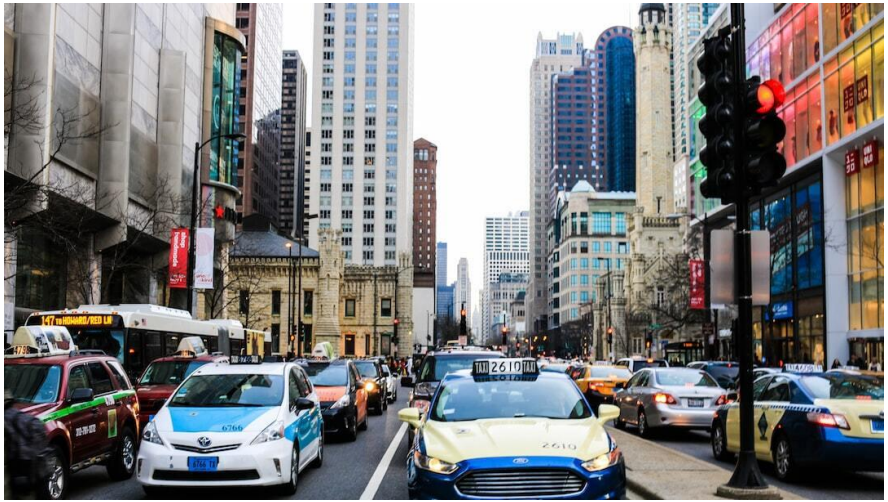
Car Accident Contributory Cause Classification in Chicago

- **Public Safety Issue:** Car accidents are a major public safety concern, necessitating effective prevention strategies.
- **Project Aim:** Build a classifier to predict the primary contributory cause of car accidents in Chicago using data on people and vehicles.
- **Stakeholder Benefits:** Insights can help the Vehicle Safety Board and City of Chicago identify key accident factors and implement preventive measures.

Business Understanding

- **Primary Goal:** Build a classifier predicting the primary contributory cause of car accidents in Chicago.
- **Target Audience:** Vehicle Safety Board and City of Chicago to reduce traffic accidents and identify patterns.
- **Challenges:** Data complexity, class imbalance and feature engineering requiring careful analysis.

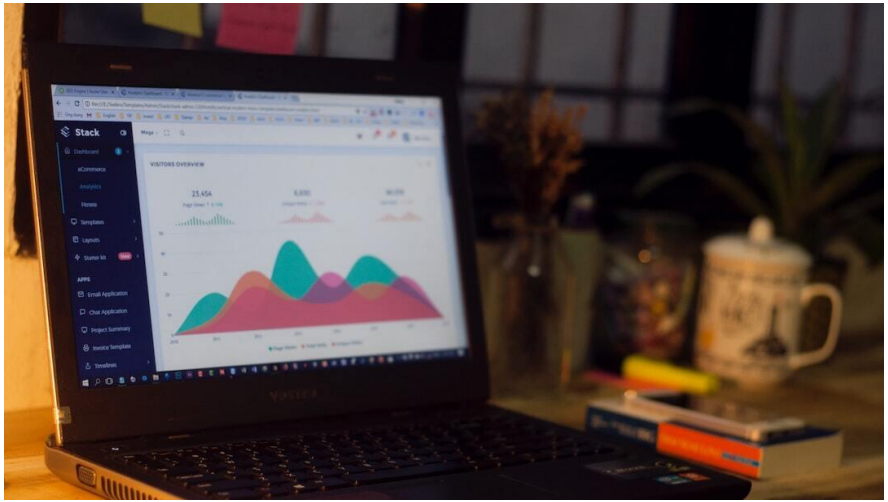


Proposed Solution and Conclusion

- **Evaluation Metrics:** Primary metric is accuracy with a target of 80%. Also considering precision, recall, and F1 score.
- **Outcome:** Reliable classifier helps identify key factors contributing to accidents.
- **Benefits:** Provides insights to Vehicle Safety Board and City of Chicago for preventive measures to enhance road safety.

Data Understanding

- **Data Sources:** People dataset and Vehicle dataset from Chicago Data Portal.
- **Initial Exploration:** Loading and examining the first few rows of both datasets.
- **Merge Datasets:** Combining People and Vehicle datasets on CRASH_RECORD_ID for comprehensive analysis.

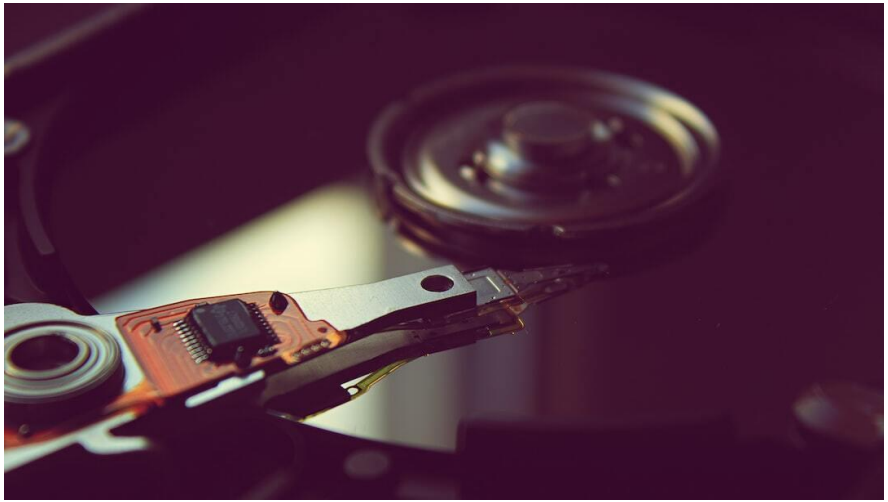


Data Preparation

- **Data Reduction:** Sampling 15% of data from each year to handle dataset size and still maintain yearly representation.
- **Choosing columns:** These were the columns used:

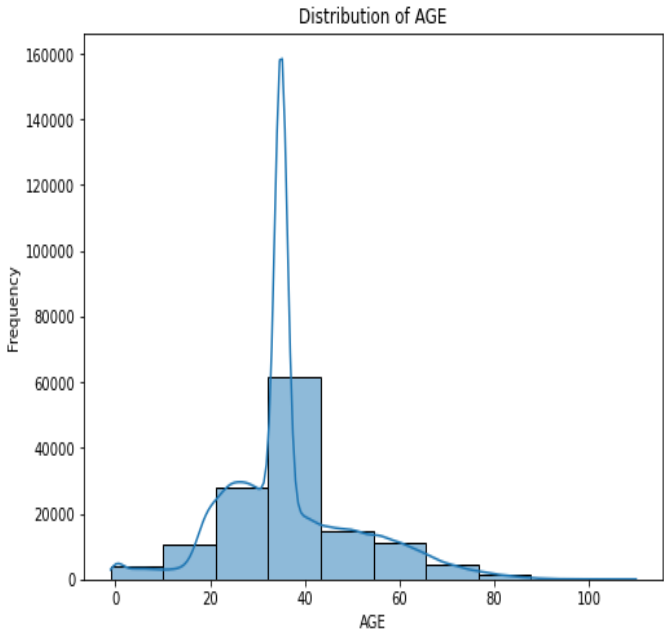
PERSON_TYPE, SEX, AGE, SAFETY_EQUIPMENT, AIRBAG_DEPLOYED, EJECTION, INJURY_CLASSIFICATION(Dependent Variable), DRIVER_ACTION, DRIVER_VISION, PHYSICAL_CONDITION, MAKE, MODEL, VEHICLE_YEAR, VEHICLE_DEFECT, VEHICLE_USE, CRASH_YEAR AND CRASH_MONTH.

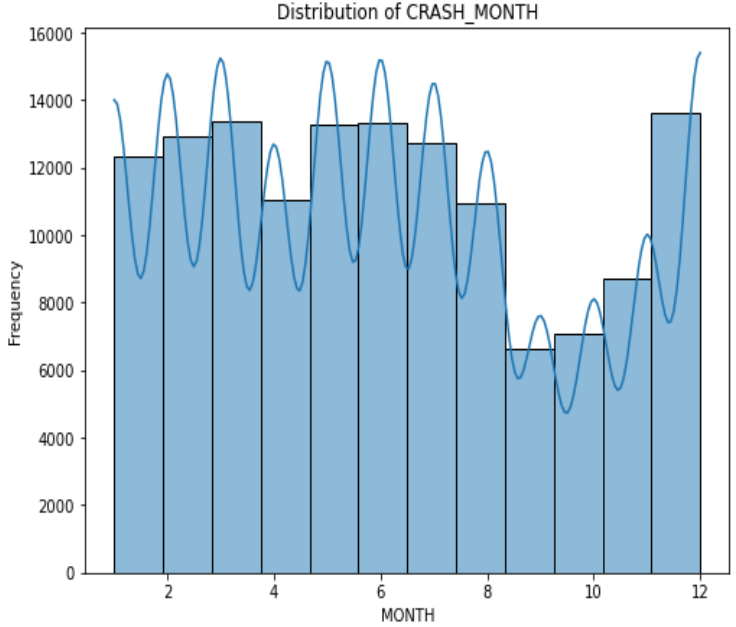
- **Final Dataset:** Resulting in a manageable dataset of 140,041 rows after cleaning and preparation.
- **Data Cleaning:** Addressing missing values and anomalies, e.g., limiting VEHICLE_YEAR between 1900 and 2024.



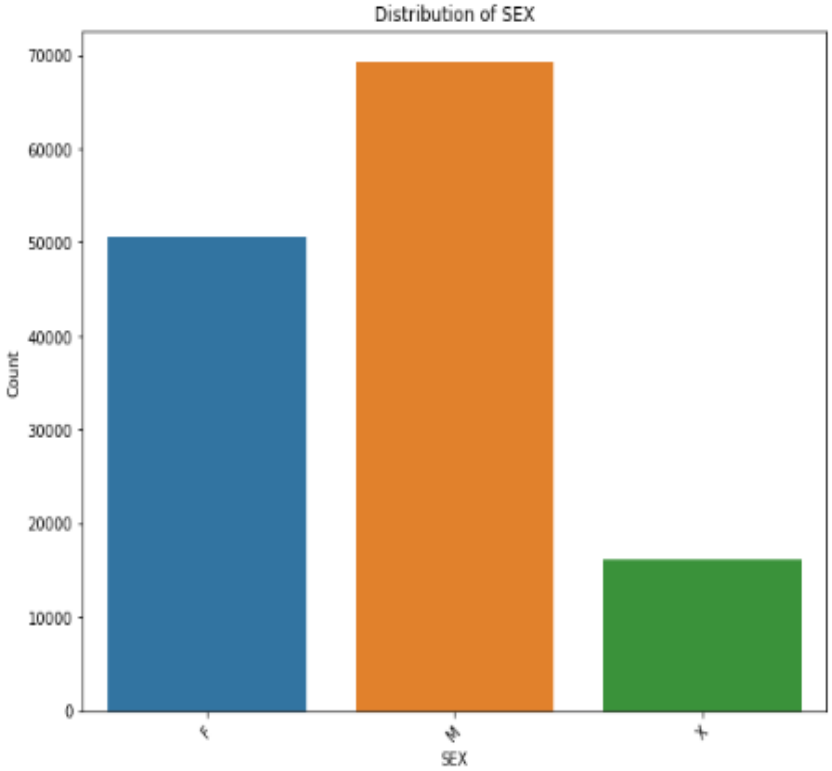
Exploratory Data Analysis (EDA)

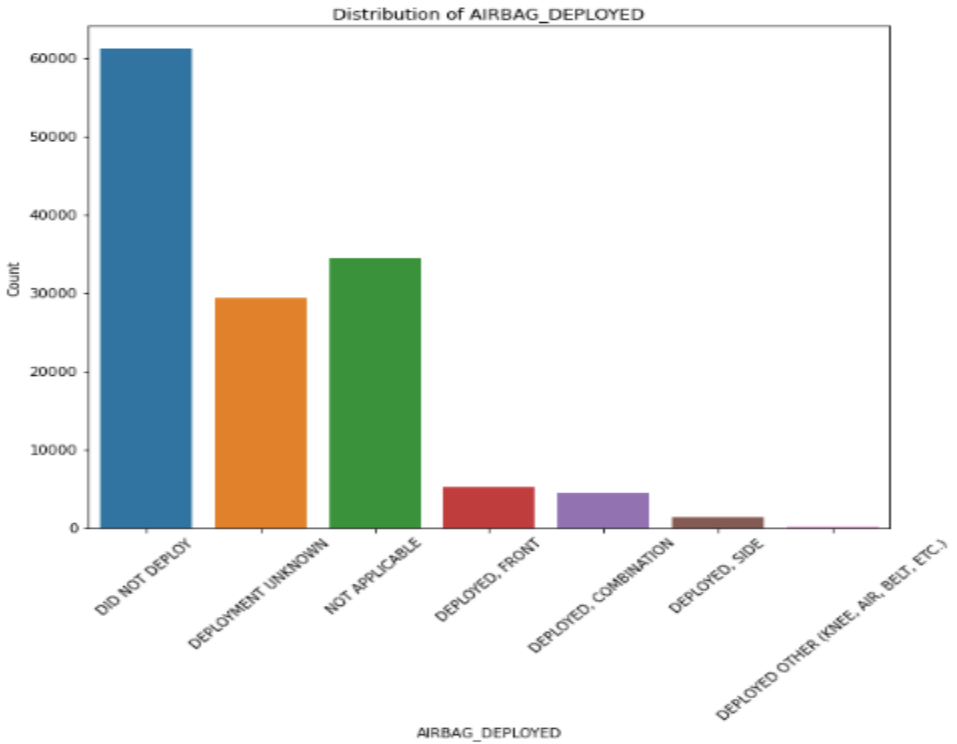
- **Univariate Analysis:** Distribution of AGE and CRASH_MONTH to understand data distribution.

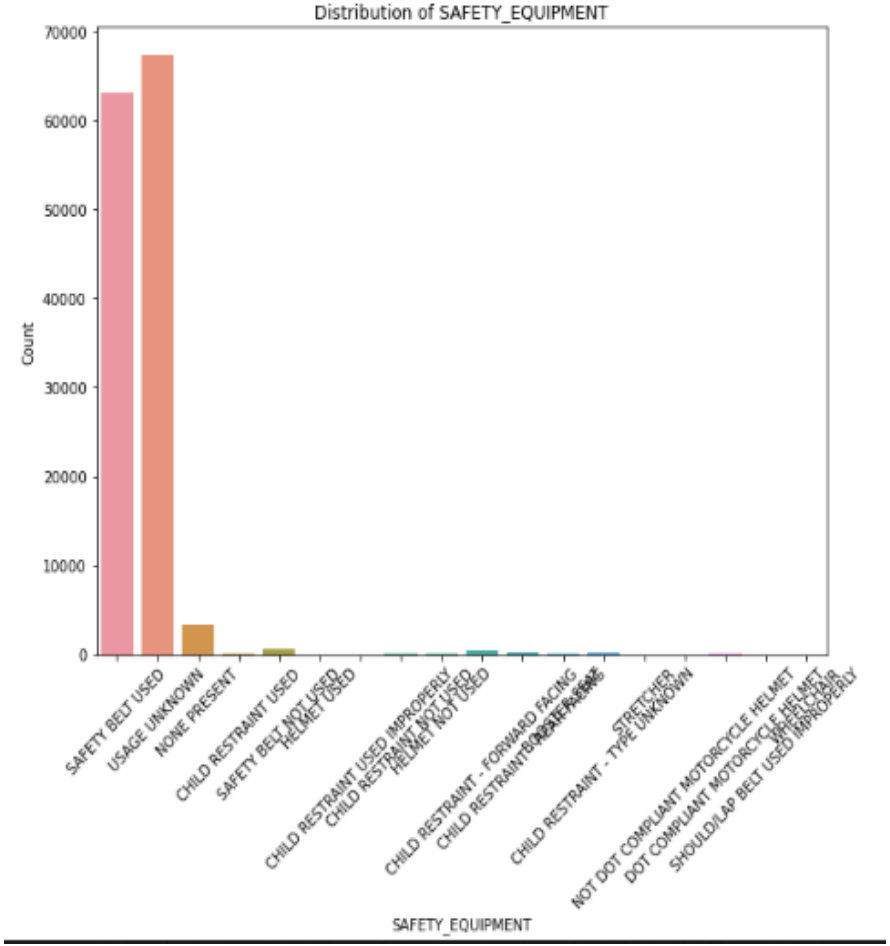


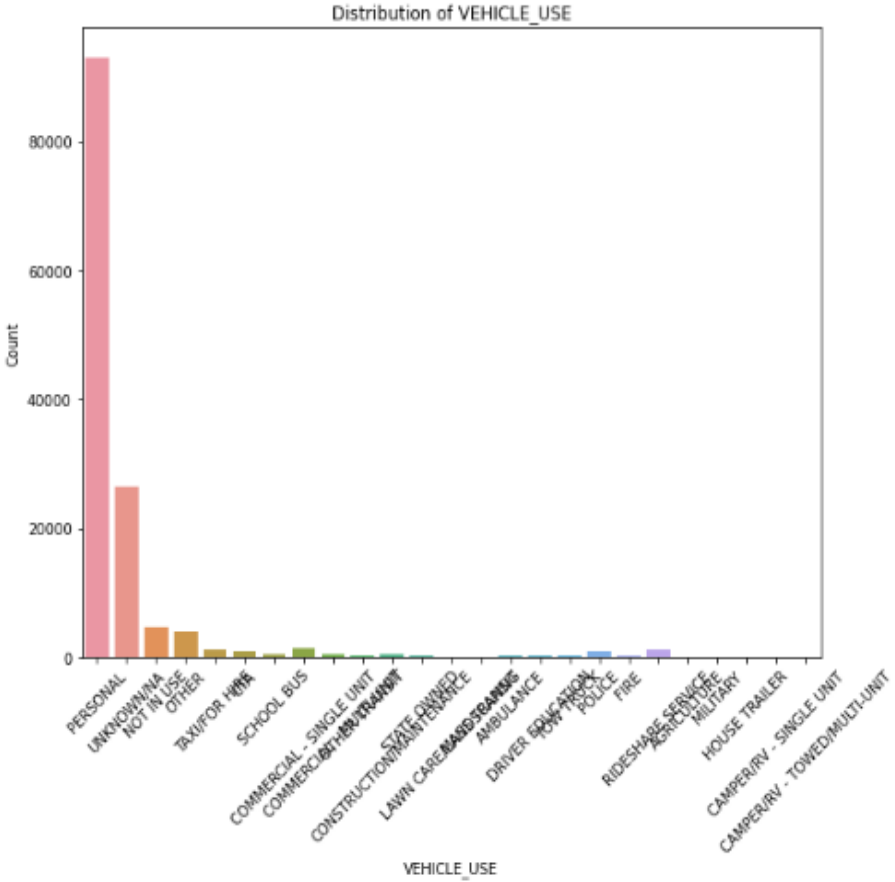


- **Categorical Variables:** Frequency analysis of variables like SEX, SAFETY_EQUIPMENT, AIRBAG_DEPLOYED and VEHICLE_USE.

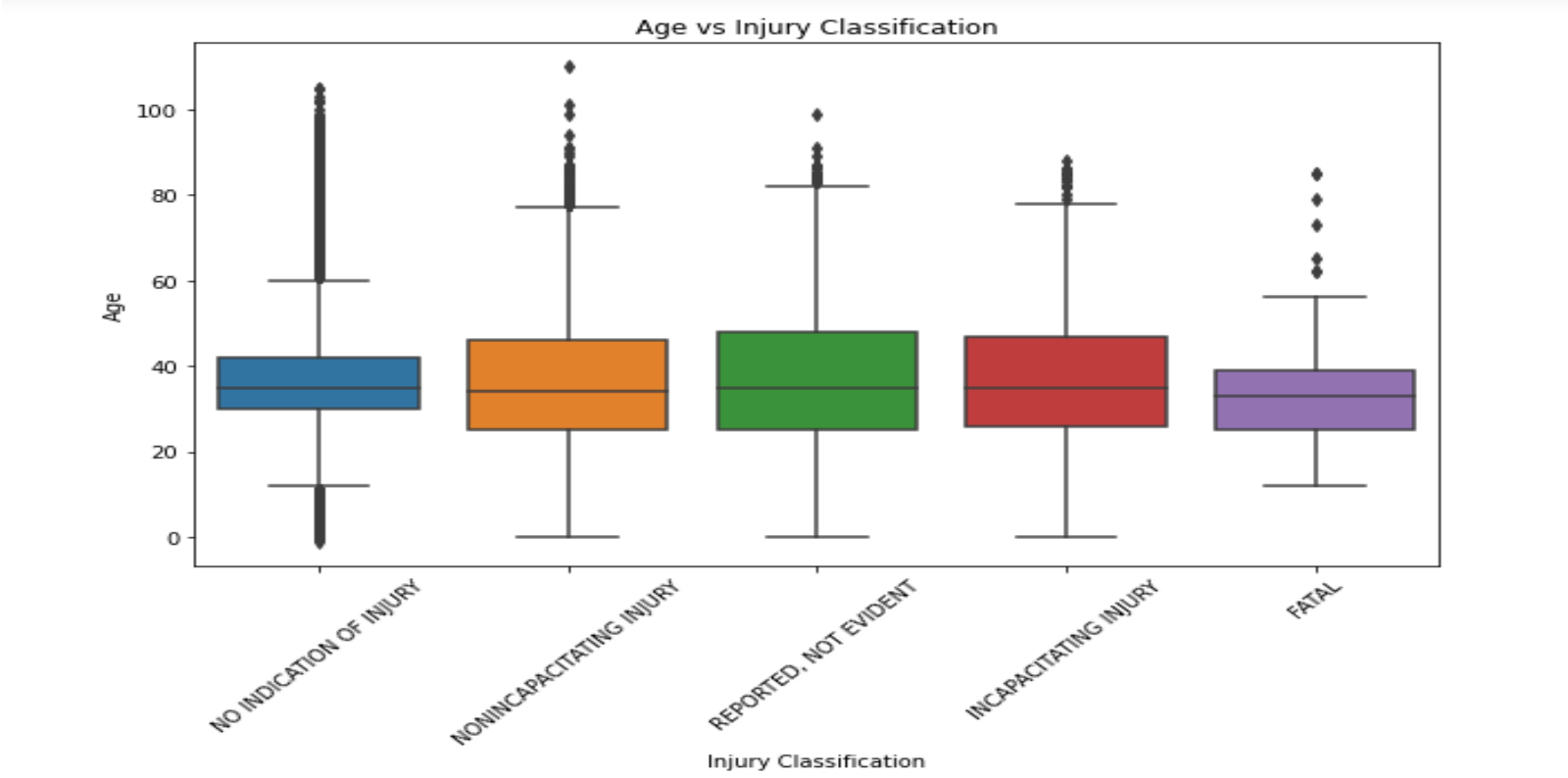





Distribution of SAFETY_EQUIPMENT
Count
SAFETY_EQUIPMENT
SAFETY BELT USED
USAGE UNKNOWN
NONE PRESENT
CHILD RESTRAINT USED
SAFETY BELT NOT USED
HELMET USED
CHILD RESTRAINT USED IMPROPERLY
CHILD RESTRAINT NOT USED
HELMET NOT USED
CHILD RESTRAINT - FORWARD FACING
STRETCHER
CHILD RESTRAINT - TYPE UNKNOWN
NOT DOT COMPLIANT MOTORCYCLE HELMET
DOT COMPLIANT MOTORCYCLE HELMET
SHOULD/LAP BELT USED IMPROPERLY


Distribution of VEHICLE_USE
Count
VEHICLE_USE
PERSONAL
UNKNOWN/NA
NOT IN USE
OTHER
SCHOOL BUS
AMBULANCE
POLICE
FIRE
MILITARY
HOUSE TRAILER
CAMPER/RV - SINGLE UNIT
CAMPER/RV - TOWED/MULTI-UNIT

- **Bivariate Analysis:** Age vs Injury Classification.



Modeling

- **Data Splitting:** Training and testing sets split with 35% test size.
- **Baseline Model:** Logistic Regression: Initial evaluation with accuracy, precision, recall, and F1 score.
- **Advanced Models:** Exploration of Random Forest and XGBoost for better performance.

NAME OF MODEL	ACCURACY	PRECISION	RECALL	F1 SCORE
Logistic Regression	0.92	0.86	0.92	0.89
Random Forest Classifier	0.92	0.88	0.92	0.88
XGBoost	0.92	0.87	0.92	0.89

Recommendations

- **Target High-Risk Age Groups:** Focus interventions on drivers aged 30 to 40.
- **Seasonal Campaigns:** Increase road safety awareness during high-risk months.
- **Vehicle Maintenance Programs:** Encourage regular maintenance to prevent defects.
- **Policy Implementation:** Address common contributory factors like driver actions and physical conditions.
- **Random Forest Classifier:** This model performed the best. Thus can be used by the Vehicle Safety Board and the city of Chicago to identify key factors contributing to car accidents and implement measures to reduce their occurrence.