**Final Project Report  
Project Title: US Accidents Data Analysis  
Pratik Harlikar, Aditya Jadhav, Yatziri Presmont, Muhammad Batikh  
  
1. Introduction :  
Objective :**Finding the main causes of traffic accidents, comprehending geographical and temporal trends, and providing practical suggestions for enhancing road safety are the goals of the analysis of the US Accidents dataset. The goal of the analysis is to find trends in the frequency of accidents, such as how they relate to weather patterns, times of day, and geographic regions. It also seeks to evaluate the seriousness of accidents in a variety of settings, including diverse regions and traffic situations. By recognizing these patterns, the analysis will offer important information about the locations and times of incidents, which will assist guide initiatives for improving road safety, lowering the number of accidents, and lessening the effects of accidents on infrastructure and public health. To make the study computationally viable while preserving typical insights, a selection of 10,000 records from the original 7 million cases was used.

**Dataset Overview:**Over 7.7 million traffic accident records, collected from multiple public data feeds between February 2016 and March 2023, make up the US Accidents Dataset. The collection, which spans 49 states in the US, is sourced from several APIs that compile information from organizations including law enforcement, traffic sensors, and cameras, as well as the US and state Departments of Transportation. Along with the time of the incident, it contains comprehensive data on the accident's location (latitude and longitude), weather (temperature, wind speed, visibility), and severity classifications (minor, severe, and deadly). This large dataset is an essential tool for traffic safety analysis, predictive modeling, and policy-making since it offers insightful information on the variables affecting traffic accidents. The dataset is available: [US Accidents](https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents/data)

**2. Data :  
2.1 What Kind of Data?**A combination of numerical, category, and temporal data are included in the dataset. Important characteristics include weather conditions (categorical), time (temporal), location (numerical, latitude and longitude), and accident severity (categorical). Comprehensive examination of accident incidence and contributory causes is made possible by this combination of features.  
  
**2.2 Dataset Information:**The dataset contains several attributes (features) and a target attribute:  
 ●   Number of Instances: The dataset contains over **7 million records**, each representing a unique traffic accident. This large number of records ensures a comprehensive dataset for analysis.  
 ●   Number of Attributes : There are a total of **45 attributes** in the dataset. These attributes are a mix of **numerical** and **categorical** data, allowing for a well-rounded analysis of factors influencing traffic accidents.  
 ●   Type of Attributes: The attributes in the dataset provide a wide-ranging view of the circumstances surrounding traffic accidents. These attributes include:

* **Numerical Attributes**: These include continuous or discrete values, such as:
  + **Temperature**: The ambient temperature at the time of the accident.
  + **Visibility**: The visibility distance during the accident event.
  + **Speed Limits**: Speed limits in the accident area.
* **Categorical Attributes**: These include data that falls into distinct categories, such as:
  + **Weather Conditions**: Categories like sunny, rainy, foggy, etc.
  + **Accident Severity**: Describes the severity of the accident, such as minor, severe, or fatal.
  + **Location Details**: Information on the type of road (e.g., highway, city street) and whether the area is urban or rural.
  + **Traffic Conditions**: Describes the density of traffic at the time of the accident, such as light, moderate, or heavy.

**Subset for Analysis:**

Given the size of the original dataset, an **analysis was performed on a subset of 10,000 instances** for manageability and computational efficiency. This subset was selected to maintain the core characteristics of the full dataset while making the analysis more feasible for exploratory and preprocessing tasks. The reduced sample allows for faster data processing and model training while still providing insights into the patterns and relationships in the data.

**2.3. Data Processing:**The data is processed in several steps:

* **Exploratory Data Analysis (EDA):** To comprehend connections and patterns, a variety of EDA approaches were used, such as data summarization (mean, median, variance) and visualization techniques including heatmaps, scatter plots, and boxplots..
* **Handling Missing Values**: Depending on the type of feature, multiple approaches were used to handling missing values. The mean value was used to impute missing values for numeric columns, while the mode was used to impute missing values for categorical columns. Rows that had substantial missing data were occasionally removed. To guarantee data quality, for instance, entries with missing values in critical fields like "Temperature(F)" and "Visibility(mi)" were eliminated because these characteristics were critical for comprehending accident conditions.
* **Outlier Detection and Treatment**: Using the Interquartile Range (IQR) approach, outliers were identified. To preserve data consistency, the outliers were either eliminated or limited at a threshold. Outliers were capped or dealt with as necessary after columns including "Temperature(F)," "Visibility(mi)," "Humidity(%)," "Windspeed(mph)," "Pressure(in)," and "Distance(mi)" were examined.
* **Categorical Feature Encoding**: To make them appropriate for building models, categorical features—like weather conditions—were encoded using label encoding or one-hot encoding.
* **Normalization**: StandardScaler and MinMaxScaler were used to normalize numerical features such that each feature contributed equally to the model. The distribution of the feature values influenced the scaler selection.

**2.4. Attributes Usage:**

* **Used Attributes**: We make use of the majority of characteristics that have a direct impact on accident analysis, including the accident time, geographical coordinates, weather, and road conditions. These have a major influence on the prediction models' results, making them essential for modeling purposes.
* **Unused Attributes:** Certain attributes are not included since they don't improve the analysis or model performance, such as the external dataset ID or redundant location information (such as exact address names when coordinates are present).

**Justification for Attribute Selection:**

**1) Reasons for Using Selected Attributes:**

Attributes like weather conditions and time of the accident are used because they are known to have a significant impact on accident likelihood and severity.

**2) Reasons for Excluding Certain Attributes :**

non-informative attributes, such as unique identifiers that do not affect accident outcomes, are excluded to simplify the model and improve computational efficiency

**3. Data Mining Task :**The main data mining task in this analysis is classification, which can be approached as either a multiclass or binary classification problem.

* **Multiclass Classification Problem**: This approach involves classifying accidents into different severity levels based on factors like weather conditions, road visibility, time of day, and location.
* **Binary Classification Problem**: Alternatively, the task can be simplified to predicting whether an accident is severe or not, such as distinguishing between severity levels 3 and 4 versus 1 and 2.

In order to prepare the dataset for the project, it was pre-processed by handling outliers, encoding categorical features, selecting features, and resolving missing values. This project's goal is to examine the US Accidents dataset in order to find important variables affecting traffic accidents, as well as to find patterns and trends and use classification models to forecast the severity of accidents. Classification is the main data mining work here, with an emphasis on forecasting the severity of accidents based on variables including location, time of day, weather, and road visibility.

**4. Analysis and Results:  
4.1 EDA and Data Preprocessing:**This section will discuss the EDA and preprocessing techniques used in our project  
 **4.1.1 Data Description:   
A screenshot of a computer program

Description automatically generated**

**4.1.2 Visualization of null values:**A graph with yellow lines and white text

Description automatically generatedA screenshot of a computer

Description automatically generated  
  
The above graph indicates several features contain missing values depicted using yellow color and the percentage of missing data for numerical and categorical columns in the dataset  
 **Findings:** The graph displays missing values in a number of columns; non-missing values are indicated by purple stripes, while missing data is represented by yellow stripes. Some columns, like 'Temperature(F)' and 'Visibility(mi)', contain fewer gaps in data than others, such as 'End\_Lat' and 'End\_Lng'. This emphasizes the necessity of cautious feature imputation or removal because incorrect treatment of missing data may result in biases or inaccuracies.

**4.1.3 Correlation Analysis:**

**Cor graph:   
A screenshot of a graph

Description automatically generated**

The heatmap provided represents the correlation between different numerical features in the dataset. Here's a brief interpretation:  
**Correlation Coefficient**: Each cell in the heatmap shows the correlation coefficient between the features, ranging from -1 to 1.

* + **Red Shades**: Indicates a strong positive correlation (close to 1). For example, 'Temperature(F)' and 'Wind\_Chill(F)' are highly positively correlated.
  + **Blue Shades**: Indicates a strong negative correlation (close to -1). For example, 'Temperature(F)' and 'Humidity(%)' are negatively correlated.
  + **Lighter Colors**: Values closer to 0 indicate a weak or no correlation between features.
* **Key Observations**:
  + Features like 'Start\_Lat' and 'Temperature(F)' have a notable negative correlation.
  + 'Severity' has low correlations with most features, suggesting that accident severity is influenced by a combination of factors rather than a single strong predictor.

**4.2 Descriptive Analysis of Key Factors**:

**4.2.1 – Overall Number of Accidents In the Country :**

**A map of the united states

Description automatically generated**

You can interact with the Zoomable version of this plot here [[file:///C:/Users/HP/AppData/Local/Microsoft/Windows/INetCache/IE/AT089W81/zoomable\_us\_accident\_map[1].html](file:///C:/Users/HP/AppData/Local/Microsoft/Windows/INetCache/IE/AT089W81/zoomable_us_accident_map%5b1%5d.html) ]

**Insight :** This map shows clustered data points, color-coded by magnitude or density, throughout the United States. Higher concentrations of data points are shown by larger clusters with warmer hues (such as orange and red), like in the Northeast, Texas, and California. In areas such as the Midwest and Mountain West, smaller clusters with cooler hues (such as green or yellow) have lower concentrations. The graphic illustrates national trends in population or event density.

**4.2.2 – Overall Number of Accidental Hotspots:**

You can interact with the Zoomable version of this plot here

[[file:///C:/Users/HP/AppData/Local/Microsoft/Windows/INetCache/IE/H4Q6B1GS/accident\_hotspots[1].html](file:///C:/Users/HP/AppData/Local/Microsoft/Windows/INetCache/IE/H4Q6B1GS/accident_hotspots%5b1%5d.html) ]

**A map of the united states with many colored circles

Description automatically generated**

**Insight :** The map combines cluster markers with a heatmap overlay to visualize data density across the United States. The heatmap highlights areas of high intensity in blue and red, with notable concentrations along the coasts (e.g., California, the Northeast) and urban hubs like Los Angeles, Dallas, and Atlanta. The clustering reinforces these hotspots while providing specific data counts for more granular insights. This view is useful for identifying both broad trends and specific dense locations.

**4.2.3 Number of Accidents per Year**:  
The total number of accidents each year provides insight into overall trends in accident occurrences. Below is a bar chart showing the number of accidents per year:

A graph of a number of accidents

Description automatically generated

A number of numbers on a white background

Description automatically generated

* The bar chart and Output results indicate an increasing trend in accidents over the years, with a peak in 2021.
* This rise could be attributed to factors such as changes in traffic volumes, road conditions, or weather patterns.
* Notably, the significant drop in 2023 may reflect incomplete data collection or changes in reporting practices.

A pie chart with numbers and a number of different colors

Description automatically generated**4.2.4 Number of Accidents per State**:  
We also examined the distribution of accidents across different states to understand which states experienced the highest proportions of accidents

The pie chart shows that California has the highest percentage of accidents (22.7%), followed by Texas (18.2%), Florida (13.6%), and New York (13.6%).

* States like Illinois, Ohio, and Oregon each have smaller shares of around 4.5%.
* This distribution indicates that highly populated and high-traffic states tend to have a higher number of accidents.

**4.3 Accident Hotspots**:  
The bar chart below shows the top 20 states with the most accidents, with California leading by a large margin.

**A graph of a number of accidents

Description automatically generated**

* The chart highlights that California, Florida, and Texas have the highest number of accidents, which may be attributed to their larger populations and extensive road  
  networks.
* This Plot makes it clear where most accidents occur, helping in the spatial analysis of accident distribution.
* Let’s choose one state from Top 20 and Do the further Analysis.

**4.4 Targeting Solo State Analysis :**

**Objective:** To Perform an in-depth analysis of accidents in a specific state .  
For this section, we target a solo state **Texas** to understand the unique patterns of accidents in that state.

**4.4.1 Number of Accidents by Month in Texas**:  
A graph of accident statistics

Description automatically generated

**Insights**: The number of accidents in Texas varies by month, with peaks in November and December, likely influenced by winter conditions and increased holiday travel.

**4.4.2 Number of Accidents by Weekday in Texas**:

A graph showing the days of week

Description automatically generated with medium confidence

**Insights**: Accidents are most frequent on weekdays, especially on Fridays, indicating higher risks during work commutes. Accidents tend to decrease on weekends.

**4.4.3 Number of Accidents by Hour in Texas**:

A graph showing the number of accident victims

Description automatically generated with medium confidence

**Insights**: Most accidents occur during rush hours, particularly between 7-9 AM and 4-6 PM, highlighting the need for targeted traffic safety measures during peak commuting times.

Overall, this section provided a detailed look into accidents occurring in Texas, with analyses by month, by day of the week, and by hour. From the analysis, it is evident that certain months such as November and December have a higher occurrence of accidents, possibly due to winter conditions and holiday travel. Accidents are also more frequent on weekdays, especially Fridays, indicating higher risks during commuting hours. Moreover, most accidents happen during morning and evening rush hours, highlighting the need for targeted interventions during these times to enhance traffic safety.  
  
**Weather Analysis of Texas:**  
**Number of Accidents Due to Weather Condition in Texas :**

The 3D scatter plot below shows the distribution of accidents in Texas based on weather conditions. The plot includes temperature (°F), longitude, and latitude to visualize how weather factors interact with accident locations. Each point is color-coded based on the weather condition during the accident, such as clear, rain, snow, fog, etc.

You can interact with the 3D version of this plot here: [ [file:///C:/Users/HP/AppData/Local/Microsoft/Windows/INetCache/IE/MWC1PA1B/texas\_weather\_3d\_plot[2].html](file:///C:/Users/HP/AppData/Local/Microsoft/Windows/INetCache/IE/MWC1PA1B/texas_weather_3d_plot%5b2%5d.html) ]

A colorful dots on a white background

Description automatically generated

The interactive visualization reveals how different weather conditions are distributed across various parts of Texas, indicating a possible influence on accident frequency.

**4.5 Anomaly Detection :**

The box plots provided for each variable ("Temperature", "Visibility," "Humidity," "wind Speed," "pressure," "Precipitation," "Distance", “Severity” etc) show the distribution of values in the dataset.

**4.5.1 Outlier Detection :**Ensuring data consistency and eliminating extreme values that could skew model training are the main goals of outlier identification. Identifying outliers in numerical features by applying the Interquartile Range (IQR) approach.

A graph of a temperature

Description automatically generated A graph of a line

Description automatically generated with medium confidence  
  
A graph with a blue rectangle

Description automatically generated A graph of a wind speed

Description automatically generated

* **Temperature(F)**: The boxplot shows that there are several extreme values on both ends. Temperatures below -50°F or above 150°F are considered unrealistic and treated as outliers.
* **Visibility(mi)**: Visibility values above 50 miles are highly unusual and were identified as outliers, as shown by the boxplot.
* **Humidity(%)**: Most of the data lies between 50% and 90%. Extreme low or high humidity values were capped to maintain consistency.
* **Wind\_Speed(mph)**: Wind speeds above 100 mph were considered outliers. These values were treated to prevent distortion in model training.

A graph of pressure

Description automatically generated A graph of a number of objects

Description automatically generated

* **Pressure(in)**: Outliers were detected in pressure values below 20 and above 35 inches. These extreme values were treated to maintain consistency in the dataset.
* **Precipitation(in)**: Precipitation values above 5 inches were considered outliers. Such extreme values were capped to avoid skewing the data.

A graph of a number of objects

Description automatically generated A graph with a line

Description automatically generated

A blue square with white text

Description automatically generatedA blue rectangle with white text

Description automatically generated

* **Distance(mi)**: Distances above 100 miles were identified as outliers, as these extreme values are unusual for typical accidents. These values were capped to ensure data consistency.
* **Severity**: The boxplot shows that severity levels 1, 3, and 4 have a few extreme values that were treated to maintain balance in the distribution.
* **Start\_Lat** and **Start\_Lng**: Latitude and longitude values were generally within expected ranges, with no significant outliers observed that required treatment.

**4.5.2 Outlier Treatment:**Outliers were capped to the nearest acceptable values within the IQR range to reduce skewness. - For example, for `'Temperature(F)'`, values beyond the calculated bounds were replaced with appropriate boundary values**.**

**Results After Outlier Handling**:

A blue rectangular object with white text

Description automatically generated A graph of a line

Description automatically generated with medium confidence

A graph of humidity

Description automatically generated A blue graph with a black line

Description automatically generated with medium confidence

* **Temperature(F)**: After outlier handling, the temperature values were brought within a more reasonable range between 20°F and 120°F, with extreme lows and highs removed. This provides a more accurate representation of the conditions during accidents.
* **Visibility(mi)**: Outliers in visibility values were handled, capping the values between 9.6 and 10.4 miles, which helps standardize data for better analysis.
* **Humidity(%)**: After handling outliers, humidity values are capped between typical ranges, reducing the influence of extreme humidity readings that could distort predictions.
* **Wind\_Speed(mph)**: Wind speed values above 100 mph were considered unrealistic for the dataset. Post-handling, values range reasonably between 5 and 17 mph.

A blue rectangle with white text

Description automatically generated A graph with lines and numbers

Description automatically generated

* **Pressure(in)**: The pressure values after handling fall within 28.5 to 31 inches, removing the unrealistic outliers that could potentially skew the analysis.
* **Precipitation(in)**: Extreme precipitation values were removed, capping them at realistic values around 0, thereby ensuring more consistent weather data.

A blue rectangle with black text

Description automatically generatedA graph with numbers and lines

Description automatically generated

A blue rectangle with white text

Description automatically generated A blue square with white text

Description automatically generated

* **Distance(mi)**: Most accidents involve short distances. Post-treatment, the maximum distance is capped at 1.2 miles, focusing on more common accident scenarios.
* **Severity**: Severity values were standardized to remove extreme instances, providing a clearer representation of accident severity trends.
* **Start\_Lat and Start\_Lng**: No significant outliers were found for these attributes, and the final values remain within expected ranges for the dataset.

Because outliers can occasionally distort the outcomes of statistical models, managing them helps to standardize the data and may make it more usable for study. It's crucial to remember that the context and any effects on the analysis should be carefully considered when deciding whether to eliminate or modify outliers.

**4.6 Classification Section:**This section depicts all the results for the binary and multiclass classifiers used in our project.

**4.6.1 Binary classifier models**

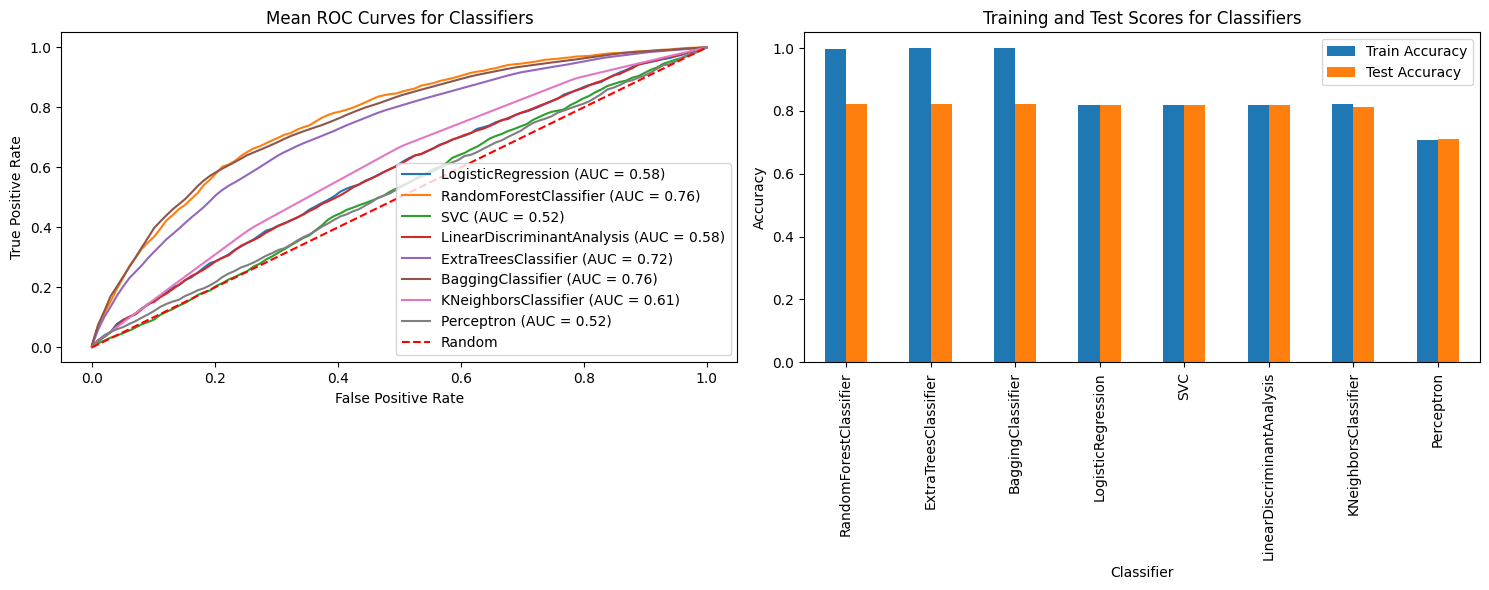
**Table 1:** Performance matrices of various binary classifiers used in our project

A white sheet with black and yellow text

Description automatically generated

* **Rankings**: The classifiers are ranked based on test accuracy, with RandomForestClassifier and ExtraTreesClassifier achieving the highest performance metrics.
* **Accuracy**: RandomForestClassifier achieves the highest train accuracy (0.99585) and a decent test accuracy (0.8224), suggesting that it effectively captures patterns in the data without overfitting.
* **Precision**: Precision is the ratio of true positives to the sum of true and false positives. It is a measure of the accuracy of the positive predictions.
* **Recall**: Recall, also known as sensitivity, is the ratio of true positives to the sum of true positives and false negatives. It shows the ability of the classifier to find all the positive samples.
* **F1 Score**: The F1 Score is the harmonic mean of precision and recall, providing a single score that balances the two metrics. It is especially useful when the class distribution is imbalanced.
* **AUC**: The Area Under the Receiver Operating Characteristic Curve (ROC AUC) measures the ability of a classifier to distinguish between classes. A value of 0.5 suggests no discriminative ability, while a value of 1 suggests perfect discrimination.

**Findings**: Based on the analysis, RandomForestClassifier and ExtraTreesClassifier showed strong performance metrics, including high test accuracy and balanced F1 scores. These models were effective at capturing complex patterns without overfitting. However, other models such as LogisticRegression and SVC showed limited recall and F1 scores, indicating challenges in identifying positive instances accurately. The combination of high accuracy and reasonable AUC values for the top models suggests they are well-suited for predicting accident severity, making them promising candidates for future applications in road safety   
interventions. A value of 0.5 suggests no discriminative ability, while a value of 1 suggests perfect discrimination.  
  
**Graph 1**: ROC and Train/Test accuracy figures for the binary classifier.



The graphs provided above offer a visual comparison of different classifiers' performance using ROC curves and accuracy measurements.

**Mean ROC Curves for Classifiers (Left Graph)**:

* The ROC curve graph displays the trade-off between sensitivity (True Positive Rate) and specificity (1 - False Positive Rate). Each line represents a classifier's performance across all possible threshold levels.
* The AUC values range from 0.52 to 0.76, with higher values indicating better overall performance. AUC values closer to 1 suggest that the classifier is better at distinguishing between the positive and negative classes.
* The Random Forest Classifier and Bagging Classifier show the highest AUC values (0.76), suggesting they have the best performance in terms of distinguishing between the classes for the given problem.
* A classifier's line closer to the top-left corner indicates higher sensitivity and specificity, meaning it is more capable of correctly classifying positive and negative cases.
* The dashed line represents the performance of a random classifier, with an AUC of 0.5. Any classifier performing close to this line is no better than random guessing.

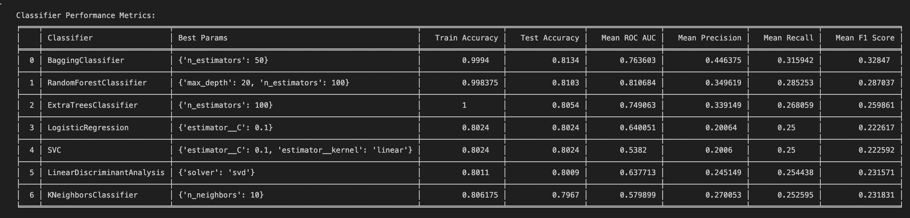
**Training and Test Accuracies for Classifiers (Right Graph)**:

* The bar chart compares the training and test accuracies of each classifier. Ideally, these should be high and close to each other to indicate a model that learns well and generalizes well.
* A large gap between training and test accuracy might indicate overfitting, where the model learns the training data too well, including noise and outliers, which do not generalize to new data.
* The classifiers in the graph mostly show a small gap between training and test accuracies, suggesting that they generalize well without significant overfitting.
* It's also important to note that while high accuracy is desirable, it is not the only measure of a good classifier, especially if the dataset is imbalanced. This is where the ROC and AUC metrics provide additional insights.

**Findings**: With a high AUC and closely matching training and test accuracies, the Random Forest Classifier is one among the top performers for the given problem, according to these interpretations. But in order to draw a firm conclusion, one would also need to take into account other elements including precision, recall, F1 score, and the problem's context, which includes class balance and the costs of various error types.

**4.6.2 Multiclass classifier models**

**Table 2**: Performance matrices of various multiclass classifiers used in our project



**Classifier**: This column names the machine learning algorithms that have been used for classification.

**Accuracy**: These metrics indicate how well the classifier performs on the training set versus the test set. Higher training accuracy compared to test accuracy might indicate overfitting.: RandomForestClassifier achieves the highest train accuracy (0.99585) and a decent test accuracy (0.8224), suggesting that it effectively captures patterns in the data without overfitting.  
**Precision:**Precision is about how precise/accurate the model is out of those predicted positives, and how many of them are actually positive. Precision is a good measure to determine when the cost of a False Positive is high. The RandomForestClassifier has the highest precision in the table.

**Recall :** Recall calculates how many of the Actual Positives our model captures by labeling it as Positive (True Positive). High recall means most of the positive examples are correctly recognized (low false negatives). Perceptron and Bagging Classifier have the highest recall.  
  
**F1 Score:** The F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. It is a good way to show that a classifier has a good value for both recall and precision. In this case, the RandomForestClassfier and BaggingClassifier have the highest F1 scores.  
  
**Hyperparameters:** This column shows the settings for each classifier. Hyperparameters are the configuration settings used to structure the machine learning model. These are set before the model runs and can affect the performance of the model. We used a random search to find the best hyperparameter for each model and listed them in Table 2.

**Graph 2**: ROC and Train/Test accuracy figures for the Multiclass classifier.

A close-up of a graph

Description automatically generated

The graphs provided above offer a visual comparison of different classifiers' performance using ROC curves and accuracy measurements.

**Mean ROC Curves for Classifiers (Left Graph)**:

* The ROC curve graph displays the trade-off between sensitivity (True Positive Rate) and specificity (1 - False Positive Rate). Each line represents a classifier's performance across all possible threshold levels.
* The AUC values range from 0.52 to 0.76, with higher values indicating better overall performance. AUC values closer to 1 suggest that the classifier is better at distinguishing between the positive and negative classes.
* The Random Forest Classifier and Bagging Classifier show the highest AUC values (0.76), suggesting they have the best performance in terms of distinguishing between the classes for the given problem.
* A classifier's line closer to the top-left corner indicates higher sensitivity and specificity, meaning it is more capable of correctly classifying positive and negative cases.
* The dashed line represents the performance of a random classifier, with an AUC of 0.5. Any classifier performing close to this line is no better than random guessing.

**Training and Test Accuracies for Classifiers (Right Graph)**:

* The bar chart compares the training and test accuracies of each classifier. Ideally, these should be high and close to each other to indicate a model that learns well and generalizes well.
* A large gap between training and test accuracy might indicate overfitting, where the model learns the training data too well, including noise and outliers, which do not generalize to new data.
* The classifiers in the graph mostly show a small gap between training and test accuracies, suggesting that they generalize well without significant overfitting.
* It's also important to note that while high accuracy is desirable, it is not the only measure of a good classifier, especially if the dataset is imbalanced. This is where the ROC and AUC metrics provide additional insights.

**Findings**: Based on the analysis, RandomForestClassifier and ExtraTreesClassifier showed strong performance metrics, including high test accuracy and balanced F1 scores. These models were effective at capturing complex patterns without overfitting. However, other models such as LogisticRegression and SVC showed limited recall and F1 scores, indicating challenges in identifying positive instances accurately. The combination of high accuracy and reasonable AUC values for the top models suggests they are well-suited for predicting accident severity, making them promising candidates for future applications in road safety interventions.  
  
**Comparison of test accuracies of various multiclass classifiers :**   
A graph of different colored bars

Description automatically generated

The bar chart compares the accuracy of various classifiers. Accuracy is a measure of how often the classifier makes the correct prediction, regardless of class.  
  
**Findings:  
Bagging Classifier :** This classifier has the highest accuracy among those listed, which is reflected by the tallest bar in the chart. It indicates that for this specific dataset, BaggingClassifier algorithm outperforms the others in making correct predictions

**Random Forest Classifier and SVC** : These classifiers have slightly lower accuracy than the Bagging Classifier but still perform relatively well compared to the others.

**Logistic** **Regression** : From Logistic Regression to KNN All the Classifiers have the relatively Fair Accuracy In Which

Logistic Regression being highest performing and Perceptron being Fairly Low Performing classifier.

**Perceptron:** This classifier has the lowest accuracy of all the classifiers shown. This could indicate that the Perceptron model is too simple to capture the complexities of the dataset or that it may need to be configured with the optimal parameters.

The bars' varying colors appear to have no other purpose except to facilitate reading of the chart and aid in differentiating between the different classifiers. It is clear from this picture that ensemble approaches such as Random Forest, Bagging Classifier, and SVC typically exhibit superior accuracy on this dataset. It's crucial to remember that when assessing classifiers, there are other metrics to consider besides accuracy. Other measures, including as precision, recall, and the F1 score, may be more pertinent depending on the application, particularly if the dataset is unbalanced.

**5. Conclusion**

The analysis of the US Accidents dataset provided comprehensive insights into factors contributing to road accidents across the United States. Key findings highlight the significant influence of factors such as weather conditions, time of day, and geographic location on accident frequency and severity. High accident rates were observed during peak traffic hours and under adverse weather conditions, emphasizing the need for targeted safety interventions.

The data mining tasks focused on both binary and multiclass classification, with RandomForestClassifier and ExtraTreesClassifier emerging as top-performing models based on metrics like accuracy, precision, and AUC. These models showed strong capabilities for predicting accident severity, suggesting their potential use in proactive road safety measures.

Overall, this analysis offers valuable insights to inform policies and strategies aimed at reducing accident rates and improving road safety.

**6. Reference :**

* **Dataset**: The US Accidents dataset used in this report can be accessed at [Kaggle - US Accidents Dataset](https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents).