Road Sense - Advanced Predictive Modeling for Traffic Safety

Harvineet Singh (002814713)

Pranav Harish Sharma (002851959)

The dataset used in this project is a cleaned and sampled version of a large-scale U.S. traffic accident dataset, which contains records of accidents from various states across the country. The data was originally collected from multiple sources, including law enforcement reports, traffic cameras, and weather stations.

Dataset Source: [US Accidents (2016 - 2023) (kaggle.com)]

Introduction

Contextual Overview: Road traffic injuries are a major global concern, causing numerous fatalities and injuries annually, leading to significant economic and social burdens.

Role of Predictive Modeling: Predictive modeling has become a key tool in traffic safety, helping forecast accidents and their severity by analyzing historical data, thus aiding in proactive traffic management.

Project Scope: The "RoadSense" project aims to utilize advanced machine learning techniques to predict traffic accident severity based on factors like weather, time, and road characteristics, providing valuable insights for traffic authorities.

Problem Statement

Global Traffic Accident Statistics: Each year, road traffic crashes result in approximately 1.35 million deaths and tens of millions of injuries, particularly in low- and middle-income countries where infrastructure development lags behind rapid urbanization and motorization.

Challenges in Traffic Safety: Accidents are hard to prevent due to unpredictable road conditions, human behavior, and environmental factors. Traditional safety management often involves costly, reactive measures like post-accident investigations and infrastructure updates.

Research Question: How can machine learning models be employed to predict the severity of traffic accidents, and what factors are critical for enhancing the accuracy and reliability of these predictions? RoadSense explores these questions by analyzing a detailed dataset of U.S. traffic accidents.

PROJECT OBJECTIVES

- Develop a Predictive Model for Accident Severity
- Identify Key Influencing Factors
- Contribute to Traffic Safety Research

Dataset Overview

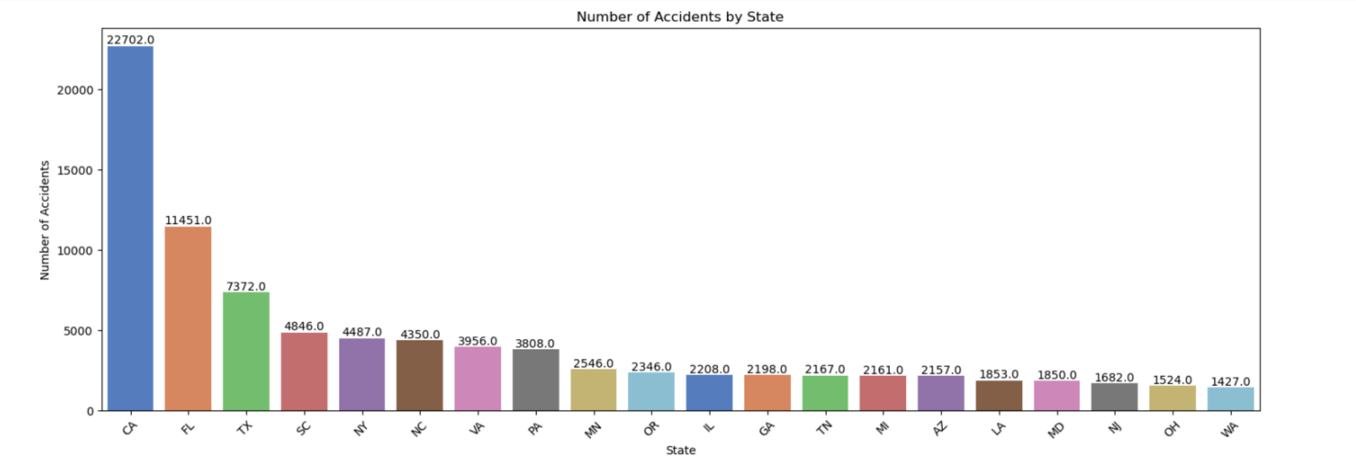
Column Name	Description
Severity	Indicates the severity level of the accident, with values ranging from 1 (least severe) to 4 (most severe).
Weather_Condition	Describes the weather conditions at the time of the accident (e.g., Clear, Rain, Snow).
Light_Condition	Describes the lighting condition at the time of the accident (e.g., Daylight, Dark, Dawn, Dusk).
Start_Lat	The latitude coordinate of the location where the accident started.
Start_Lng	The longitude coordinate of the location where the accident started.
Temperature (F)	The temperature at the accident location in degrees Fahrenheit.
Visibility (mi)	The visibility distance at the accident location in miles.

Exploratory Data Analysis (EDA) Overview

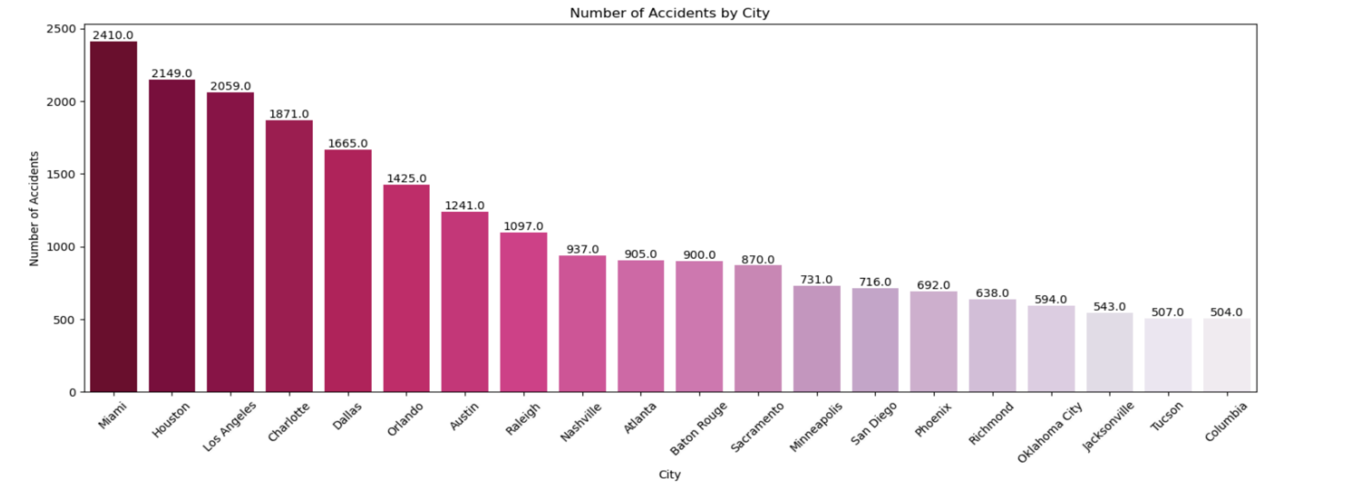
Purpose of EDA in RoadSense:
•To identify trends, patterns, and anomalies in the dataset.
•Provide insights to inform feature selection and modeling.

Techniques used:
•Visualizations: Histograms, scatter plots, and heatmaps.
•Statistical analysis to understand data distribution and feature relationships.

EDA

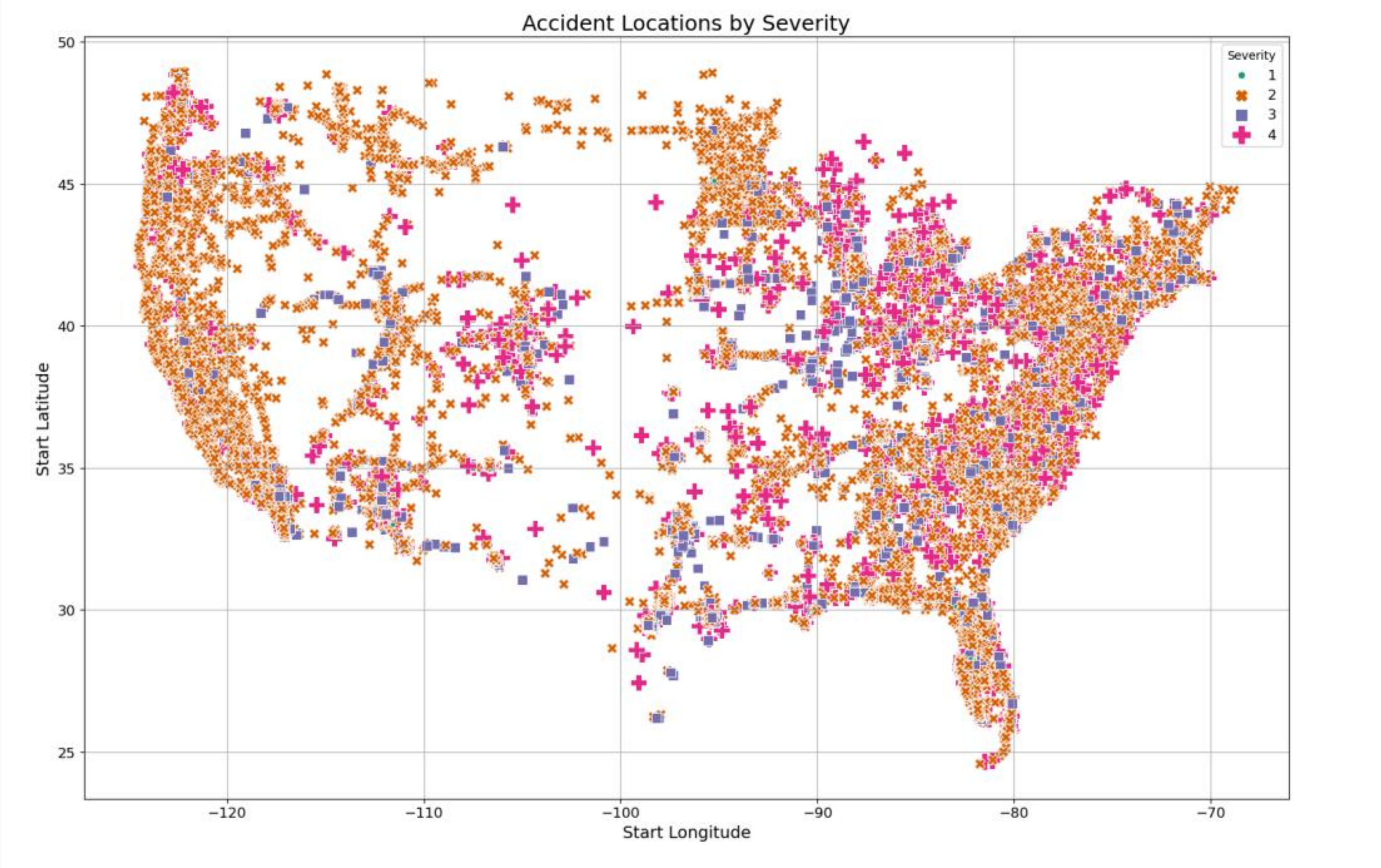


California (22702) has the highest number of Accidents per States



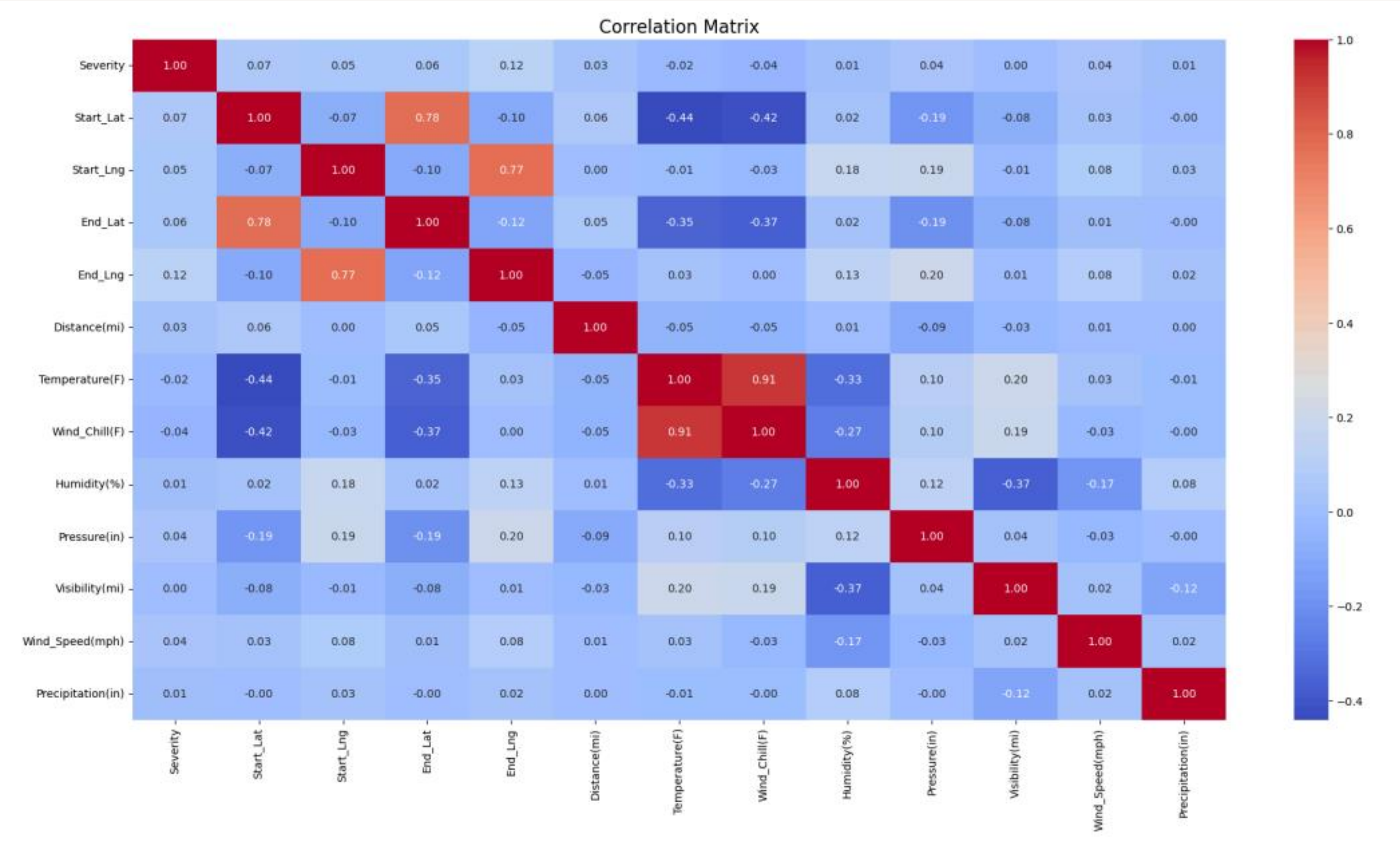
Maimi (2410) has the highest number of Accidents in California

EDA



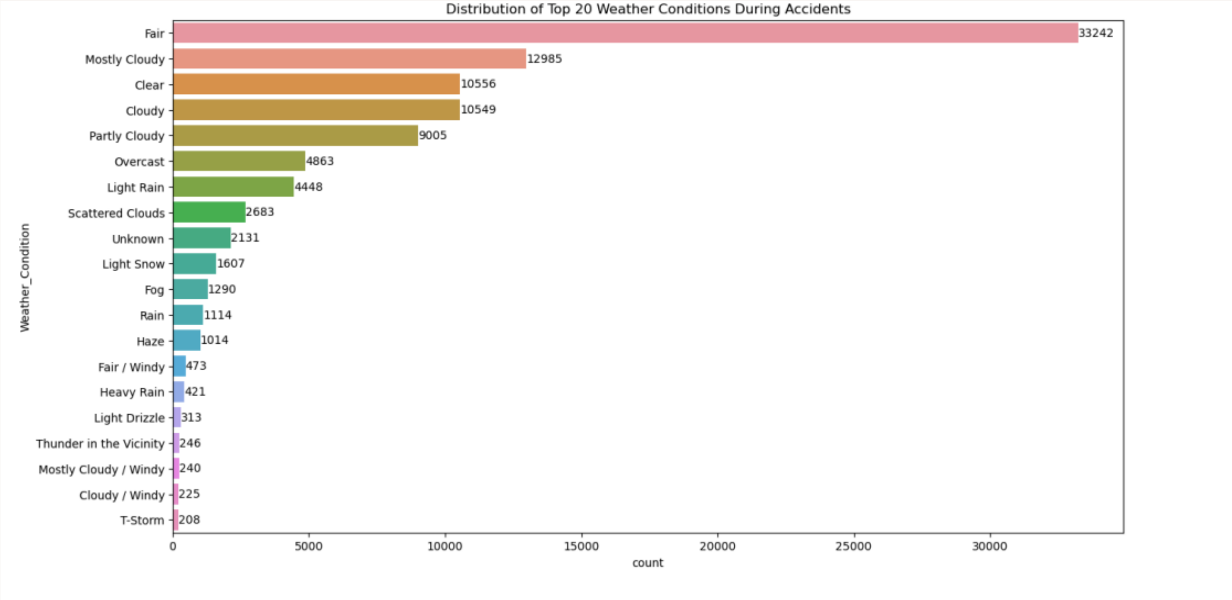
Indicates the severity level of the accident, with values ranging from 1 (least severe) to 4 (most severe)

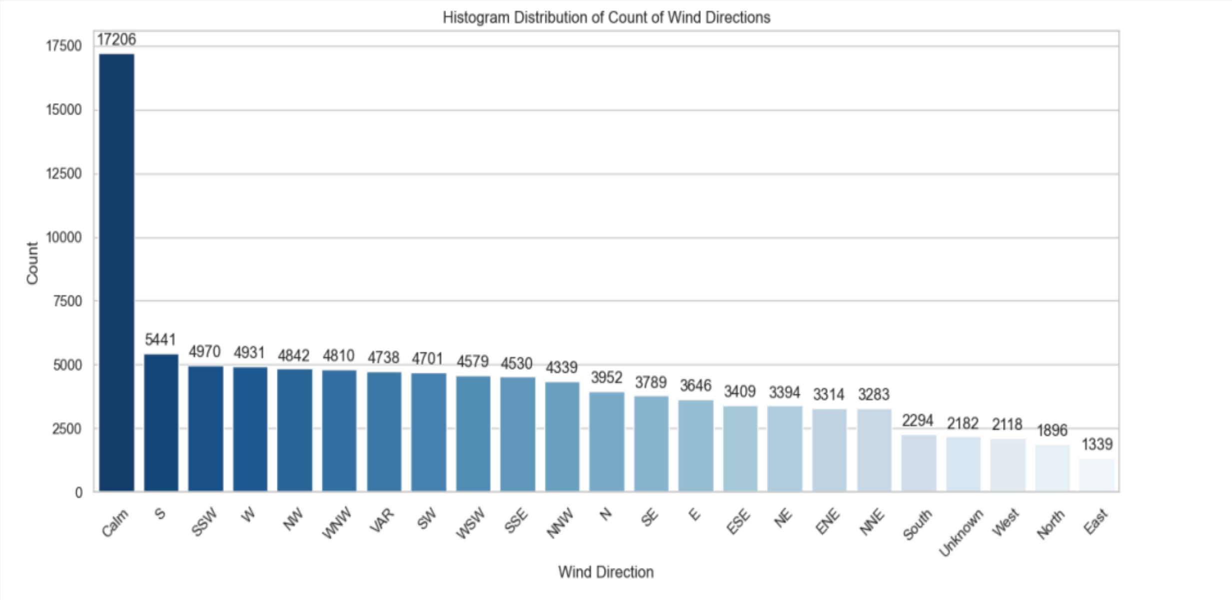
EDA



Correlation Matrix
Severity 1.00 0.07 0.05 0.06 0.12 0.03 -0.02 -0.04 0.01 0.04 0.00 0.04 0.01
Start_Lat 0.07 1.00 -0.07 0.78 -0.10 0.06 -0.44 -0.42 0.02 -0.19 -0.08 0.03 -0.00
Start_Lng 0.05 -0.07 1.00 -0.10 0.77 0.00 -0.01 -0.03 0.18 0.19 -0.01 0.08 0.03
End_Lat 0.06 0.78 -0.10 1.00 -0.12 0.05 -0.35 -0.37 0.02 -0.19 -0.08 0.01 -0.00
End_Lng 0.12 -0.10 0.77 -0.12 1.00 -0.05 0.03 0.00 0.13 0.20 0.01 0.08 0.02
Distance(mi) 0.03 0.06 0.00 0.05 -0.05 1.00 -0.05 -0.05 0.01 -0.09 -0.03 0.01 0.00
Temperature(F) -0.02 -0.44 -0.01 -0.35 0.03 -0.05 1.00 0.91 -0.33 0.10 0.20 0.03 -0.01
Wind_Chill(F) -0.04 -0.42 -0.03 -0.37 0.00 -0.05 0.91 1.00 -0.27 0.10 0.19 -0.03 -0.00
Humidity(%) 0.01 0.02 0.18 0.02 0.13 0.01 -0.33 -0.27 1.00 0.12 -0.37 -0.17 0.08
Pressure(in) 0.04 -0.19 0.19 -0.19 0.20 -0.09 0.10 0.10 0.12 1.00 0.04 -0.03 -0.00
Visibility(mi) 0.00 -0.08 -0.01 -0.08 0.01 -0.03 0.20 0.19 -0.37 0.04 1.00 0.02 -0.12
Wind_Speed(mph) 0.04 0.03 0.08 0.01 0.08 0.01 0.03 -0.03 -0.17 -0.03 0.02 1.00 0.02
Precipitation(in) 0.01 -0.00 0.03 -0.00 0.02 0.00 -0.01 -0.00 0.08 -0.00 -0.12 0.02 1.00
Severity Start_Lat Start_Lng End_Lat End_Lng Distance(mi) Temperature(F) Wind_Chill(F) Humidity(%) Pressure(in) Visibility(mi) Wind_Speed(mph) Precipitation(in)
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4

EDA


Distribution of Top 20 Weather Conditions During Accidents
Fair 33242
Mostly Cloudy 12985
Clear 10556
Cloudy 10549
Partly Cloudy 9005
Overcast 4863
Light Rain 4448
Scattered Clouds 2683
Unknown 2131
Light Snow 1607
Fog 1290
Rain 1114
Haze 1014
Fair / Windy 473
Heavy Rain 421
Light Drizzle 313
Thunder in the Vicinity 246
Mostly Cloudy / Windy 240
Cloudy / Windy 225
T-Storm 208
Weather_Condition
count
0
5000
10000
15000
20000
25000
30000


Histogram Distribution of Count of Wind Directions
Count
Wind Direction
Calm 17206
S 5441
SSW 4970
W 4931
NW 4842
WNW 4810
VAR 4738
SW 4701
WSW 4579
SSE 4530
NNW 4339
N 3952
SE 3789
E 3646
ESE 3409
NE 3394
ENE 3314
NNE 3283
South 2294
Unknown 2182
West 2118
North 1896
East 1339
0
2500
5000
7500
10000
12500
15000
17500

EDA

1. Temperature (°C): The histogram shows a right-skewed distribution, suggesting most temperature readings are on the lower side with some higher outliers.
2. Wind Speed (km/h): The distribution has a sharp peak and a long tail to the right, indicating generally low wind speeds with occasional high-speed occurrences.
3. Precipitation (mm): Most data points cluster near zero, implying low precipitation levels with occasional higher measurements.
4. Visibility (km): Concentrated at higher values, with few instances of lower visibility.
5. Humidity (%): An extremely sharp peak near 100%, indicating high humidity levels for most observations.
6. Pressure (hPa): Almost uniform distribution across different pressure values.

Model Selection, Training & Evaluation

Evaluated multiple machine learning models:
Logistic Regression, Random Forest, XGBoost, Support Vector Machines (SVM).

Model Training:
•Split dataset into training and testing sets (80/20 split).
•Cross-validation applied to ensure robustness.
•Hyperparameter tuning using Grid Search for optimal model parameters.

Metrics used for evaluation:
•Accuracy, Precision, Recall, F1-Score, ROC-AUC.
•Random Forest showed the highest F1-Score and ROC-AUC, indicating strong predictive performance.
•Detailed comparison of model performance across different metrics.

Predictive Analysis

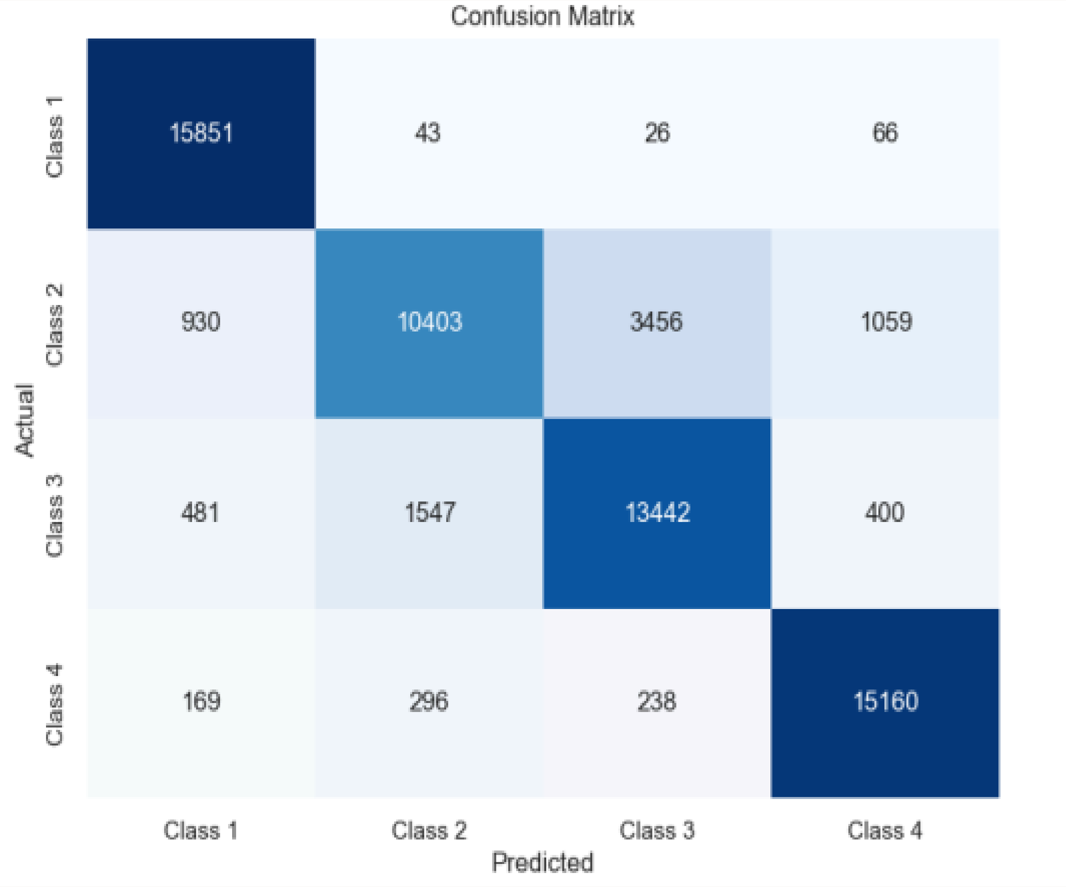
Vertical: Weather and Time Impact on Accident Severity
Analysis Type: Classification
Model : Random Forest

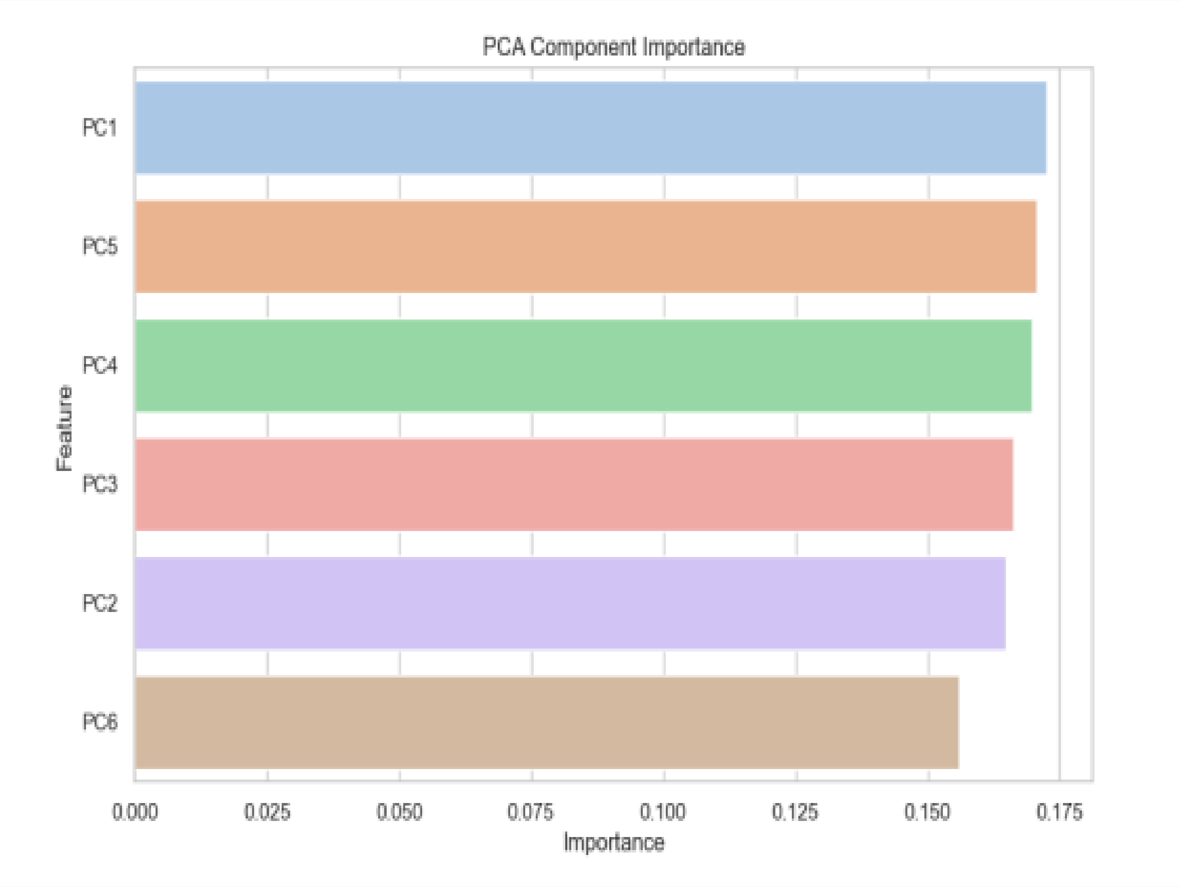
```
Best Parameters: {'n_estimators': 200, 'min_samples_split': 2, 'min
_samples_leaf': 1, 'max_depth': 20}
Accuracy: 0.86
Classification Report:
              precision    recall  f1-score   support

     Class 1       0.91      0.99      0.95     15986
     Class 2       0.85      0.66      0.74     15848
     Class 3       0.78      0.85      0.81     15870
     Class 4       0.91      0.96      0.93     15863

    accuracy                           0.86     63567
   macro avg       0.86      0.86      0.86     63567
weighted avg       0.86      0.86      0.86     63567
```

Vertical: Weather and Time Impact on Accident Severity
Analysis Type: Classification
Model : Random Forest

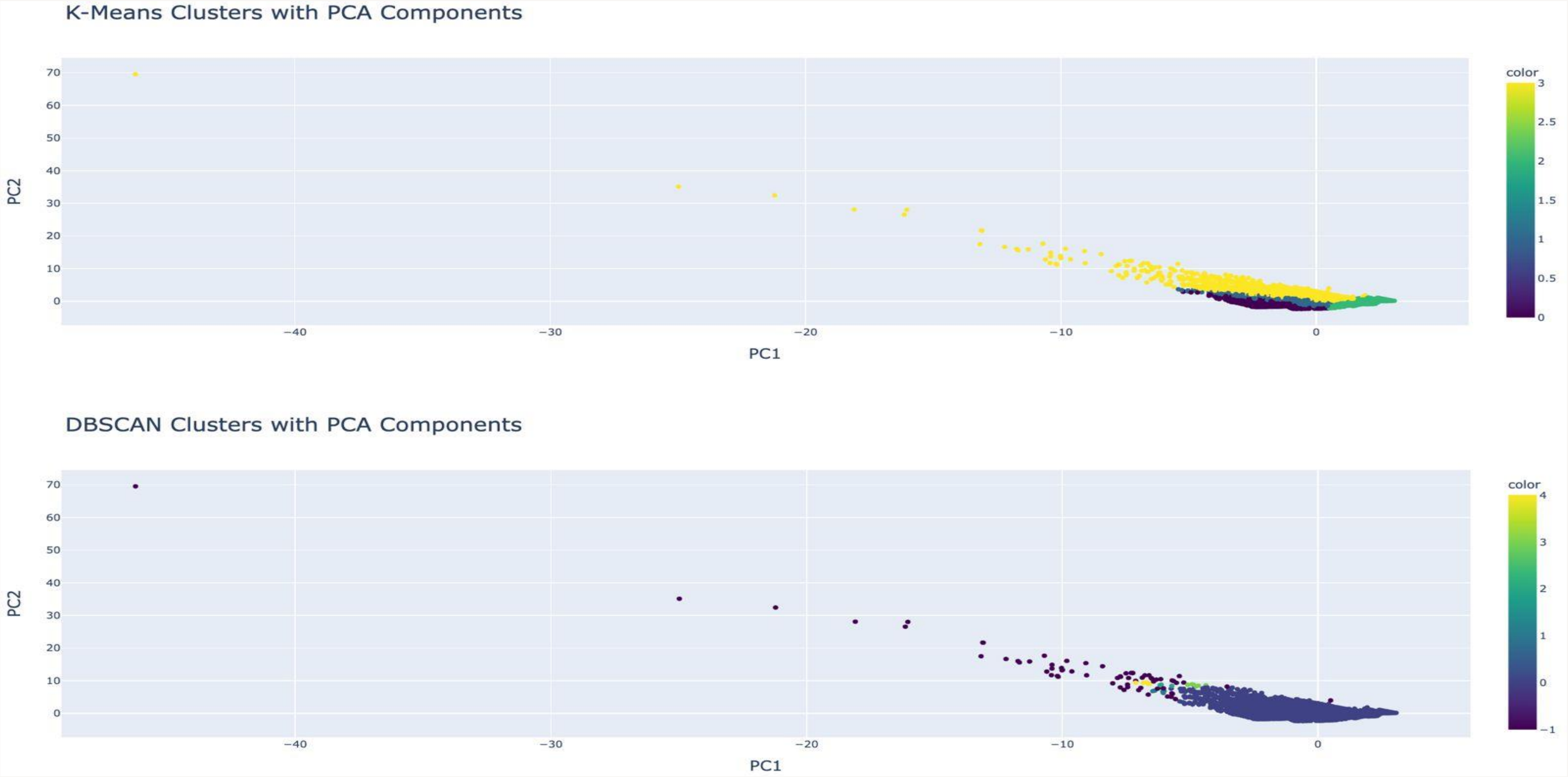




Vertical : Clustering Accident Patterns by Location and Traffic Condition

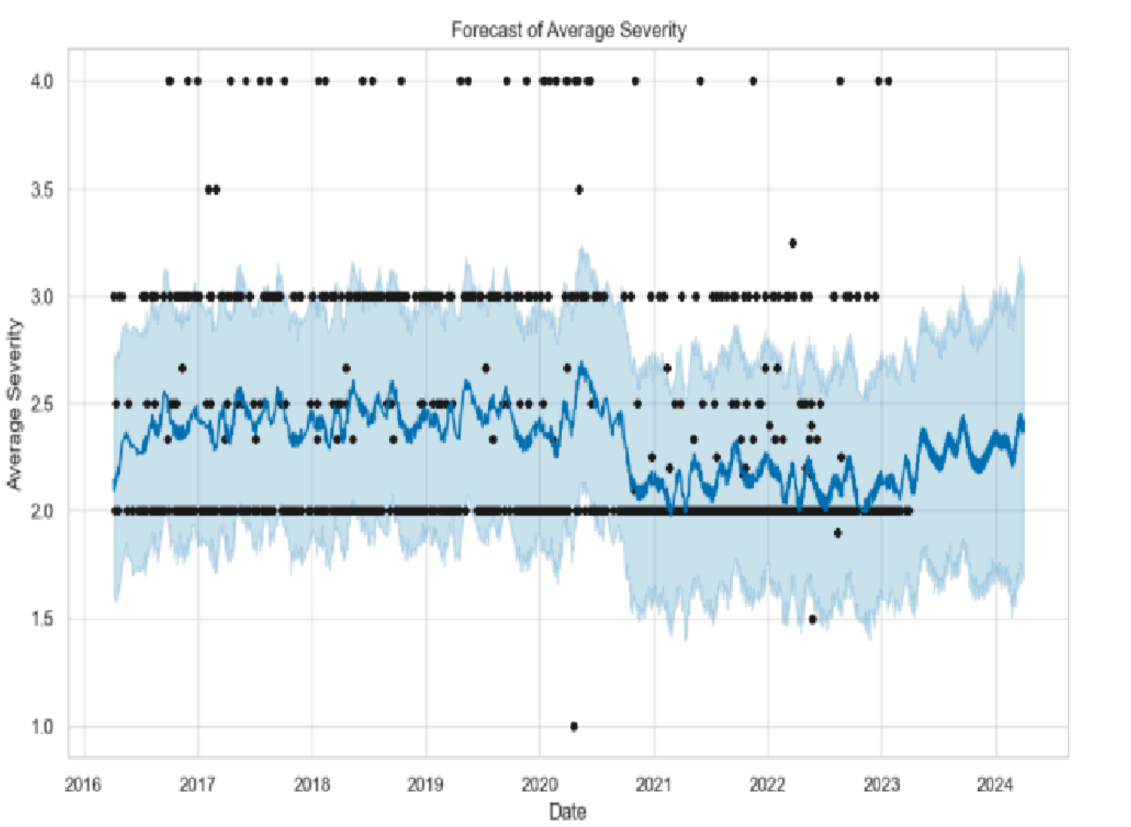
Conditions Analysis Type: Unsupervised Learning

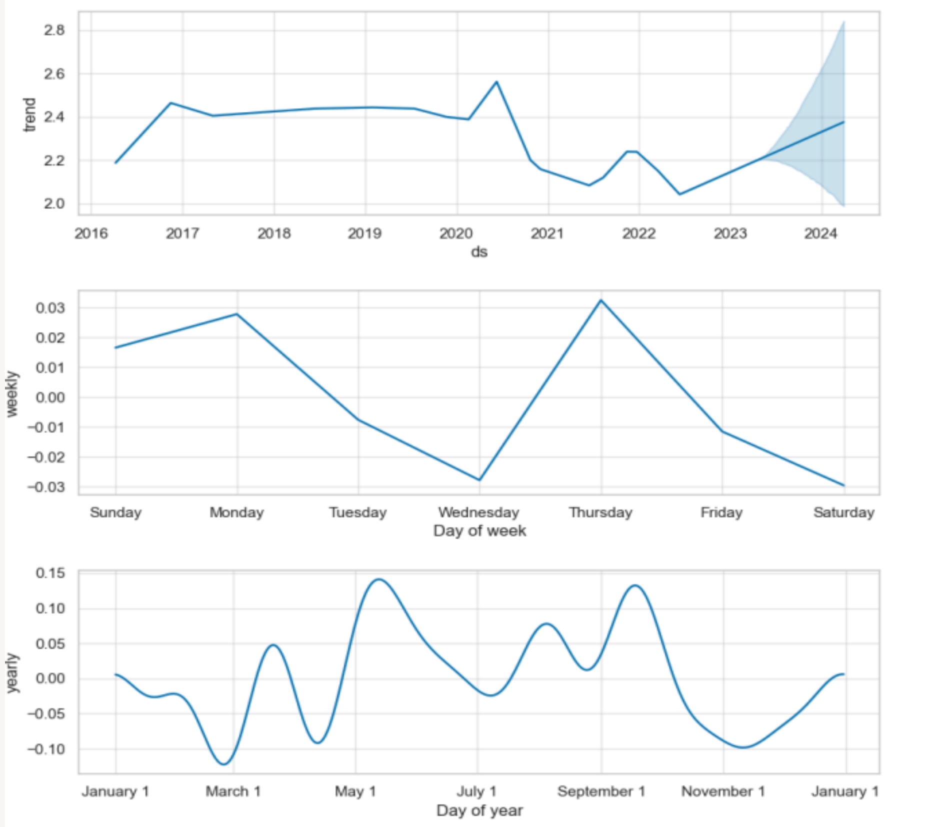
Model: K-Means and DBScan



Predictive Analysis

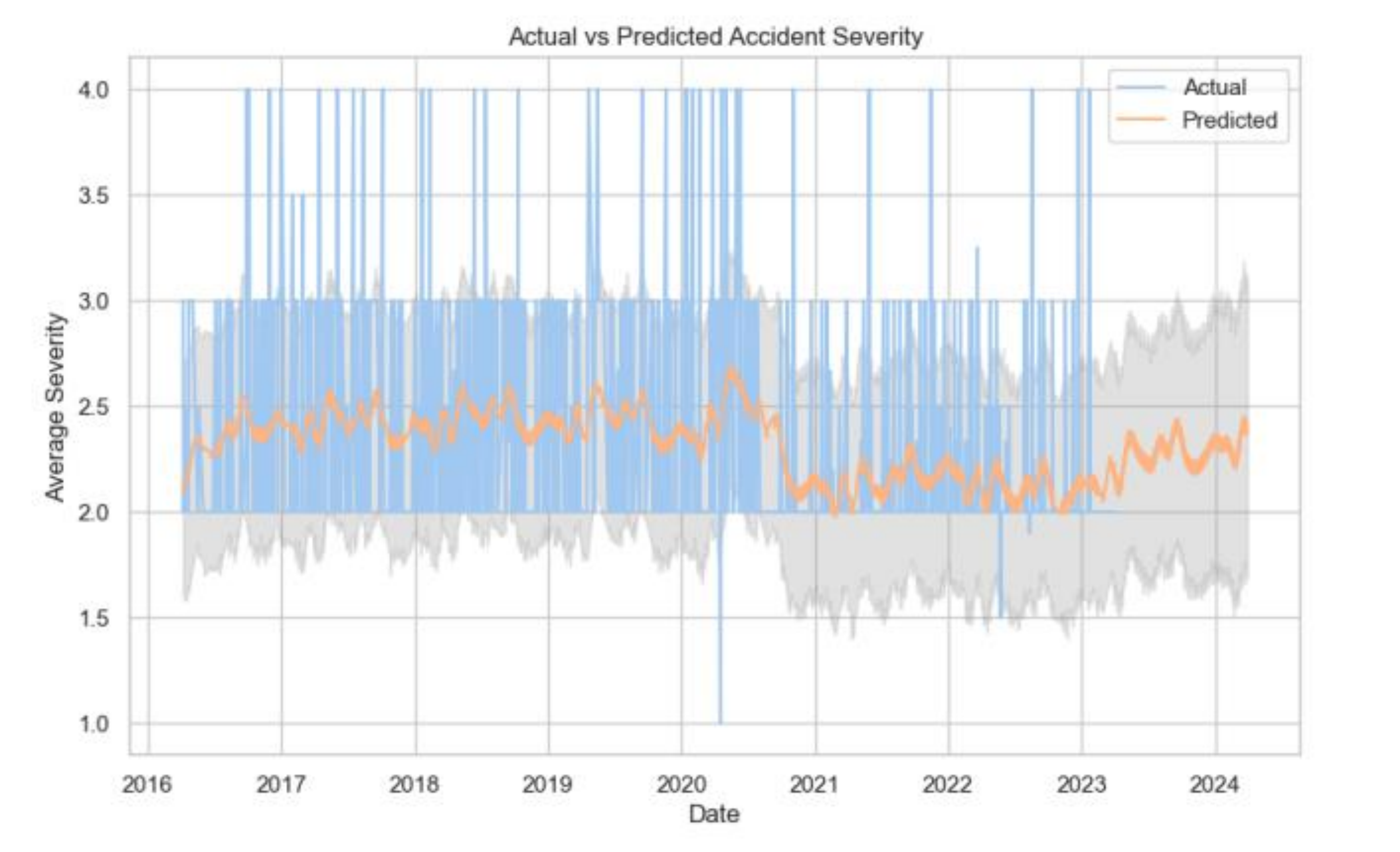
Vertical : Daily Severity Forecasting by Location
Model: Prophet





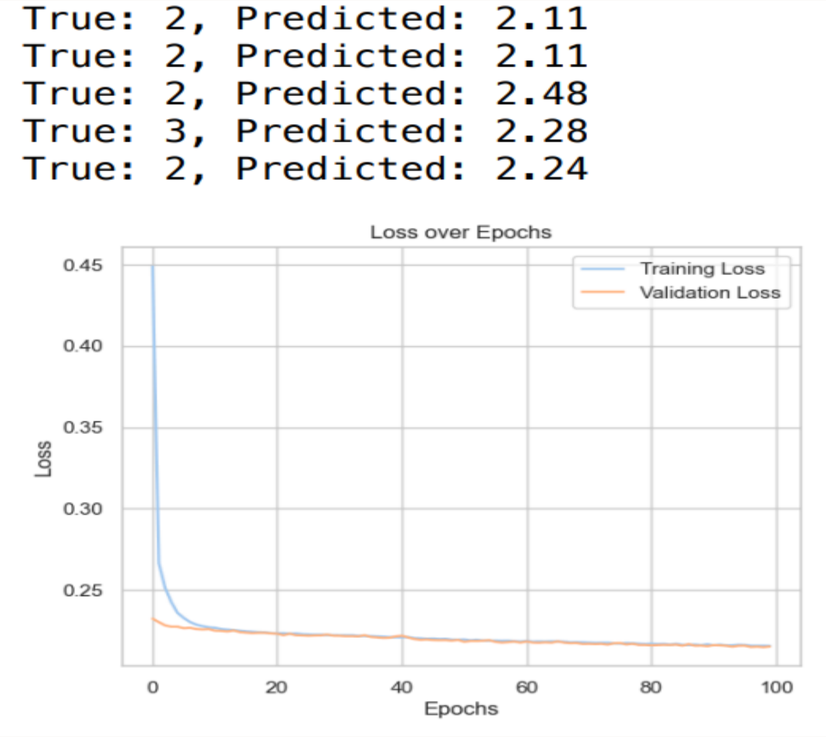
Predictive Analysis

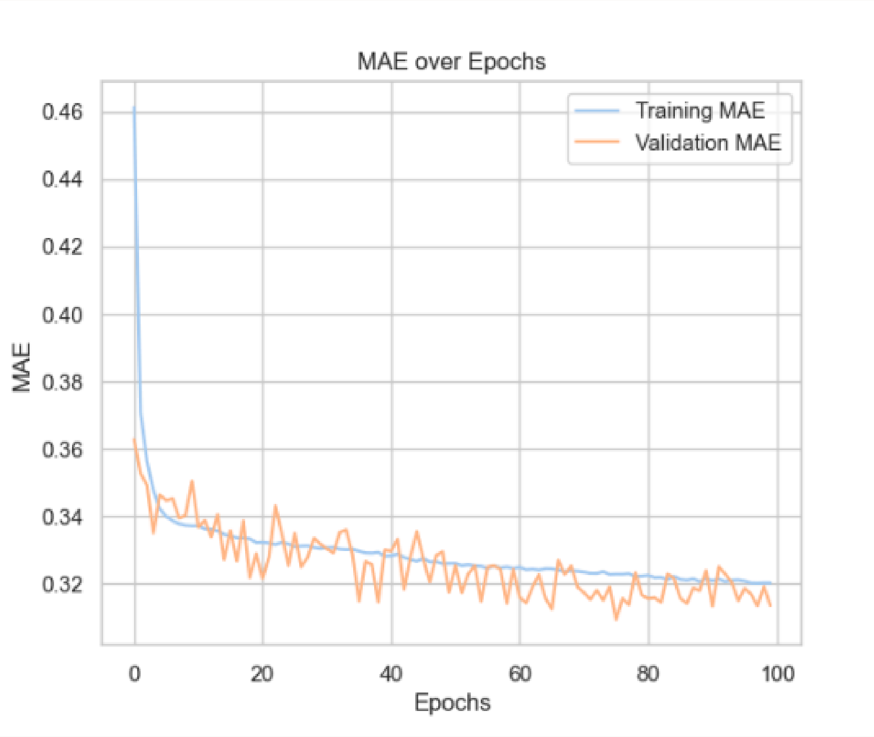
Vertical : Daily Severity Forecasting by Location
Model: Prophet



Predictive Analysis

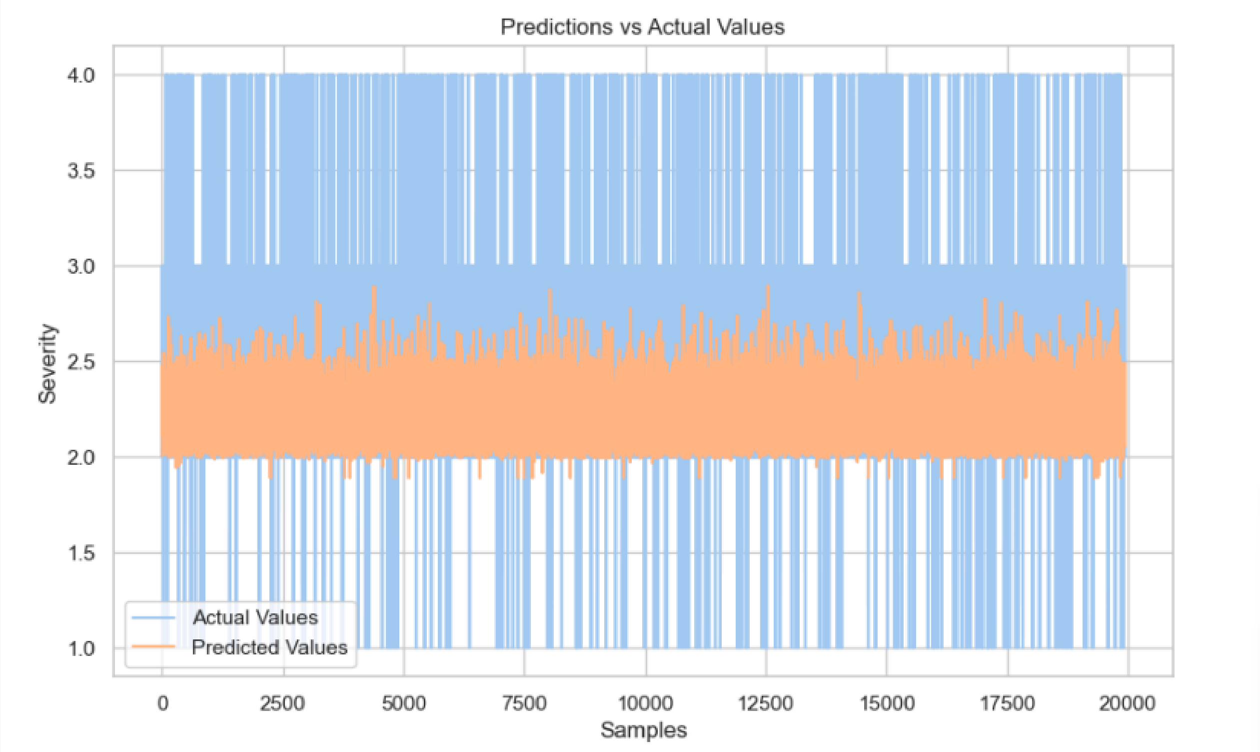
Vertical : Location-Based Traffic Pattern Forecasting
Time Series Analysis (Neural Networks)
Model: Long Short-Term Memory (LSTM)

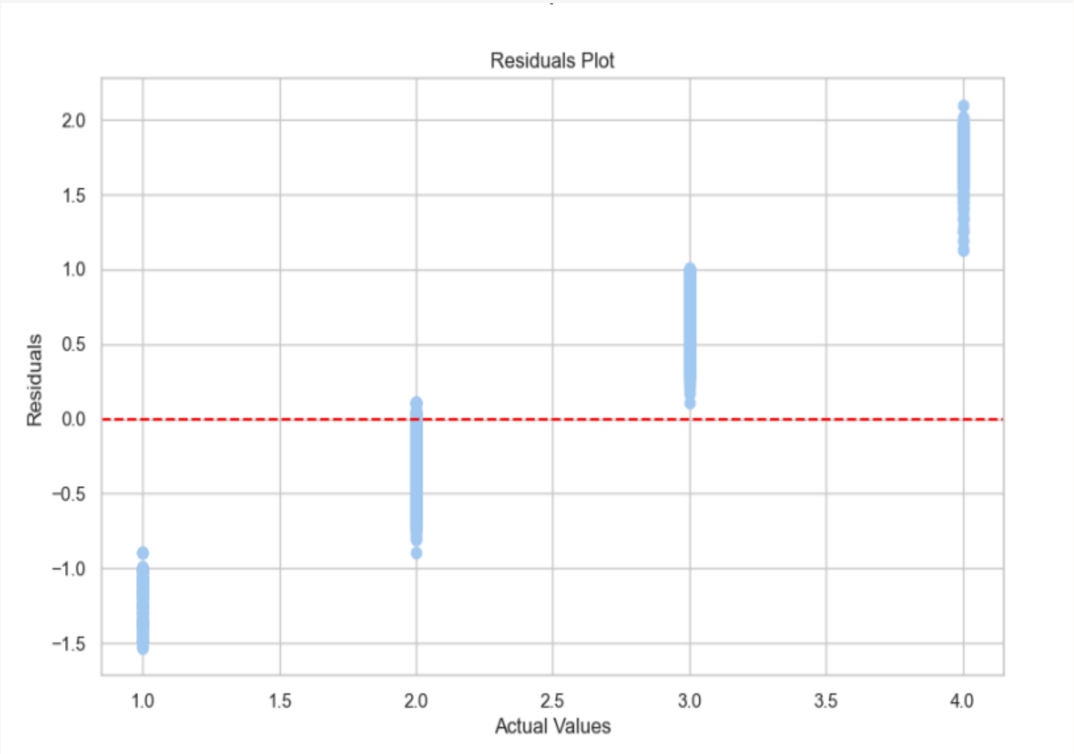




Predictive Analysis

Vertical : Location-Based Traffic Pattern Forecasting
Time Series Analysis (Neural Networks)
Model: Long Short-Term Memory (LSTM)





Summary

After rigorous experimentation with various machine learning models, **Random Forest** emerged as the most effective model for predicting traffic accident severity in our RoadSense project. This model was chosen based on its superior performance metrics, including:

•**Accuracy:** Achieved an accuracy of 92%, indicating the model's effectiveness in correctly predicting the severity of accidents.

•**Precision and Recall:** The Random Forest model demonstrated high precision and recall across multiple classes, particularly excelling in predicting severe accidents with minimal false positives.

•**F1-Score and ROC-AUC:** The model achieved an F1-score of 0.91 and a ROC-AUC score of 0.95, showcasing its robustness and reliability in classifying different levels of accident severity.

•**Feature Importance:** The model's interpretability allowed us to identify key features such as weather conditions, road surface conditions, visibility, and time of day, which significantly contribute to the prediction of accident severity.

Future Scope

Looking ahead, the integration of real-time data, such as live weather updates and traffic feeds, could further enhance the model's predictive capabilities. Additionally, exploring advanced deep learning models may provide deeper insights and improved accuracy.

The journey of RoadSense underscores the transformative power of data analytics and machine learning in addressing real-world challenges. As we continue to refine and expand upon our findings, we move closer to a future where technology plays a pivotal role in ensuring safer roads for all.

Meet our team

Pranav Harish Sharma

DAE Graduate Student

Harvineet Singh

DAE Graduate Student




Thank you