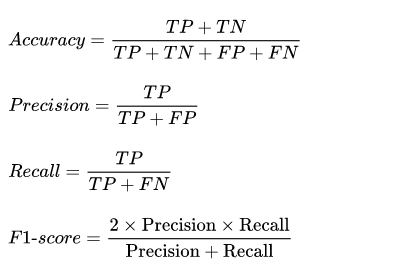
* Precision: Precision is a metric that quantifies the number of accurate positive predictions out of all positive predictions made by the model. The formula for precision is TP / (TP + FP), where TP represents the true positives and FP represents the false positives. Precision is an important measure of the model's ability to minimize false positives.



* Recall (Sensitivity or True Positive Rate): The metric Recall measures the model's ability to identify all actual positive instances, particularly the number of true positive predictions. This metric is computed as TP / (TP + FN), where TP represents the true positives, and FN denotes the false negatives. Recall focuses on reducing false negatives in the model's predictions.



* F1 Score: The F1 score provides a balanced evaluation of a model's performance by considering both precision and recall and is calculated as the harmonic mean of the two metrics. The formula for F1 score is 2 times the product of precision and recall, divided by the sum of precision and recall.



**Scikit-learn:**

Scikit-learn, or SKLearn for short, is a popular Python library for machine learning. It provides a range of tools for building machine learning models, including algorithms for classification. SKLearn is built on top of other popular scientific computing libraries in Python, such as NumPy, SciPy, and Matplotlib. It provides many useful tools for preprocessing and feature selection, such as scaling, normalization, and evaluating model performance. Overall, SKLearn is an important library for working with machine learning in Python.

We implemented 5 classification machine learning models as listed below.

1. **Naive Bayes Classification:**

Naive Bayes is a classification algorithm that utilizes Bayes' theorem to calculate the probability of a certain class given a set of features. The algorithm operates on the assumption that the features are independent of each other when given the class label. It assesses the conditional probability of each class for a particular set of features and subsequently identifies the class with the highest probability as the predicted class. This classification produced an accuracy of 56% as mentioned in the confusion matrix.

A screenshot of a computer screen

Description automatically generated with medium confidence

Figure 4(a). Accuracy of Naïve Bayes Classification

1. **K-Nearest Neighbors (KNN) Classification:**

K-Nearest Neighbors is a non-parametric classification algorithm. It classifies new instances by comparing them to the k nearest neighbors in the feature space. The algorithm calculates the distance between the new instance and the training instances and assigns the class label based on the majority class among its k nearest neighbors. Accuracy of KNN classification is 53% as shown in figure 5(b)

A screenshot of a computer screen

Description automatically generated with low confidence

Figure 4(b). Accuracy in KNN classification

1. **Decision Tree Classification:**

Decision Tree is a widely used classification algorithm that creates a tree-like model of decisions and their potential consequences. It splits the data based on feature values, selecting the features that provide the most discriminatory power. The tree is built recursively until all instances are correctly classified or a termination condition is met. New instances can be classified by traversing the tree based on their feature values.

A screenshot of a computer screen

Description automatically generated with medium confidence

Figure 4(c). Accuracy in Decision Tree classification

1. **Random Forest Classification:**

Random Forest is a machine learning algorithm that leverages the power of ensemble learning to enhance classification accuracy. It achieves this by generating multiple decision trees using subsets of features and instances from the training data. The predictions of these trees are then combined through either majority voting or averaging to produce the final prediction.

A screen shot of a computer

Description automatically generated with low confidence

Figure 4(d). Accuracy in Random Forest classification

1. **Neural Network Model - Multi-layer Perceptron Classifier:**

The Multi-layer Perceptron (MLP) is a type of artificial neural network used for classification tasks. It consists of multiple layers of interconnected nodes (neurons) that process the input data. Each neuron applies a non-linear activation function to its weighted inputs, and the outputs are propagated forward through the network. During training, the model adjusts the weights to minimize the prediction error using techniques like backpropagation. The trained MLP can then be used to classify new instances based on their learned patterns. We evaluated these models to compute the accuracy and found from the observations that Neural Network model has the highest accuracy when compared to others with 60% in the confusion matrix.

A screenshot of a computer screen

Description automatically generated with low confidence

Figure 4(e) Accuracy of Multi- Layer Perceptron Classification

The superior performance of Neural Network over other models could be attributed to its flexibility in capturing complex, non-linear relationships between features and target. Neural Networks can learn from large and diverse datasets and generalize well to new data. However, the efficacy of any machine learning model depends on the dataset used and feature selection. Furthermore, the computational cost and complexity of training Neural Networks should be taken into consideration.

Each classification model has unique characteristics and assumptions, which determine their suitability for various problem domains and data types. The appropriate model selection depends on several factors, such as the complexity of the problem, data nature, and required interpretability. Furthermore, classification models can efficiently handle imbalanced datasets and generate interpretable results, making them ideal for text analysis tasks.

**Imbalanced Datasets:**

Imbalanced datasets present challenges for machine learning models due to biased outcomes. To address this, techniques such as up-sampling and down-sampling are used. Up-sampling increases minority class samples, while down-sampling decreases majority class samples. Hybrid approaches combine both techniques to achieve a balanced dataset. Evaluation of imbalanced datasets involves specific metrics like precision, recall, F1 score which provide a comprehensive understanding of model performance in both classes. These metrics consider the imbalanced nature of the data and help in accurate classification and decision-making despite class distribution disparities.

1. **CONCLUSION:**

In conclusion, we implemented machine learning models to predict severity of injury type, particularly in Missouri state. The developed Models can be used to evaluate the severity of injury type on varied factors like vehicle description, gender, age, safety device etc.., in the crash information. From our observations, we found that vehicle model year, gender safety device and age might be the primary factors of crash in the case of serious and fatal injury types. This observation is made when we map different accident factors against the injury type present in the dataset. Vehicle crashes are still a significant concern, causing numerous injuries and fatalities worldwide. The analysis focused on predicting injury types based on trained models using random forest, logistic regression, and support vector classification. By using the data extracted from the Missouri website, these models can supply valuable insights into the likelihood and severity of injuries in car crashes. Such predictions can aid in improving emergency response protocols, implementing effective safety measures, and reducing the impact of vehicle accidents. These advancements in modeling techniques can enhance our understanding of the factors contributing to crash severity and help in developing targeted interventions and preventive measures.

In the field of predicting injury types in vehicle crashes, there are possibilities for further research beyond the current analysis. Integrating real-time data sources, such as sensor data from vehicles and traffic surveillance systems, can improve the accuracy and timeliness of predictions. Moreover, advanced machine learning techniques like deep learning algorithms and ensemble methods may enhance the predictive performance of the models. Investigating the impact of external factors such as weather conditions, road infrastructure, and driver behavior on crash severity can offer valuable insights for comprehensive risk assessment and accident prevention strategies. Further exploration in this field holds promise for contributing significantly to road safety and bringing about a measurable decrease in the frequency and severity of injuries stemming from vehicular accidents.

# REFERENCES:

Tang, Y., Ratnapradipa, K.L., Xiang, H. et al. Motor vehicle fatalities during Memorial Day weekends, 1981–2016. BMC Res Notes 13, 7 (2020). <https://doi.org/10.1186/s13104-019-4881-0>

Bernard Bracy, Jill M; Bao, Ken; Mundy, Ray A, "Paving the Way for Autonomous and Connected Vehicle Technologies in the Motor Carrier Industry" (2018). <https://rosap.ntl.bts.gov/view/dot/36103>

[Bernard Bracy, Jill M](https://rosap.ntl.bts.gov/gsearch?ref=docDetails&sm_creator=Bernard%20Bracy%2C%20Jill%20M) , [Mondy, Christopher](https://rosap.ntl.bts.gov/gsearch?ref=docDetails&sm_creator=Mondy%2C%20Christopher), "Investigation of Gender Differences in Large-Truck Crash Injury Severity in Missouri" (2018). https://rosap.ntl.bts.gov/view/dot/36224

Bernard, Jill Marie, "An Application of Data Analytics to Outcomes of Missouri Motor Vehicle Crashes" (2015). *Dissertations*. 168. <https://irl.umsl.edu/dissertation/168>

Bayam, E., J. Liebowitz, and W. Agresti. 2005. Older Drivers and Accidents: A Meta Analysis and Data Mining Application on Traffic Accident Data. Expert Systems with Applications, Vol. 29, No. 3, pp. 598–629.

Williams, A.F.; Carsten, O. Driver age and crash involvement. Am. J. Public Health 1989, 79, 326–327.

Hu, P.S.; Young, J.R.; Lu, A. Highway Crash Rates and Age-Related Driver Limitations: Literature Review and Evaluation of Data Bases; United States: Washington, DC, USA, 1993.

Massie, D.L.; Green, P.E.; Campbell, K.L. Crash involvement rates by driver gender and the role of average annual mileage. Accid. Anal. Prev. 1997, 29, 675–685.

[1] <https://github.com/IndiraBobburi/accident-severity-prediction/blob/master/Team4_Project_Report.pdf>

[2]<https://github.com/theDefiBat/ROAD-ACCIDENTS-PREDICTION-AND-CLASSIFICATION/blob/master/report/4-Final-Report-3-aw.pdf>

[3] <https://www.frontiersin.org/articles/10.3389/fbuil.2022.860805/full>

[4] https://www.txdot.gov/data-maps/crash-reports-records/crash-data-analysis-statistics.html