Forecasting the Severity of Road Accidents in the USA

Dharanidhar Reddy Banala, Kalyan Vikkurthi, Manidedeepya Chennapragada and Nikhil Mylarusetty

Department of Applied Data Science, San Jose State University

Data 245: Machine Learning

Professor Dr. Shih Yu Chang

May 18, 2023

Table of Contents

Abstract

Efficient prediction of road accident severity plays a crucial role in enhancing safety measures and reducing the impact of accidents in the transportation sector. To address this challenge, machine learning models can be utilized to estimate the severity of accidents based on various factors such as weather conditions, road type, time of day, and other relevant features. Accurate forecasting of accident severity enables authorities and emergency services to allocate appropriate resources and respond effectively. This project focuses on utilizing four distinct machine learning models, namely Decision Tree, Random Forest, K-Nearest Neighbors (KNN), and Logistic Regression, to compare their predictive performance in determining accident severity. The Decision Tree model is a versatile algorithm that creates a tree-like model of decisions and their possible consequences. Random Forest combines multiple decision trees to make more accurate predictions. KNN classifies instances based on their proximity to other instances in the feature space. The evaluation of these models will be conducted using appropriate metrics such as accuracy, precision, recall, and F1-score. These metrics will provide insights into the performance of each model in terms of correctly identifying the severity levels of road accidents. By comparing the results, it was found that Random Forest performed the best with an accuracy of 86.52%. The findings of this project can contribute to developing a reliable system for predicting accident severity, which can be utilized by transportation authorities, emergency services, and other stakeholders to implement proactive measures for accident prevention, resource allocation, and improved emergency response. Ultimately, the goal is to reduce the severity and frequency of road accidents, enhancing overall safety on US roads.

1. Problem Understanding and Formulation

In numerous regions of the globe, accident cases are ignored and not considered seriously. According to statistics, 20 to 50 million accidents occur annually. According to the National Highway Traffic Safety Administration, there are 64 million traffic incidents in the United States every year. Every day, ninety individuals die in traffic accidents. This is the primary cause of mortality in the United States. Regarding this matter, action is required. These statistics grabbed our interest and inspired us to investigate the causes of accidents.

Reducing traffic accidents is a crucial public safety policy that requires in-depth analysis. Since this is one of the most significant problems in the modern world, we decided to analyze the Kaggle dataset on accidents. We extensively analyzed traffic accident data and performed analytics on the top 20 cities in the United States to identify contributing factors based on trends and propose solutions. And then implemented algorithms to make predictions on the severity of the accidents.

1.1 Problem Statement

The objective of this project is to develop machine learning models, including logistic regression, random forest, decision tree, and K-nearest neighbors (KNN), to predict the severity of accidents in the United States. Accurate prediction of accident severity plays an important role in immediate emergency response and resource allocation, thereby enhancing the accident managing system for better performance than the existing and potentially reducing the number of fatalities and injuries.

1.2 Project Background

Road traffic accidents pose a significant threat to public safety and result in substantial economic losses. Therefore, predicting the severity of accidents is of utmost importance to assess their

potential impact and enable timely response measures. Machine learning algorithms have shown promise in analyzing accident data and providing accurate severity predictions. In this project, we will leverage logistic regression, random forest, decision tree, and KNN algorithms to build models that can predict the severity of accidents based on various factors such as weather conditions, road type, time of day, and other relevant features.

1.3 Literature Survey

Yan and Shen (2022) look into how the random forest algorithm can be used to predict how bad traffic accidents will be. In their research, the authors look at the different things that can affect how bad an accident is and suggest a prediction model based on the random forest method. The study talks about how accurate predictions of severity are important for improving traffic safety. It also talks about how their results could be used to promote sustainable transportation systems. In the end, the paper gives useful information about how machine learning methods can be used to predict how bad traffic accidents will be. This shows how important it is to promote sustainable practices in the transportation field.

Nasri et al. (2022) focuses on the causes of severe pedestrian crash injuries in Victoria, Australia, and the variables that increase those risks. In order to examine the data and determine the factors that lead to varying degrees of injury severity, the authors use both ordered and unordered logistic regression models. The study's primary objective is to identify major risk variables that contribute to pedestrian crash injuries so that they can be reduced or eliminated. Findings from this research add to the body of knowledge on pedestrian safety and can be used to craft more effective interventions and policies that mitigate pedestrian injuries.

M et al. (2022) explores the use of the K-Nearest Neighbors (KNN) algorithm for accident prediction. The authors look into how historical accident data and various variables

might be used to train the KNN model and forecast the likelihood of accidents in specific places. The study's goal is to improve accident prevention and road safety measures through the use of machine learning techniques. The research highlights the potential of predictive modeling in identifying high-risk locations prone to accidents by applying the KNN algorithm. The data can be used to design proactive tactics to reduce accidents and improve overall transportation safety.

VA and AA (2014) explains how using neural networks and decision trees to examine the causes of road accidents. The authors investigate how these machine learning strategies might be used to draw conclusions from accident records. This paper uses decision trees to help identify key contributors to accidents and categorize accident severity. Further, neural networks are used to model the connection between accident characteristics and outcome prediction. The research highlights the significance of using cutting-edge computational approaches for traffic accident investigation, shedding light on the potential of decision trees and neural networks for better comprehending and tackling road safety issues.

1.4 CRISP-DM Methodology

This report aims to provide a comprehensive analysis of US accident severity using the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology.

Business Understanding

During this phase, the project team conducted meetings to discuss project ideas and objectives. Team members suggested various ideas and conducted a literature survey to derive a clear project objective. Each team member shared their interests in implementing different machine learning techniques. Once the project topic was finalized, the team conducted extensive literature surveys to gain a better understanding of potential approaches, technologies, and

methodologies that could be employed. A well-defined plan for data mining was also created during this phase

Data Understanding

In the second phase, the team focused on gathering reliable and authenticated data for the project. A dataset from a reputable source, "Kaggle," was chosen. The team performed exploratory data analysis to assess the dataset's quality, including checking for null values and other data quality measures. This phase laid the foundation for subsequent phases, as the team gained a better understanding of the dataset and planned the project accordingly.

Data Preparation

Before proceeding with the analysis, it is necessary to prepare the dataset for further processing. This stage involves performing data cleaning tasks, handling missing values, and transforming variables as required. Additionally, feature engineering techniques will be applied to extract relevant features that could potentially influence accident severity. Exploratory data analysis (EDA) is a crucial step in gaining initial insights and understanding the characteristics of the data. The utilization of statistical techniques and visualizations, we can explore the relationships between different variables and accident severity. This analysis will help uncover potential risk factors that are associated with severe accidents.

Modeling

In this phase, predictive models will be developed to forecast accident severity based on the available dataset. Machine learning algorithms, such as decision trees, random forests, KNN and logistic regression, will be trained using the prepared dataset. These models will then be evaluated and fine-tuned using appropriate performance metrics to ensure their accuracy and reliability.

Evaluation

The performance of the developed models are assessed based on their predictive accuracy and ability to correctly classify accident severity based on confusion matrix . Additionally, feature importance analysis will be conducted to identify the most influential factors that contribute to severe accidents. These findings will be crucial in prioritizing interventions and safety measures.

Deployment

The insights gained from the analysis will be presented in a clear and concise manner, providing actionable recommendations to stakeholders. The report will highlight the key factors contributing to accident severity and suggest preventive measures and policy changes that can be implemented to reduce the occurrence of severe accidents in the United States.

1.5 System Architecture

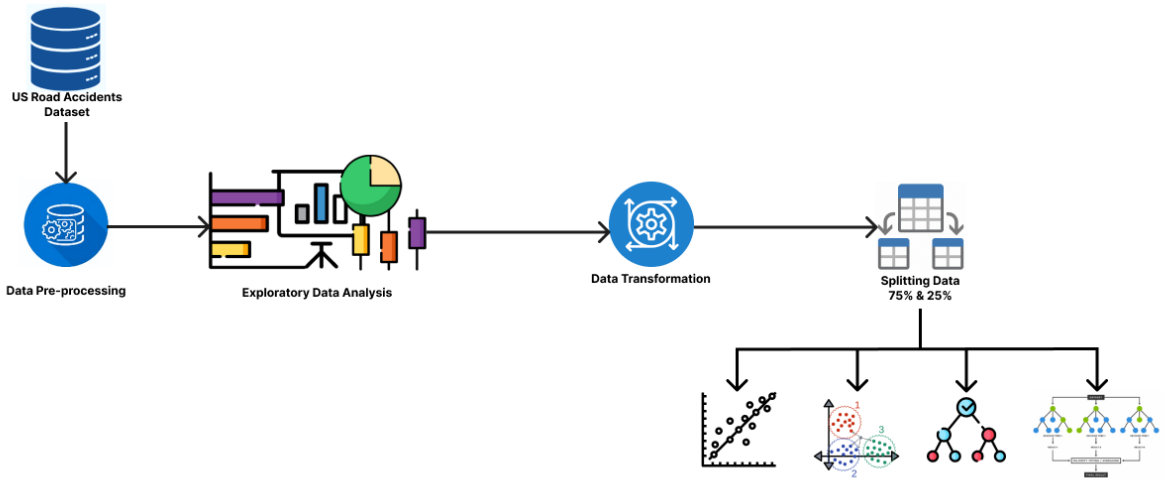
The system architecture for predicting the severity of US road accidents using Machine Learning involves several key phases as shown in **Figure** 1. The first phase is dedicated to collecting the necessary data from a reliable source, in this case, the US road accidents dataset from Kaggle. This dataset is downloaded and stored for further processing.

In the data preprocessing phase, the collected data undergoes a series of steps to ensure its quality and suitability for analysis. Once the data is preprocessed, the Exploratory Data Analysis (EDA) phase begins. This phase focuses on understanding the dataset through visualizations, statistical measures, and data exploration techniques. Following EDA, the data is transformed into suitable formats for model training. The dataset is split into two parts: the independent variables (features) and the dependent variable (severity of accidents). The data is further divided into training and testing sets, typically with a 75% - 25% split ratio. In the modeling phase, various machine learning algorithms are employed to predict accident severity.

In this case, the Random Forest, Logistic Regression, K-Nearest Neighbors (KNN), and Decision Tree algorithms are utilized using appropriate libraries or frameworks, such as scikit-learn in Python. The models are trained using the training set and their performance is evaluated using suitable evaluation metrics like accuracy, precision, recall, and F1-score. Hyperparameters of each model are optimized through techniques like cross-validation or grid search to enhance their predictive capabilities. Once the models have been evaluated, the best-performing model is selected based on its performance on the testing set.If necessary, further enhancements or optimizations are made to improve its performance and accuracy. Finally, the selected model is deployed in a suitable environment or platform. This typically involves creating an application or API that allows users to input relevant features and obtain predictions for accident severity.

Figure 1

System Architecture



2. Data Understanding and Exploration

2.1 Data Collection

The dataset is obtained from Kaggle which contains information about US road accidents. This data was collected from various sources, including traffic cameras, weather stations, emergency response systems, and other related sources. It covers accidents that occurred in different regions of the United States. The dataset includes a wide range of attributes, such as the location and severity of the accident, weather conditions, road conditions, time and date of the accident, presence of any traffic signals, and other relevant factors. These attributes provide valuable insights into the circumstances surrounding each accident. To address missing data, we calculated the percentage of missing values in each column. This helped us identify any discrepancies or columns with a high proportion of missing values. By understanding the extent of missing data, we can determine the appropriate strategies to handle them, such as imputation or removal.The sample dataset can be seen in **Figure 2**

Figure 2

Sample Dataset

	ID	Severity	Start_Time	End_Time	Start_Lat	Start_Lng	End_Lat	End_Lng	Distance(mi)	Description	...	Roundabout	Station	Stop	Traffic_Calmir
0	A-1	3	2016-02-08 00:37:08	2016-02-08 06:37:08	40.108910	-83.092860	40.112060	-83.031870	3.230	Between Sawmill Rd/Exit 20 and OH-315/Olentang...	...	False	False	False	Fal
1	A-2	2	2016-02-08 05:56:20	2016-02-08 11:56:20	39.865420	-84.062800	39.865010	-84.048730	0.747	At OH-4/OH-235/Exit 41 - Accident.	...	False	False	False	Fal
2	A-3	2	2016-02-08 06:15:39	2016-02-08 12:15:39	39.102660	-84.524680	39.102090	-84.523960	0.055	At I-71/US-50/Exit 1 - Accident.	...	False	False	False	Fal
3	A-4	2	2016-02-08 06:51:45	2016-02-08 12:51:45	41.062130	-81.537840	41.062170	-81.535470	0.123	At Dart Ave/Exit 21 - Accident.	...	False	False	False	Fal
4	A-5	3	2016-02-08 07:53:43	2016-02-08 13:53:43	39.172393	-84.492792	39.170476	-84.501798	0.500	At Mitchell Ave/Exit 6 - Accident.	...	False	False	False	Fal

Figure 3

Columns in the US Accidents Dataset

```
Index(['ID', 'Severity', 'Start_Time', 'End_Time', 'Start_Lat', 'Start_Lng',
       'End_Lat', 'End_Lng', 'Distance(mi)', 'Description', 'Number', 'Street',
       'Side', 'City', 'County', 'State', 'Zipcode', 'Country', 'Timezone',
       'Airport_Code', 'Weather_Timestamp', 'Temperature(F)', 'Wind_Chill(F)',
       'Humidity(%)', 'Pressure(in)', 'Visibility(mi)', 'Wind_Direction',
       'Wind_Speed(mph)', 'Precipitation(in)', 'Weather_Condition', 'Amenity',
       'Bump', 'Crossing', 'Give_Way', 'Junction', 'No_Exit', 'Railway',
       'Roundabout', 'Station', 'Stop', 'Traffic_Calming', 'Traffic_Signal',
       'Turning_Loop', 'Sunrise_Sunset', 'Civil_Twilight', 'Nautical_Twilight',
       'Astronomical_Twilight'],
      dtype='object')
```

2.2 Data Exploration and Pre-processing

During the data exploration phase, we examined a dataset to understand its structure, attributes, and any issues it has. We checked the data types, identified missing values, and conducted statistical summaries for insights. We also addressed missing values by imputing them or removing affected rows. Data cleaning focused on quality issues like outliers or inconsistent values, using transformations or removal. We ensured data remains accurate by correcting formatting inconsistencies or erroneous values. This prepares the dataset for analysis by improving its quality and usability. These steps made the data reliable, consistent, and suitable for further tasks. The Figure 4 and Figure 5 below shows how the data cleaning was done. Specifically Figure 4 shows the NULL values in the data and the Figure 5 depicts the cleaned data without any NULL values.

Figure 4

Null Value Counts Before Removal

```
ID                       0  Pressure(in)            59200
Severity                 0  Visibility(mi)          70546
Start_Time               0  Wind_Direction          73775
End_Time                 0  Wind_Speed(mph)        157944
Start_Lat                0  Precipitation(in)      549458
Start_Lng                0  Weather_Condition       70636
End_Lat                  0  Amenity                     0
End_Lng                  0  Bump                        0
Distance(mi)             0  Crossing                    0
Description              0  Give_Way                    0
Number             1743911  Junction                    0
Street                   2  No_Exit                     0
Side                     0  Railway                     0
City                   137  Roundabout                  0
County                   0  Station                     0
State                    0  Stop                        0
Zipcode               1319  Traffic_Calming             0
Country                  0  Traffic_Signal              0
Timezone              3659  Turning_Loop                0
Airport_Code          9549  Sunrise_Sunset           2867
Weather_Timestamp    50736  Civil_Twilight           2867
Temperature(F)       69274  Nautical_Twilight        2867
Wind_Chill(F)       469643  Astronomical_Twilight    2867
Humidity(%)          73092  dtype: int64
```

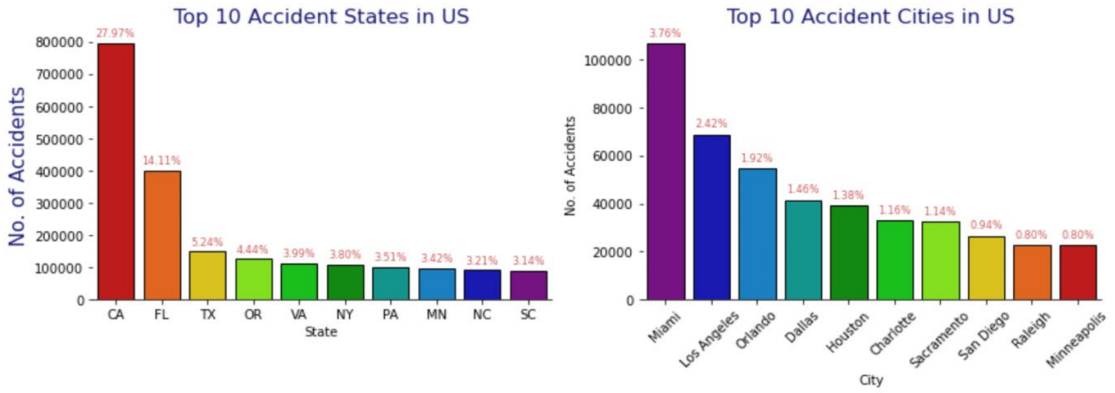
Figure 5

Null Value Counts After Removal

```
Severity              0
Start_Lat             0
Start_Lng             0
Distance(mi)          0
Side                  0
County                0
State                 0
Timezone              0
Airport_Code          0
Temperature(F)        0
Humidity(%)           0
Pressure(in)          0
Visibility(mi)        0
Wind_Direction        0
Wind_Speed(mph)       0
Weather_Condition     0
Amenity               0
Bump                  0
Crossing              0
Give_Way              0
Junction              0
No_Exit               0
Railway               0
Roundabout            0
Station               0
Stop                  0
Traffic_Calming       0
Traffic_Signal        0
Turning_Loop          0
Sunrise_Sunset        0
Civil_Twilight        0
Nautical_Twilight     0
Astronomical_Twilight 0
Year                  0
Month                 0
Hour                  0
DelayTime             0
Day                   0
dtype: int64
```

Figure 6

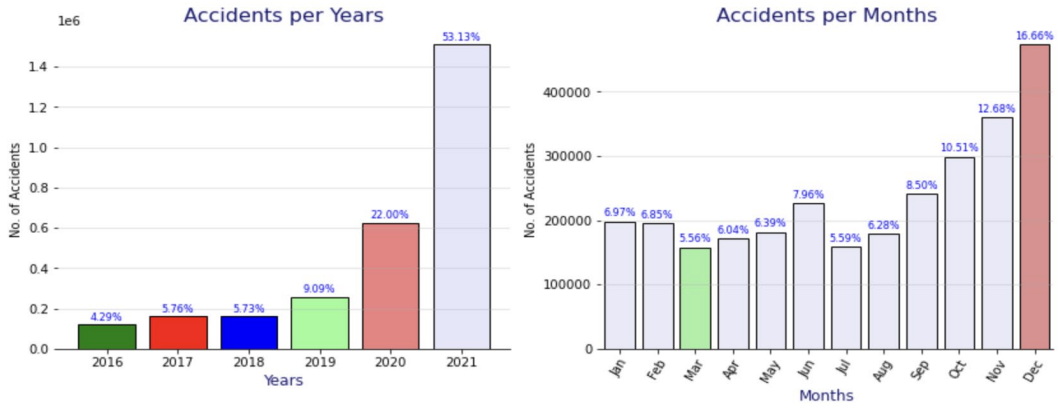
Top 10 Accident States and Cities



The above Figure 6 represents the Top 10 highest number of accidents occurring in states and cities. The state of California has the highest number of accidents with a high margin accounting for 27.97 percent followed by Florida with 14.11 percent. Apart from these, the remaining cities contribute less to the number of accidents. Coming to accidents by cities Miami stands at the first position followed by Los Angeles which is part of California.

Figure 7

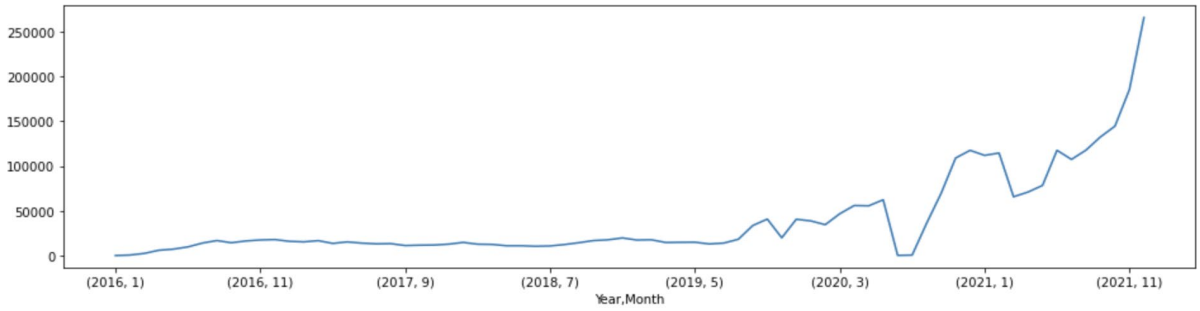
Number of Accidents per Years and Months



The Figure 7 shows the accidents per year staring from 2016 to 2021. In the year 2016, the number of accidents are less as the number of vehicles are less compared to 2021. In the covid year 2020, Even though there are lock down restrictions there is rapid increase of accidents which is astonishing. After lifting of lockdown the there is sudden spike in the number of accidents from the year 2020 to 2021. Coming to accidents per month, the number of accidents are highest for the month December due to Christmas and New Year.

Figure 8

Line Graph For Number of Accidents from 2016-2021

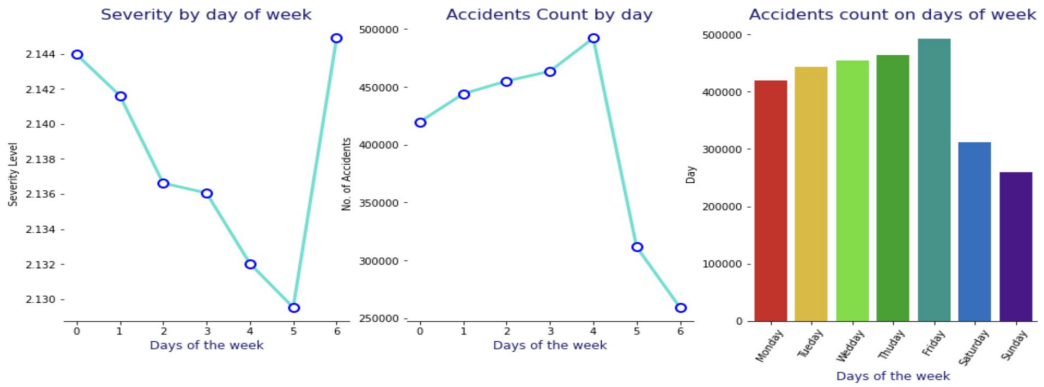


The Figure 8, it shows the trend from the year 2016 where the number of accidents from 2016 to 2019 are constant and there is slight increase from there. However, in the year 2020

there is a slight decrease in the number of accidents due to Covid cases. In the subsequent year, there was a sudden increase after easing lockdown restrictions.

Figure 9

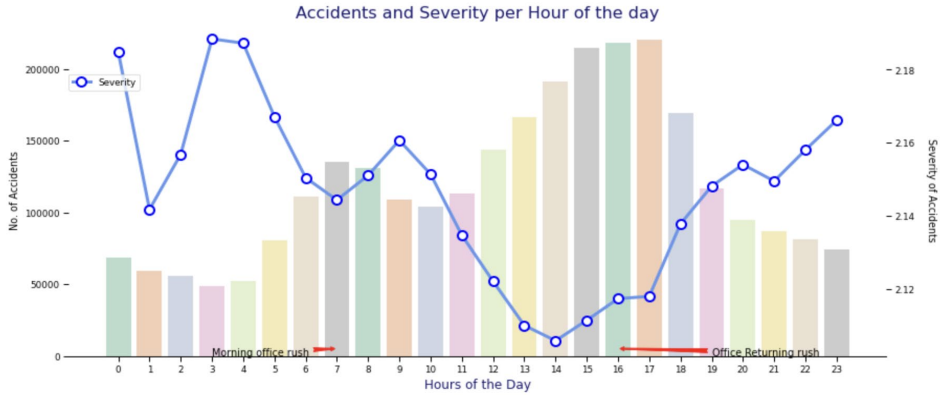
Severity, Accident Counts based on the Weekday



These are three graphs in Figure 9 that are depicted above that includes a bar and line plot. The first graph shows the Severity by the day of week which shows that the Severity of accidents on the Saturdays are lower compared to other days. However, coming to the number of accidents for every day of the week Sunday has the lowest number of accidents whereas it is highest for Friday as it is the start of the weekend. On the working days the number of accidents remained constant.

Figure 10

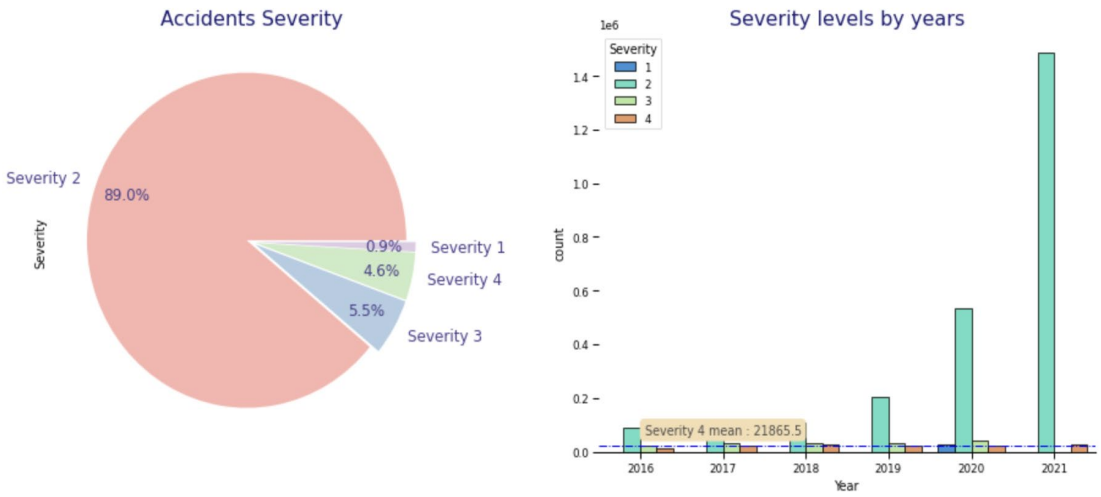
Accidents and Severit per Hour of the Day



In Figure 10 we can see the crucial task where the number of accidents are plotted for each hour. We can clearly observe there are two peak hours that include Morning peak hour and the Evening peak hour. The reason for the morning peak hour is due to office hours especially from 7 am to 9 am whereas the number of accidents in the peak hour at evening is highest as the people will be returning from the offices.

Figure 11

Effect of Severity Levels



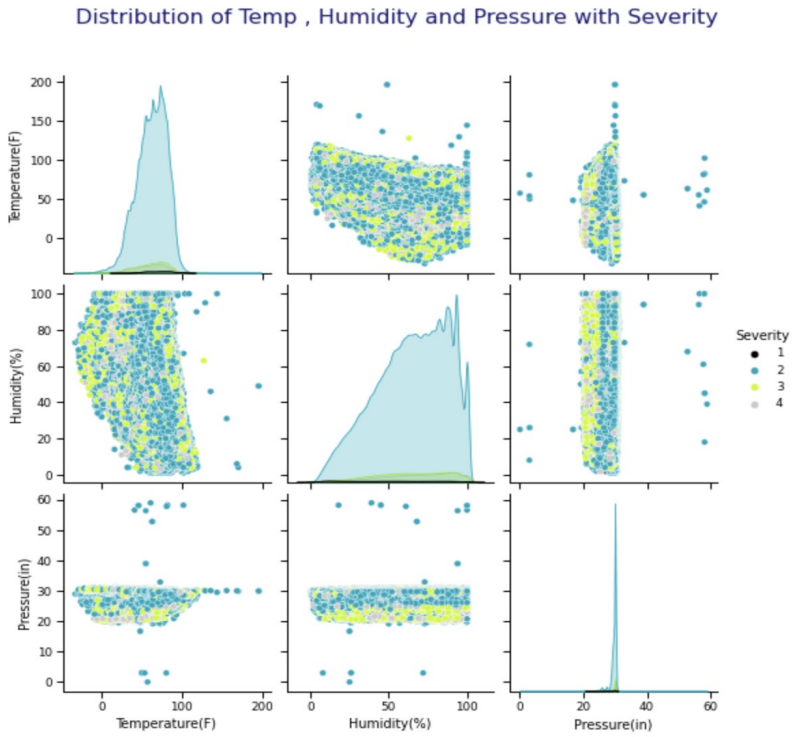
In Figure 11 the graphs are especially helpful in analyzing the severity level which is very helpful in the Data Modelling as target variable for this project is Severity. Moreover, we can clearly observe the value for severity 2 which is abnormally high when compared to the other severity levels. It is clearly evident that the target variable is completely skewed towards one severity level. In the second graph is shows that severity level 2 got drastically increased due to increase in the number of mild accidents.

The pairplot graph in Figure 12 allows us to visually analyze the relationships between temperature, humidity, pressure, and accident severity, providing insights into potential correlations. The diagonal plots display the individual distributions of each variable, revealing

trends and variations in temperature, humidity, and pressure across different severity levels. The off-diagonal plots show the bivariate relationships between pairs of variables, helping us understand how they interact and their potential impact on accident severity. The use of different colors for each severity level aids in distinguishing between severe and less severe accidents, highlighting any patterns or trends in the distribution of weather variables. In summary, the pairplot graph serves as a valuable exploratory tool, enabling researchers and stakeholders to gain insights and identify factors influencing accident severity under different weather conditions.

Figure 12

Distribution of Temperature, Humidity and Pressure with Severity



The Figure 13 of the above graph is intended to provide a visual representation of the correlations between accident-related characteristics. We can identify potential associations or dependencies between variables by scrutinizing the heatmap. This information is essential for

comprehending the factors that contribute to the severity of accidents and can aid in the development of prevention and mitigation strategies. For the severity there is no much correlation with any column in the dataset.

Figure 13

Correlation Matrix

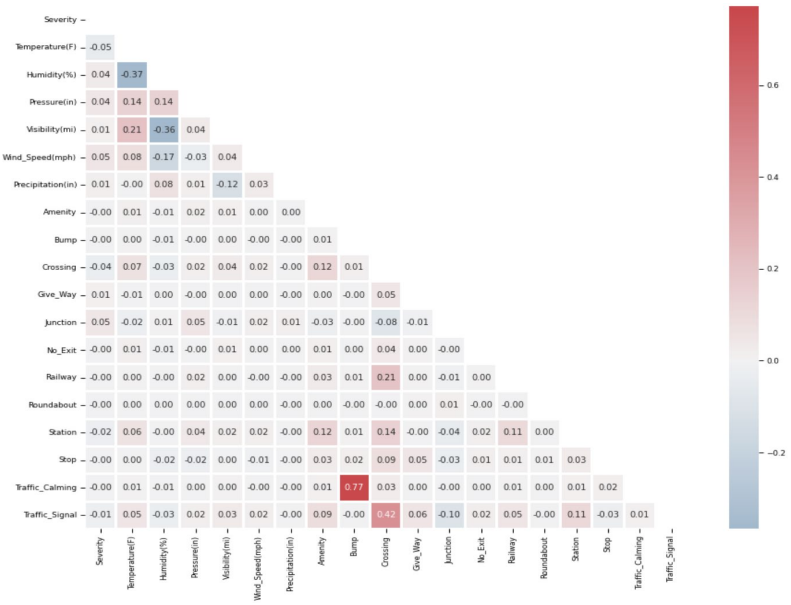
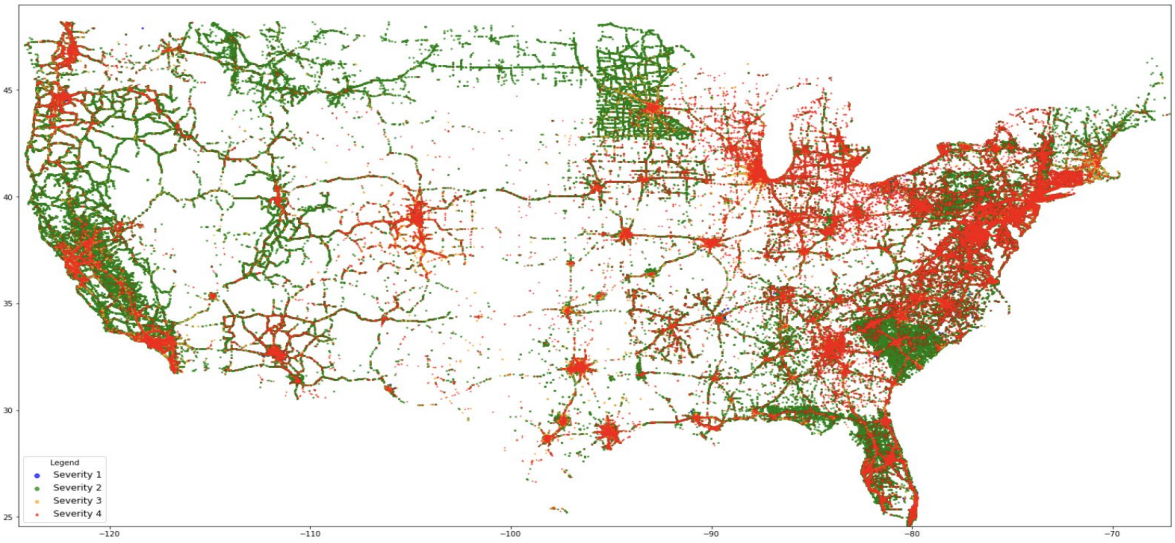


Figure 14

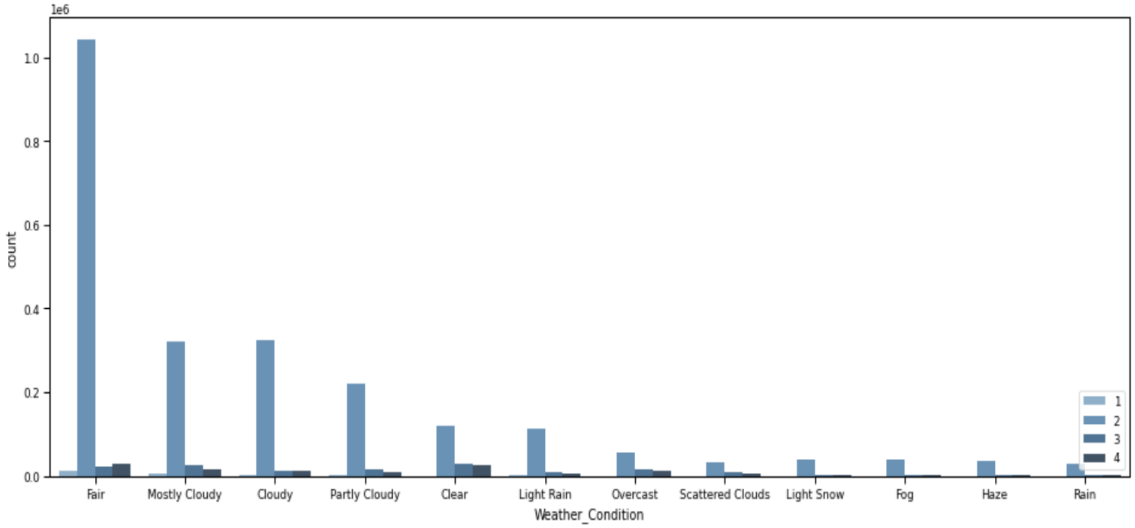
Severity Graph of Geospatial Data



The above Figure 14 represents the severity in the whole United States of America. The Severity 4 is highest in the Eastern region as there is large traffic that is involved in the eastern regions. However, in the complete western region only the Bay area has large severity as that involves more traffic.

Figure 15

Bar Graph of Severity based on the weather



The objective of depicting the graph in Figure 15 is to show the connection between meteorological conditions and accident severity. By examining the countplot, we can determine which meteorological conditions are most frequently associated with various accident severity levels. The color-coding of the bars enables a clear distinction between severity levels, allowing for a visual comparison. This information can be useful for comprehending the impact of weather conditions on the severity of accidents and, potentially, instituting targeted safety measures or interventions to mitigate the effects of adverse weather conditions.

3. Data Preparation

3.1 Imbalanced Classification

In our attempt to predict severity, we encountered imbalanced data, which indicates that severity levels are not distributed uniformly across the dataset. This imbalance can affect the efficacy and precision of our model.

The dataset consists predominantly of severity levels 1, 2, and 3, while severity level 4 is uncommon. This unequal distribution presents a difficulty because the model may not be exposed to sufficient examples of severity level 4 accidents to learn and make accurate predictions and also the severity of 2 is very huge compared to the other levels 2 and 3.
The issue of imbalanced data is significant because it can result in inaccurate forecasts. Our model may have a tendency to predict the majority class more frequently, thereby underestimating the frequency and severity of severe accidents. As a consequence, there is a risk of misallocation of resources, delayed response to serious accidents, and compromised safety measures. To perform binary classification, we merged severity level 1, severity level 2 as 0 and severity 3 and 4 as 1. The below Figure 16 shows the binary classification after merging the severity levels

Figure 16

Binary Classified Data

```
0    2407971
1     247912
Name: Severity, dtype: int64
```

We implemented strategies to resolve the imbalance in the dataset to address this concern. We used oversampling and undersampling to construct a more accurate representation of each severity level. This allowed the model to learn from a wider variety of examples, thereby

enhancing its ability to accurately predict severity across all levels. In evaluating the efficacy of our model, we considered a number of metrics, including accuracy, precision, recall, and F1-score. These metrics provide insight into the performance of the model for both majority and minority classes. Even with the imbalanced data, we were able to evaluate the model's ability to capture severe accidents by concentrating on multiple evaluation metrics.

It is essential to recognize the limitations posed by unbalanced data. Despite the fact that our efforts to correct the imbalance have enhanced model performance, there may be space for additional enhancements. Collecting additional data specifically for incidents of severity level 4 could enhance the model's ability to reliably predict uncommon and critical events.

In conclusion, we acknowledge the imbalanced character of the severity prediction dataset and have taken measures to mitigate its effect on the efficacy of our model. By addressing this issue and contemplating the implications of imbalanced data, we hope to provide more accurate predictions across all severity levels and contribute to improved risk management and decision-making in the assessment of accident severity.

3.2 Dataset Split into Train, Test and Validation

In the beginning, the data are split into two sets: a training set that contains 80% of the data, and a temporary set that contains 20% of the data. The temporary set is then further separated into a validation set, which consists of 25% of the data, and a test set, which contains 75% of the data. The test set is then used.

4. Model Evaluation Metrics

4.1 Precision

Precision, in the context of classification problems, quantifies the proportion of correctly predicted positive instances relative to the total number of positive instances predicted by the model. It is concerned with the accuracy of positive predictions, regardless of the number of actual positive instances. A higher precision value indicates a lower rate of false positives, making it a useful metric for assessing the model's ability to avoid incorrectly classifying negative instances as positive. The formula for this is given in Equation 1.

$$precision = \frac{TP}{TP + FP} \tag{1}$$

4.2 Recall

In the context of classification problems, recall quantifies the proportion of correctly predicted positive instances relative to the number of actual positive instances in the data set. It prioritizes capturing all positive instances while minimizing false negatives. A higher recall value indicates a reduced rate of missing positive instances, making it a useful metric for assessing the model's capacity to identify all relevant positive instances. The formula for this is given in Equation 2.

$$recall = \frac{TP}{TP + FN} \tag{2}$$

4.3 F1 Score

The F1 score is a metric used to evaluate a classification model's performance. Precision and recall are considered to provide a proportionate measure of the model's accuracy. The F1 score is the harmonic mean of precision and recall, ranging from 0 to 1, with a higher value indicating a more effective model. It is especially useful when there is a disparity between classes or when

both precision and recall must be considered simultaneously. The formula for this is given in Equation 3

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{3}$$

4.4 Accuracy

Accuracy, in the context of classification problems, is a measurement of how well a machine learning model reliably predicts the class labels of the test dataset. It represents the proportion of correctly classified instances relative to the total number of instances in a dataset. A greater accuracy indicates the better accuracy of the model in predicting class labels. The formula for this is given in Equation 4

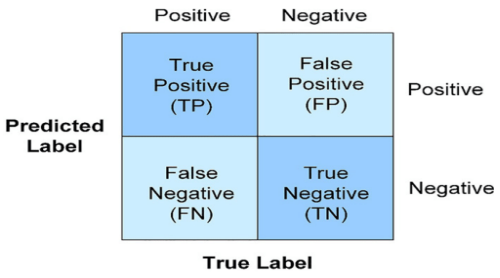
$$accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{4}$$

4.5 Confusion Matrix

In classification problem, a confusion matrix tabulates a machine learning model's performance. It compares model predictions to class labels in detail. The matrix's four values are true positives, true negatives, false positives, and false negatives as shown in Figure 17 help evaluate the model's accuracy, precision, recall, and other performance parameters. The confusion matrix helps evaluate a classification model and suggest areas for improvement.

Figure 17

Confusion Matrix



5. Modeling

5.1 Random Forest

The algorithm we used to predict accident severity is the random forest algorithm. Random forest is a potent and widely-used technique for ensemble learning that provides several benefits for our undertaking. Random forest extends the concept of decision trees by constructing multiple trees and aggregating their predictions. This ensemble method offers numerous advantages for our severity prediction project. Random forest is recognized for its superior predictive accuracy. By combining predictions from multiple trees, the risk of overfitting is reduced and the model's ability to generalize is enhanced. This ensures that our severity predictions are accurate and robust, despite the complexity and diversity of accident data.

In addition, random forest is effective in dealing with imbalanced datasets, a prevalent challenge in severity prediction problems. Class imbalance is mitigated by the algorithm's random sampling of features and observations during tree building. This improves predictions across all severity levels, including minority classes, and precludes model bias towards the majority class. The ability of random forest to capture variable importance is another advantage. The algorithm evaluates the contribution of each feature to the accuracy of the predictions to determine the importance of each feature. This analysis helps us identify the most influential factors on accident severity, allowing us to prioritize interventions and allocate resources more efficiently.

The random forest algorithm provides high predictive accuracy, adaptability to imbalanced datasets, feature importance analysis, robustness to noise and anomalies, and the capacity to manage missing data.

In this project, the random forest classifier is trained using the training set, which allows it to learn the patterns and relationships present in the data. After training the classifier, it is utilized to forecast labels for the validation set. The model's predicted labels are compared to the actual labels, and various evaluation metrics are computed to assess its performance. The metrics that are included are accuracy, recall, precision, F1-score, and a confusion matrix. The evaluation metrics that have been calculated are printed to the console. These metrics provide an overview of the performance of the random forest classifier on the validation set. Afterwards, the classifier that has been trained is utilized to make predictions on the test set's labels. Evaluation metrics are computed for the test set by comparing the predicted labels to the true labels in a similar manner. The metrics include accuracy, recall, precision, F1-score, and a confusion matrix and results are shown in Figure 18 and Figure 19. The evaluation metrics for the random forest classifier's performance on the test set are printed to the console, which provides insights into the model's expected ability to generalize to new, unseen data. This is a significant step towards evaluating the model's effectiveness. Random forest is the highest performing model for us with 86.52% accuracy. All the mentioned advantages played a crucial role and made the random forest standout from the other models.

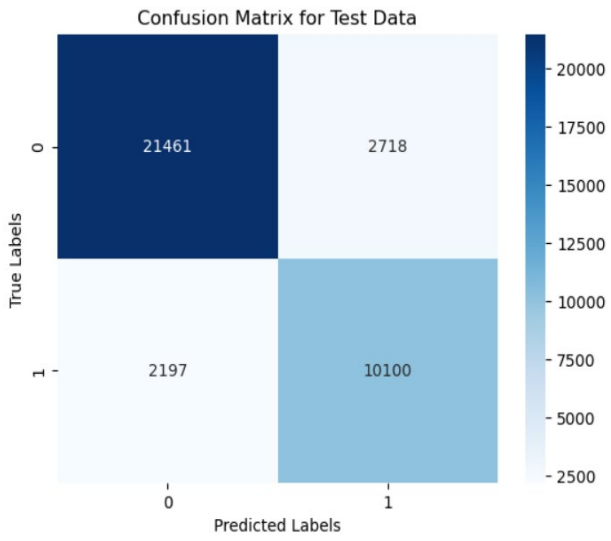
Figure 18

Metric Results for Random Forest

```
Results of Random Forest Classifier on Validation Set
Accuracy    :  0.8662292325407125
Recall      :  0.8662292325407125
Precision   :  0.8674689872698589
F1-score    :  0.8667382581098381
Confusion Matrix   :
[[64519  7919]
 [ 6719 30269]]
Results of Random Forest Classifier on Test Set
Accuracy    :  0.8652538655554337
Recall      :  0.8652538655554337
Precision   :  0.8669561650397047
F1-score    :  0.8659177209363987
Confusion Matrix   :
[[21461  2718]
 [ 2197 10100]]
```

Figure 19

Confusion Matrix of Random Forest



5.2 Decision Tree

We predicted accident severity using the decision tree algorithm. Due to its unique characteristics and capabilities, the decision tree methodology is of vital importance for this project. The interpretability of decision trees is one of their primary advantages. The tree structure represents the decision-making process clearly and intuitively. This clarity enables us to comprehend and explain the factors that contribute to accident severity with ease. The variables that play a significant role in determining severity levels can provide stakeholders, such as traffic authorities and policymakers, with valuable insights. Moreover, decision trees excel at identifying the most crucial severity prediction features. By analyzing the tree's structure, we can identify the variables that have the greatest impact on classifying accidents according to their severity. This information is useful for guiding future data collection efforts, resource allocation, and targeted interventions to effectively prevent severe accidents.

The ability of decision trees to capture nonlinear relationships between variables is another crucial aspect. Accidents in our study are influenced by complex relationships and patterns that may not be readily captured by linear models. The decision tree algorithm is capable of handling these complexities and accurately modeling the various severity-contributing factors. Moreover, decision trees are ideally adapted for dealing with imbalanced datasets, as is the case with our severity prediction project. Without requiring explicit data balancing techniques, the algorithm can learn from both the majority and minority classes. This adaptability ensures accurate forecasts for all severity levels, even when imbalance is present. Additionally, decision trees are resilient to outliers and absent data, which are prevalent obstacles in real-world accident datasets. The algorithm can manage these anomalies without compromising the accuracy of our severity predictions. This robustness enables us to work with actual accident data, allowing us to make accurate and insightful predictions.

The original split for this project's logistic model was 80%/20%; once the model was trained, it was applied to the validation set to predict labels. The accuracy of the model is evaluated by comparing the predicted labels to the true labels and computing a number of evaluation measures. Among these are the F1-score, a confusion matrix, and measures of accuracy, recall, and precision. The decision tree classifier's performance on the validation set is summarized by printing the evaluation metrics it has calculated to the console. Then, the classifier is put to use to make predictions about the test set's labels. Metrics for the test set are also derived through a comparison of expected and actual labels. Accuracy, recall, precision, the F1-score, and a confusion matrix are the measures used. In the end, the test set evaluation metrics for the decision tree classifier are output to the console, providing insight into the model's expected generalization performance on data. So the accuracy we got by implementing

decision tree is 80.92%. Decision tree was predicted the actual values good and their confusion matrix also shows how efficient decision tree is.

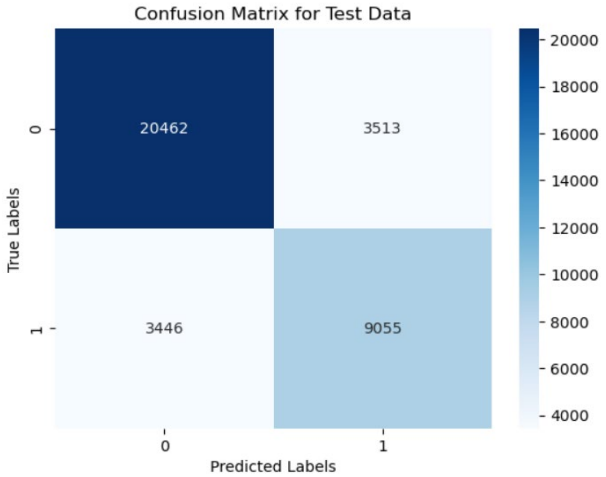
Figure 20

Metric Results for Decision Tree Classifier

```
Results of Decision Tree Classifier on Validation Set
Accuracy    :  0.810209639391004
Recall      :  0.810209639391004
Precision   :  0.8109385932137072
F1-score    :  0.8105519829314608
Confusion Matrix   :
[[61726 10663]
 [10105 26932]]
Results of Decision Tree Classifier on Test Set
Accuracy    :  0.809217019410023
Recall      :  0.809217019410023
Precision   :  0.8094656952116089
F1-score    :  0.8093384889847655
Confusion Matrix   :
[[20462  3513]
 [ 3446  9055]]
```

Figure 21

Confusion Matrix of Decision Tree



5.3 K-Nearest Neighbors

In our project, we also used the K-Nearest Neighbors (KNN) algorithm to predict the severity of accidents. KNN is a non-parametric, instance-based classification algorithm that provides distinct benefits to our modeling strategy. The KNN algorithm's simplicity and

convenience of implementation is one of its most significant assets. It makes no assumptions about the distribution of the underlying data, making it applicable to a wide variety of datasets, including our accident severity dataset. In addition, KNN does not necessitate extensive model training or parameter optimization, saving computational time and resources.

When working with imbalanced datasets, KNN is particularly useful. Due to the fact that KNN makes predictions based on the primary class of the adjacent neighbors, it can effectively manage imbalanced classes. By adjusting the value of K, we are able to influence the balance between sensitivity and specificity, thereby addressing the unbalanced character of our severity prediction data. Using the KNN algorithm's strengths, we can accurately predict the severity of an accident based on its adjacent neighbors and their majority class. This information is invaluable for enhancing safety measures and implementing targeted interventions. Due to these reasons as it may translate their advantages to the project, we chose KNN.

The model is set up in this case to make predictions based on the five closest neighbors. After that, the KNN model is trained on the training data. The model learns from examples of training data, identifying patterns and correlations between input attributes and corresponding labels. The model is ready to make predictions once it has been trained. To obtain the predicted labels, the algorithm applies the trained KNN model to the test data. These are the conclusions reached by the model based on the test data. The accuracy we obtained by using KNN is 78.94%.

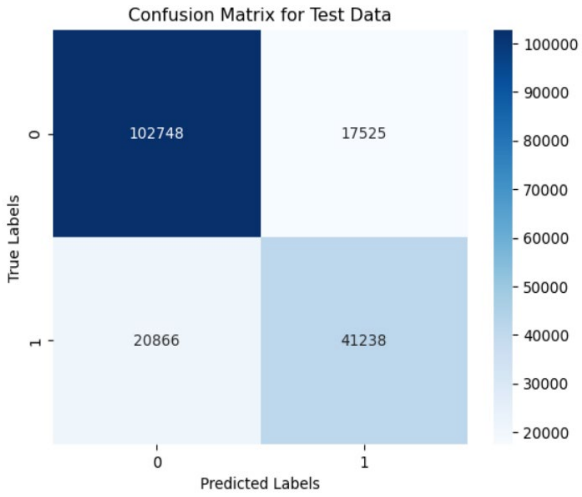
Figure 22

Metric Results for KNN

```
Results of KNN Classifier
Accuracy    :  0.7894964825608493
Recall      :  0.7894964825608493
Precision   :  0.787125388175605
Confusion Matrix   :
[[102748  17525]
 [ 20866  41238]]
```

Figure 23

Confusion Matrix of KNN



5.4 Logistic Regression

The final algorithm used is logistic regression to predict the severity of an accident. Logistic regression is a widely employed statistical modeling technique that provides several benefits for our project. Logistic regression is ideally suited for binary classification problems in which one of two outcomes must be predicted. In our case, we are predicting the severity of accidents, which can be divided into two categories: accidents with high severity and accidents with low severity. Logistic regression also performs reasonably well with imbalanced datasets. Depending on the requirements of our project, we can determine the balance between sensitivity and specificity by modifying the decision threshold. This flexibility enables us to adapt the model to the imbalance nature of our severity prediction data and achieve an optimal balance between accurately predicting accidents with high severity and minimizing false positives.

In addition, logistic regression is computationally efficient and easy to implement. Comparatively to other classification algorithms, it does not demand extensive computational resources or complicated parameter tuning. This makes logistic regression a viable option for our

severity prediction task, particularly when dealing with large data sets. For instance, we can identify incidents with a high likelihood of being severe and accordingly prioritize our resources. By utilizing logistic regression, we can gain insight into the factors influencing the severity of accidents, make informed decisions, and enhance overall safety measures.

In this project the logistic model has been trained using the initial split of 80% and 20% then once the model is trained, it is applied to the validation set to predict labels. Several evaluation metrics are calculated to assess the model's performance based on a comparison between the predicted and actual labels. These metrics consist of precision, recall, F1-score, and a confusion matrix. The evaluation metrics are then output of the logistic regression model on the validation set. Next, labels for the test set are predicted based on the trained model. In a similar fashion, evaluation metrics for the test set are calculated by comparing the predicted labels to the actual labels. The metrics consist of precision, recall, F1-score, and a confusion matrix. Then the evaluation metrics for the performance of the logistic regression model on the test set are output to the console, indicating how well the model is expected to generalize to new, unobserved data. For this project, logistic regression was the weakest model among the others with an accuracy of 68.23%.

Figure 24

Metric Results For Logistic Regression

```
Results of Logistic Regression on Validation Set
Accuracy    :  0.6819311680953338
Recall      :  0.6819311680953338
Precision   :  0.6571835781544563
F1-score    :  0.6454155313446214
Confusion Matrix   :
[[64472  7868]
 [26937 10149]]
Results of Logistic Regression on Test Set
Accuracy    :  0.6823390722666959
Recall      :  0.6823390722666959
Precision   :  0.6576157397795589
F1-score    :  0.6453063837105761
Confusion Matrix   :
[[21534  2587]
 [ 9000  3355]]
```

6. Deployment

These models are deployable on both websites and mobile applications. This may facilitate an intuitive interface for accessing the predictions. This interface enables users to input pertinent information, such as accident details, and receive real-time predictions of severity based on the deployed model.

The developed models can also be integrated into transportation authorities and traffic control centers existing traffic management systems. By predicting the severity of accidents in real-time, these systems may help authorities in prioritizing emergency response, optimizing traffic flow, and implementing measures that reduce the effects of accidents.

Our predictive models can also be used by insurance companies to assess accident severity risks and modify their pricing and coverage accordingly. By integrating precise severity predictions into their risk assessment models, insurers can offer their clients more personalized and customized insurance policies.

7. Discussions

7.1 Results

Table

Comparison Table for the Results

Algorithm	F1-Score	Accuracy
Logistic Regression	0.6453	0.6823
K-Nearest Neighbors (KNN)	0.7735	0.7894
Decision Trees	0.8093	0.8092
Random Forest	0.8659	0.8652

From the above Table it is very clear that Random Forest and Decision Tree performed better than the other algorithms we implemented in our project, including Logistic Regression and K-Nearest Neighbors (KNN), when it came to predicting accident severity. There are several factors which are contributing to their performance advantage over other models.

First, the Random Forest and Decision Tree algorithms belong to the family of ensemble learning algorithms. Random Forest combines multiple decision trees to make predictions, whereas Decision Tree stands alone. It is well known that ensemble methods can reduce bias and variance in predictions, resulting in increased precision. By combining the predictions of multiple trees, Random Forest is able to capture a wider variety of patterns and accomplish improved predictive performance.

Both Random Forest and Decision Tree models can capture nonlinear relationships between features and the objective variable. This is particularly essential for our goal of

predicting accident severity, as there are complex and nonlinear interactions between the various factors contributing to accident severity. These algorithms have the ability to understand and represent these nonlinear relationships, enabling them to capture the data's underlying patterns more accurately. Random Forest and Decision Tree models also support hyperparameter tuning, which involves choosing the optimal combination of hyperparameters to optimize the model's performance. By fine-tuning these parameters using techniques such as Grid Search or Randomized Search, we can increase the predictive ability and effectiveness of the models.All these might be the reasons for the better performance of random forest and decision trees over others.

7.2 Conclusion

In our project, the aim was to predict the severity of accidents in the United States using various machine learning methods and evaluate the performance of Random Forest, Decision Tree, Logistic Regression, and K-Nearest Neighbors (KNN) models. After analysis, it was found that Random Forest and Decision Tree models outperformed the other models in predicting accident severity with accuracies 86.59 % and 80.93% respectively. These models were beneficial because they could capture complex relationships and nonlinearities in the data, enabling more accurate predictions. Random Forest's ensemble approach, which combines multiple decision trees, contributed to its superior performance. However, it's worth noting that Logistic Regression and KNN models also provided valuable insights and predictions, albeit not as strong as the other algorithms. With fair data and no label bias, these models could potentially improve their performance. Future enhancements could be made to refine their formulas and make them more effective. Overall, our project demonstrated the usefulness of different machine

learning techniques in predicting accident severity, with Random Forest and Decision Trees showing particularly promising results.

7.3 Future Scope

For future work, we wanted to add additional features that could improve our models' predictive ability. Exploring external factors such as the weather, roadways, and time of day could provide valuable insight into the severity of accidents. In addition, performing in-depth feature analysis and selection techniques could aid in identifying the most informative variables for accurately predicting severity.In addition, it would be advantageous to evaluate the efficacy of other advanced algorithms for machine learning, such as Support Vector Machines (SVM) and Gradient Boosting models, in predicting accident severity. Given their capability to manage complex data patterns and relationships, these algorithms may provide additional enhancements and insights.

7.4 Source Code

Github Link: https://github.com/KalyanVikkurthi002/Machine_Learning_GRoup_7

References

M, A., K, A., K, A., M, A., & R, C. K. (2022). *Accident Prediction Using KNN Algorithm*. https://doi.org/10.1109/icerect56837.2022.10059746

Nasri, M., Aghabayk, K., Esmaili, A., & Shiwakoti, N. (2022). Using ordered and unordered logistic regressions to investigate risk factors associated with pedestrian crash injury severity in Victoria, Australia. *Journal of Safety Research*, *81*, 78–90. https://doi.org/10.1016/j.jsr.2022.01.008

VA, O., & AA, E. (2014). Traffic Accident Analysis Using Decision Trees and Neural Networks. *International Journal of Information Technology and Computer Science*, *6*(2), 22–28. https://doi.org/10.5815/ijitcs.2014.02.03

Yan, M., & Shen, Y. (2022). Traffic Accident Severity Prediction Based on Random Forest. *Sustainability*, *14*(3), 1729. https://doi.org/10.3390/su14031729