US Accidents

Big Data Computing (2022-2023) Project

Elena Jiang 1846716

Project Goal

Studying the *US Accidents dataset* is important to identify patterns and factors contributing to accidents, enabling the development of effective safety measures and policies, ultimately saving lives and reducing injury rates.

My project focuses on the **classification** of accidents, allowing users to input information about an accident. The system will then provide information on the severity of the accident.

Roadmap

Dataset

Data processing

Cluster

Classification

Demo

US Accidents (2016 -2023)

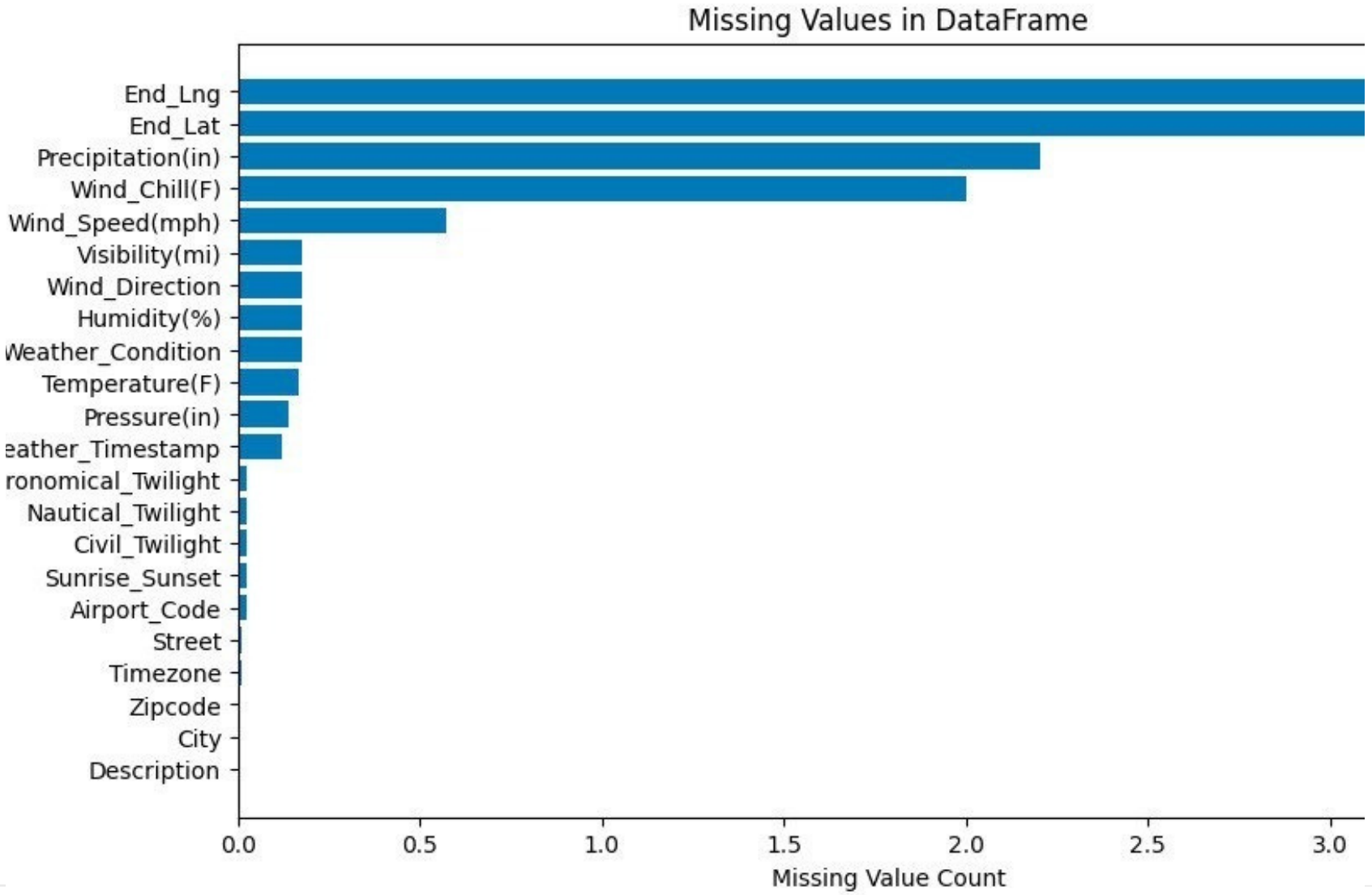
The dataset already exists on **Kaggle**.
It is a *countrywide car accident* dataset that covers **49 states of the USA**. The accident data were collected from **February 2016 to March 2023**, using multiple APIs that provide streaming traffic incident (or event) data. The dataset currently contains approximately **7.7 million**accident records.

Accident's information

(46 features)

The dataset has a total of 46 columns that provide detailed information about each accident record.

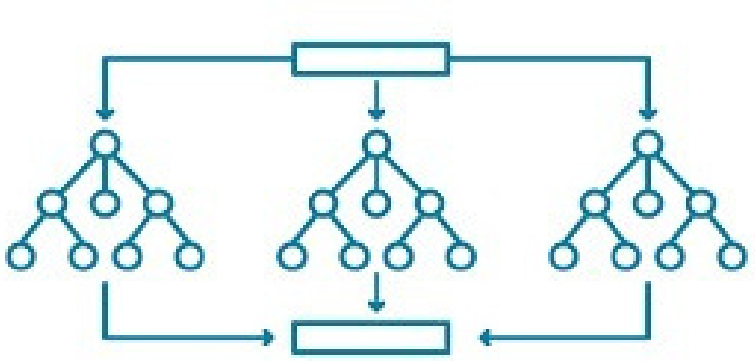
- **Severity**: Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic
- **Start_Lat, Start_Lng, End_Lat, End_Lng:** Shows in GPS coordinate of the start and end point.
- **Street, City, County, State, Zipcode**: give more information about the position
- **Weather-related** variables such as: Temperature (F), Wind Chill (F), Humidity (%), Precipitation (in), Visibility (mi), and more.
- **Traffic-related** features, all of which are boolean indicators indicating their presence in or near the location: Roundabout, Station, Stop, Traffic Calming, Traffic Signal, and others.
- features that determine the **time of day**, including Civil Twilight, Nautical Twilight, and Astronomical Twilight.



Data Processing: Missing Value

Data Processing:

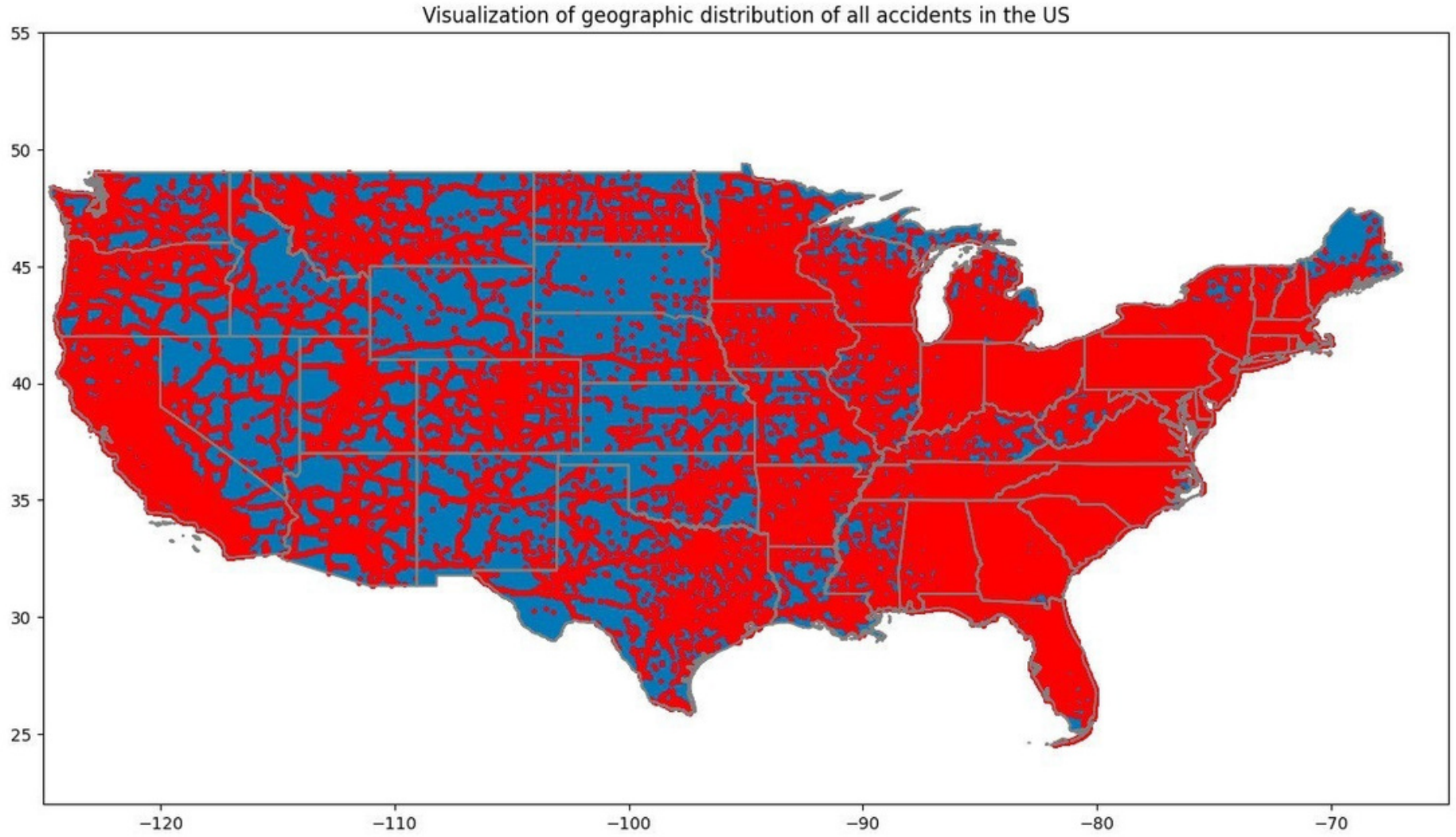
Feature Importance



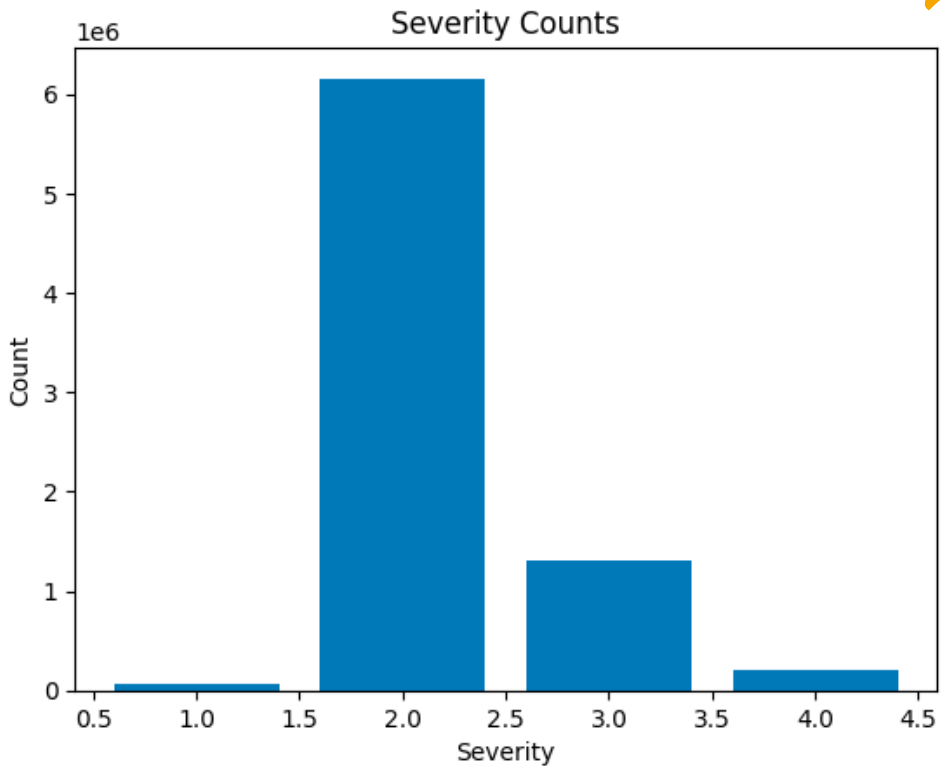
Feature importance is a technique used to determine the impact or significance of different features on a machine learning model's predictions.

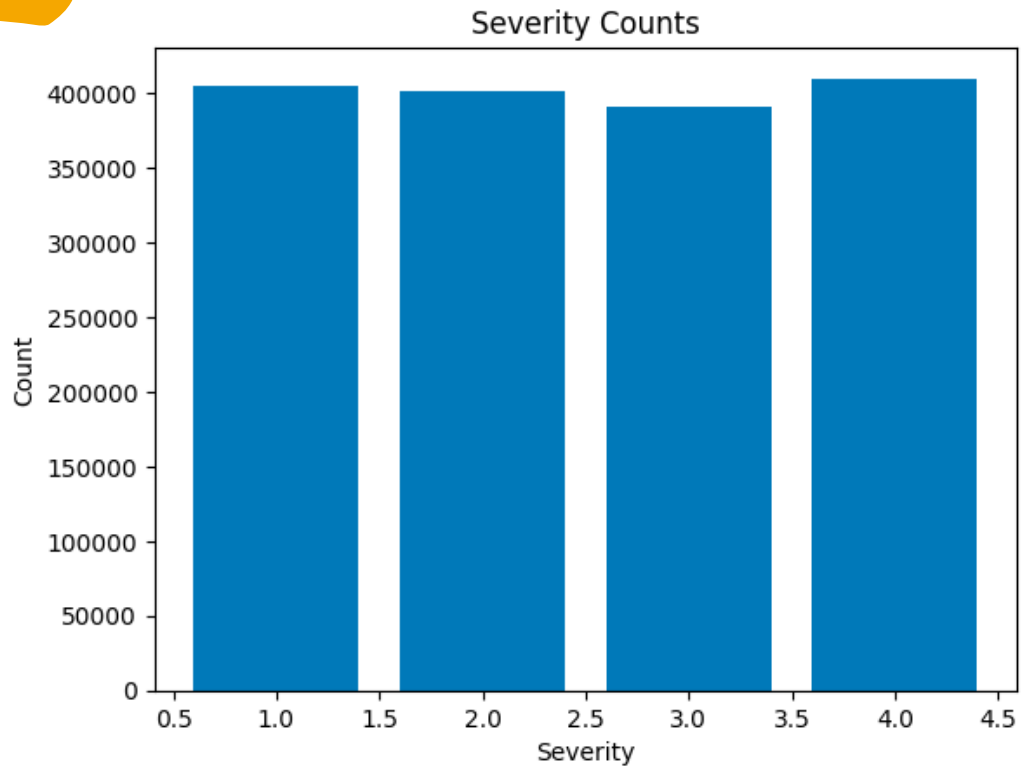
I use a **Random Forest classifier** to calculate and display the importance of each feature in predicting the *'Severity*' label.

Data Processing: Accidents Distribution Graph



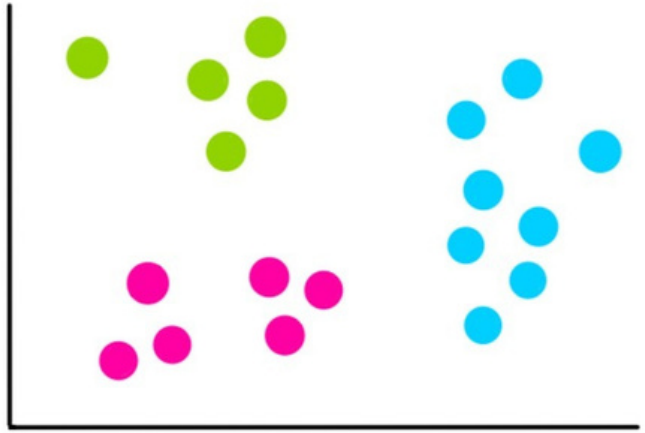
RESAMPLING





Handle Imbalanced Dataset

Clustering: K-means



I use **K-means algorithm** which cluster accidents based on different features for studying and using them in accident classification.

For each feature, I perform clustering using both **Euclidean** and **cosine** distance measures. The data has been clustered **from 5 up to 50 clusters** for each measure, with each method involving a maximum of **20 iterations.**

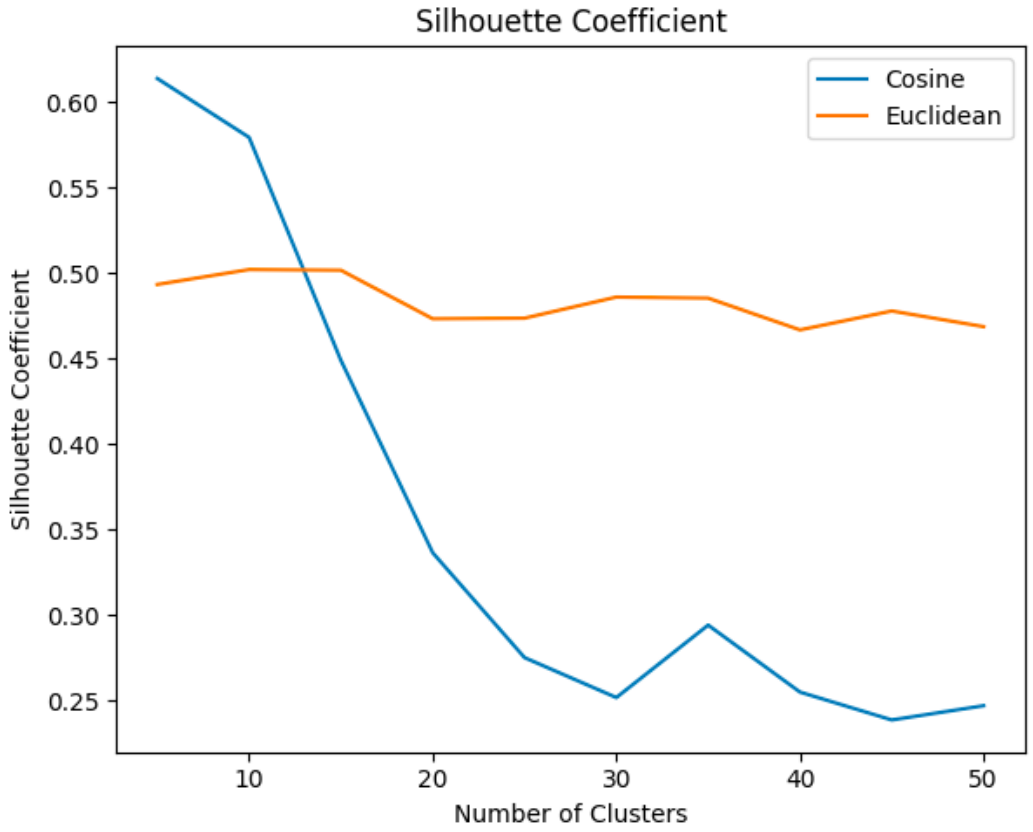
Clustering: K-means evaluation

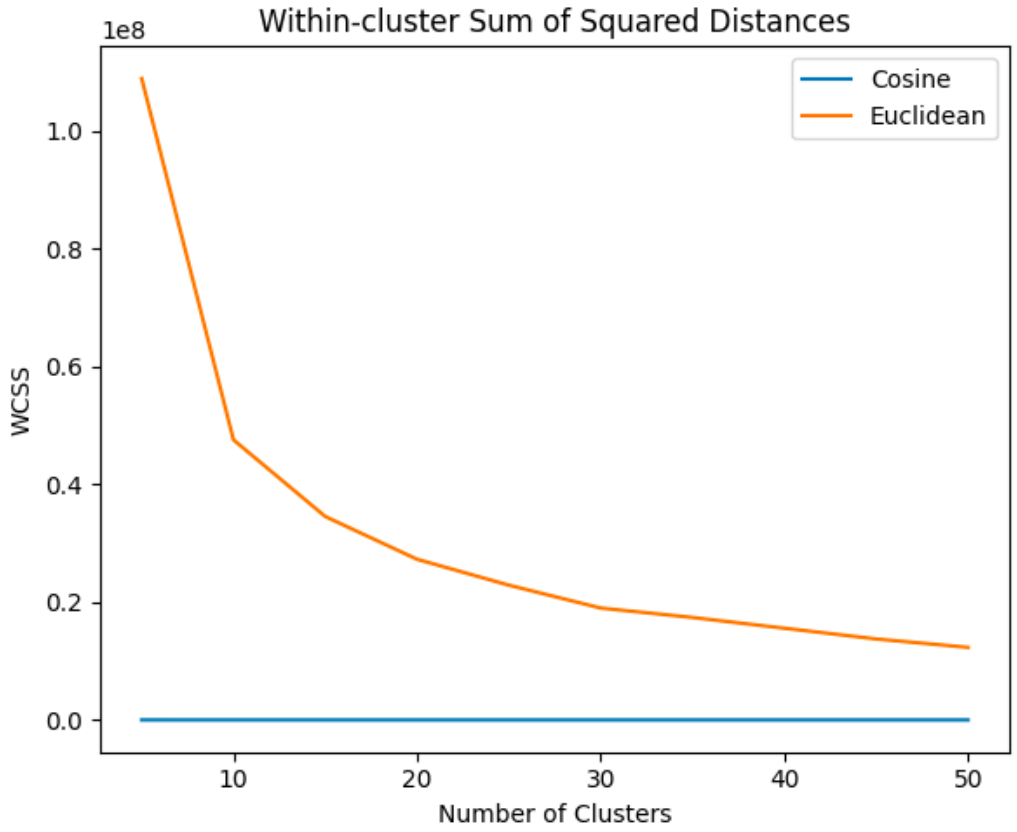
The **Silhouette score** is a valuable metric for assessing the quality of clusters. A higher Silhouette score typically indicates better-defined and well-separated clusters.

The **Within-cluster Sum of Squared Distances (WSSD)** provides a measure of how tightly the data points within each cluster are grouped. A lower WSSD indicates more compact and well-defined clusters.

Cluster of weather-related features

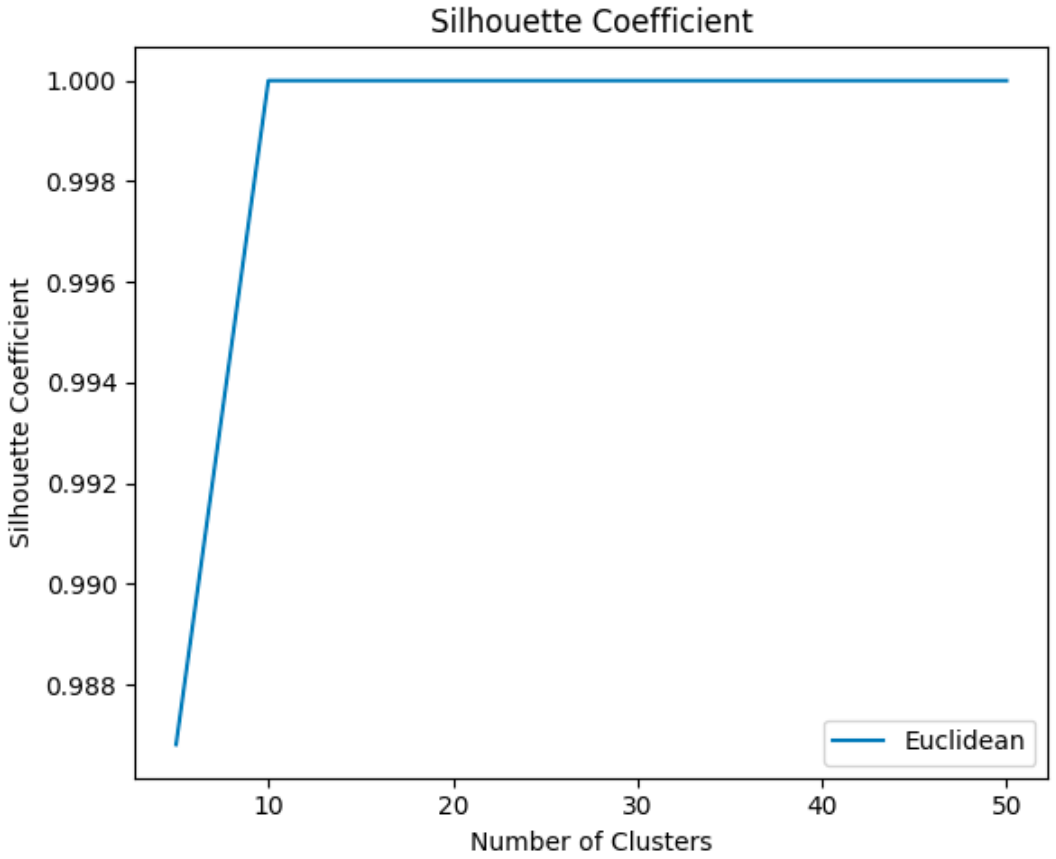
I utilize the features **Wind_Chill (F), Wind_Speed (mph), Wind_Direction, Precipitation (in), Weather_Condition** for weather clustering analysis

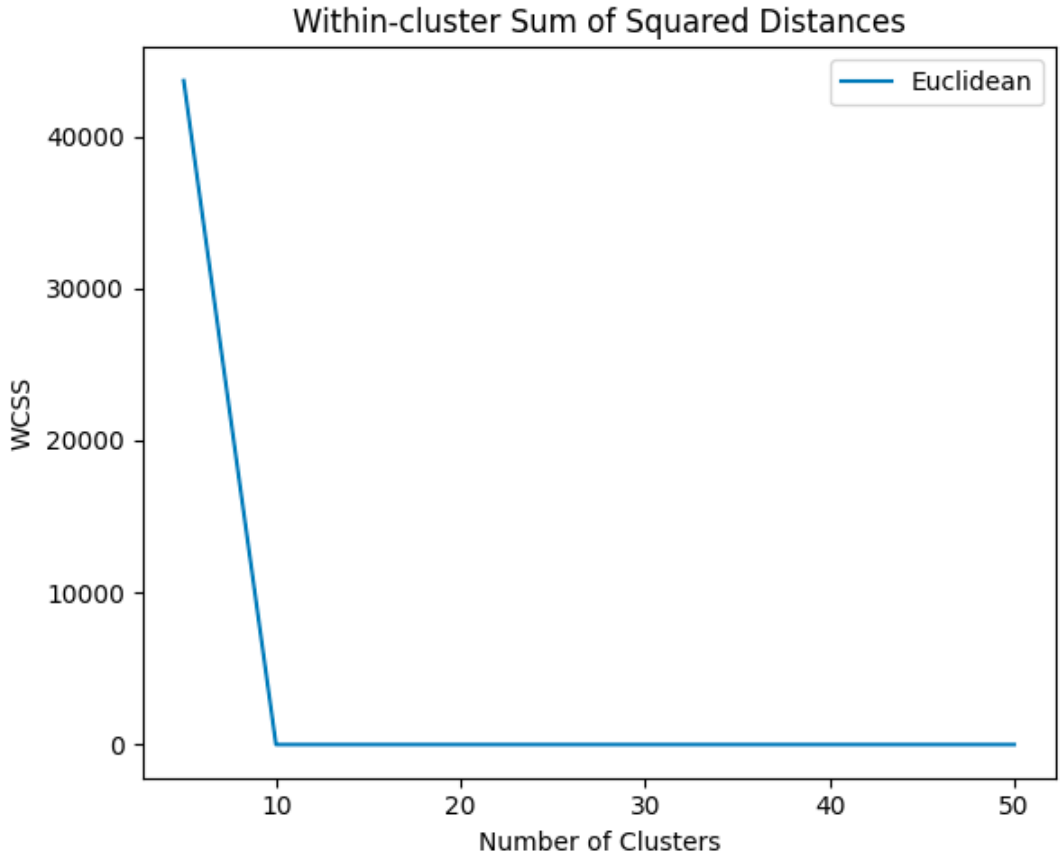




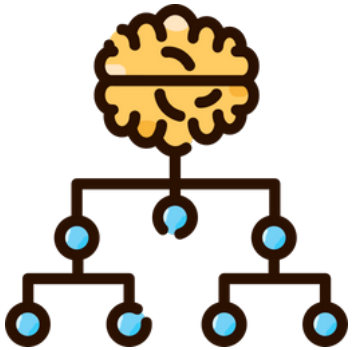
Cluster of Traffic-related features

I utilize the features **Crossing, Stop** and **Traffic_signals** for traffic clustering analysis. They are boolean values indicating their presence in or near the accident location.





Classification



For classification, I utilize the results from previous clustering and the accident description. Now, I am employing various machine learning models to train the system, enabling it to predict the severity of accidents.

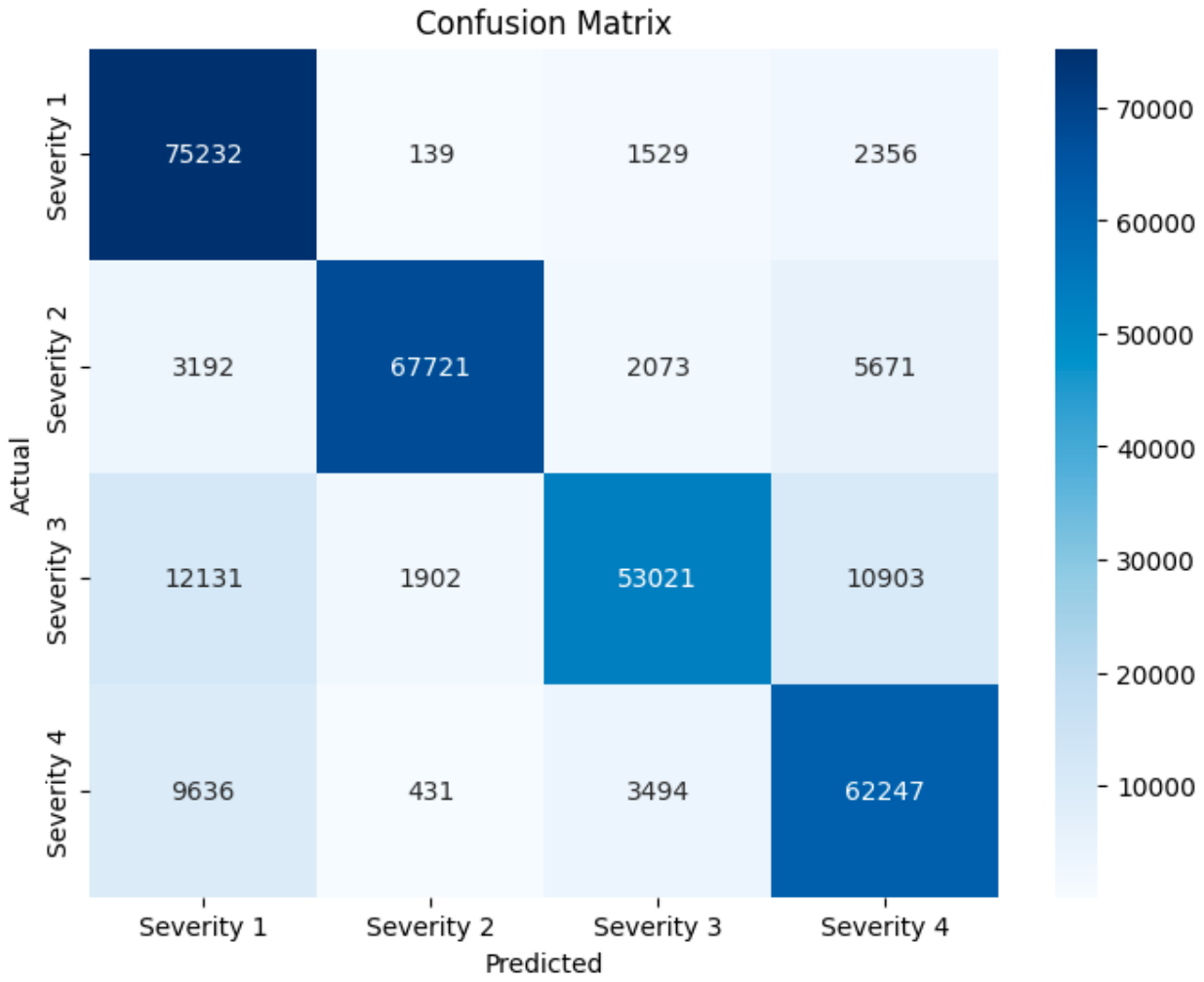
Two types of classification

Random Forest

Multinomial logistic regression

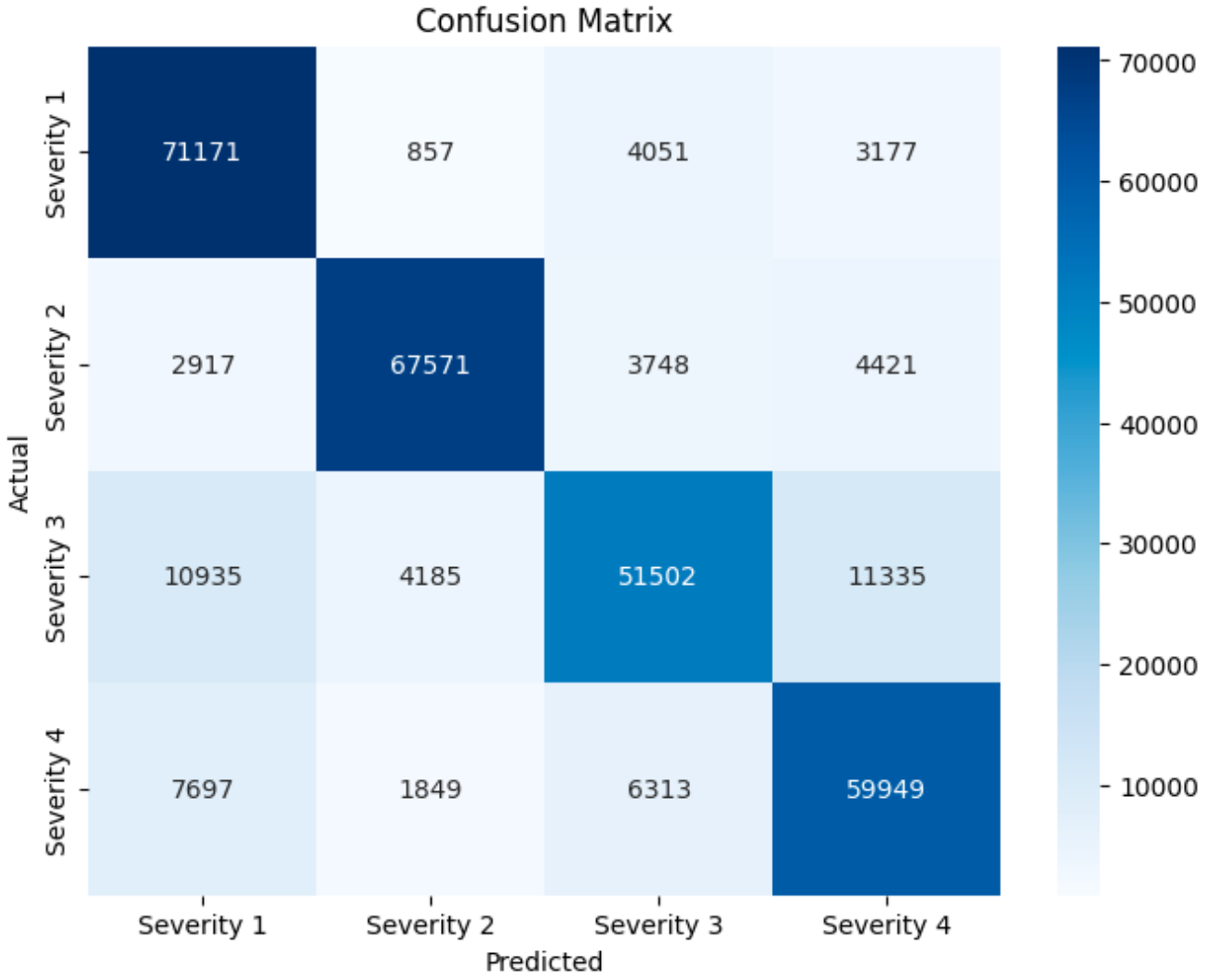
Random Forest

Using the Random Forest model, I achieved an **accuracy** of **82.85%**. The **weighted precision** is **84.15%** **the weighted recall** is **82.85%,** and **the weighted F1-score** is **82.78%.**



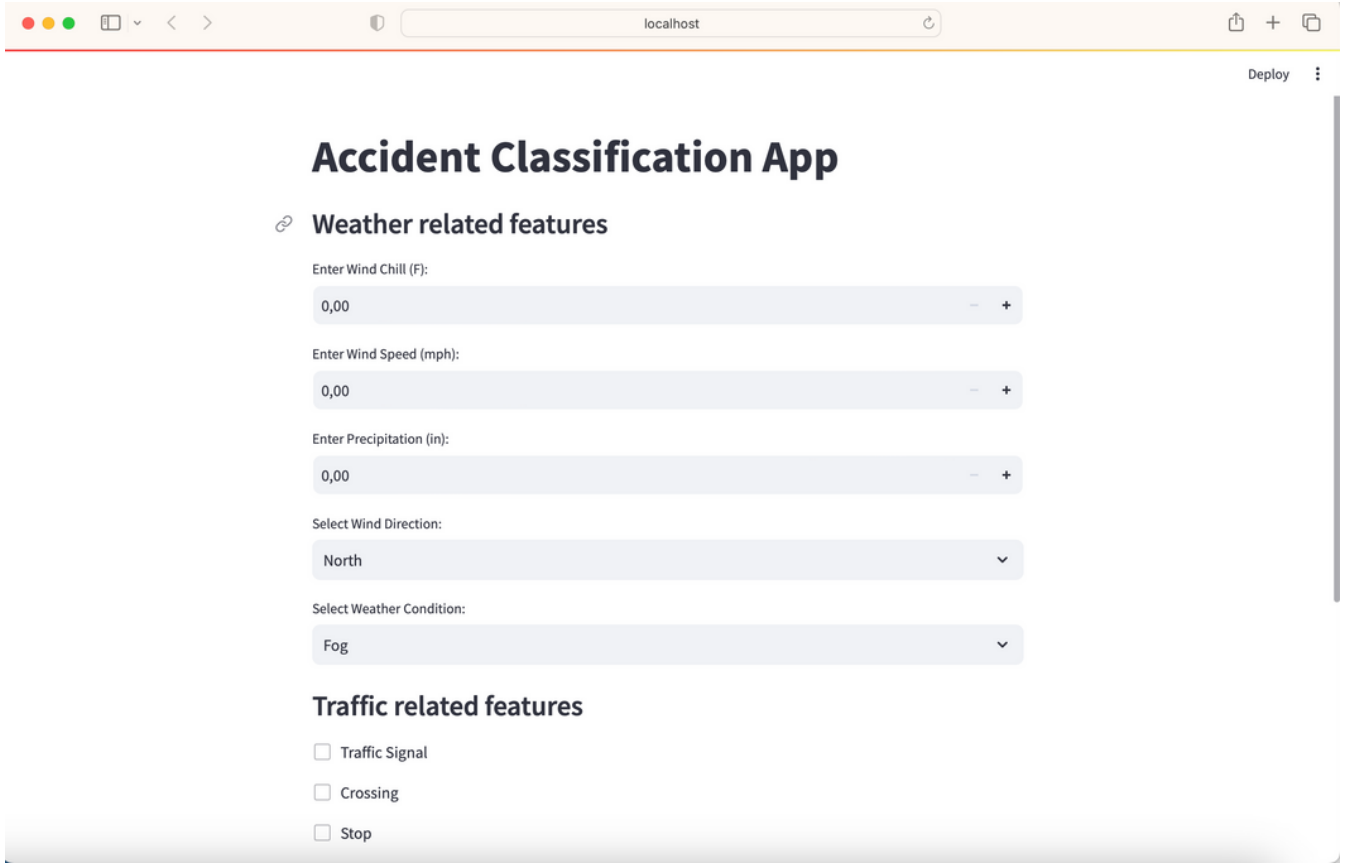
Multinomial Logistic Regression

Using the Random Forest model, I achieved an **accuracy** of **80.27%**. The **weighted precision** is **80.54%** **the weighted recall** is **80.27%,** and **the weighted F1-score** is **80.12%.**



Web App

Use Streamlit to build a web app where users can input various features to classify the severity of the accident.



THANKS

FOR WATCHING