PREDICTING CAUSES OF CRASHES IN THE CITY OF CHICAGO

Presented by Patricia Ngari

OVERVIEW

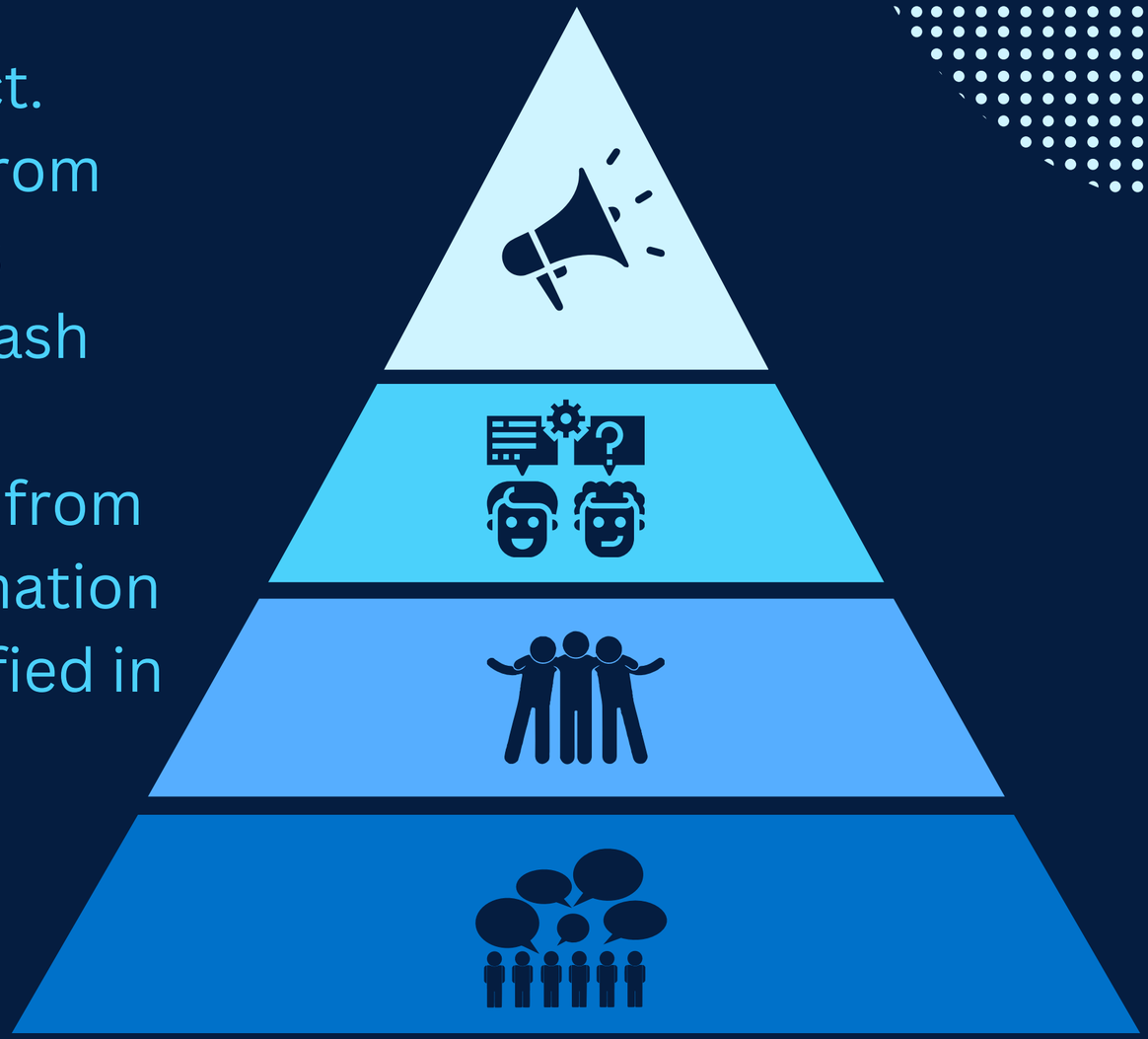
- The project utilizes advanced machine learning techniques, including classification algorithms and feature engineering, to create a robust model that can provide insights into accident causation

Objectives

- Systematically explore and clean the Chicago crash datasets to address missing values, outliers, and inconsistencies.
- Identify and engineer the most relevant features that significantly influence crash causes
- Develop and train a robust multi-class classification model
- Rigorously evaluate the model’s performance using key metrics such as F1-score, precision, and recall, and optimize it through hyperparameter tuning and cross-validation to ensure it generalizes effectively to new, unseen data.
- Analyze the model’s output to uncover patterns or common factors linked to specific crash causes, providing actionable insights that can guide policy decisions, enhance road safety measures, and shape public awareness campaigns.

Data Choice and Understanding

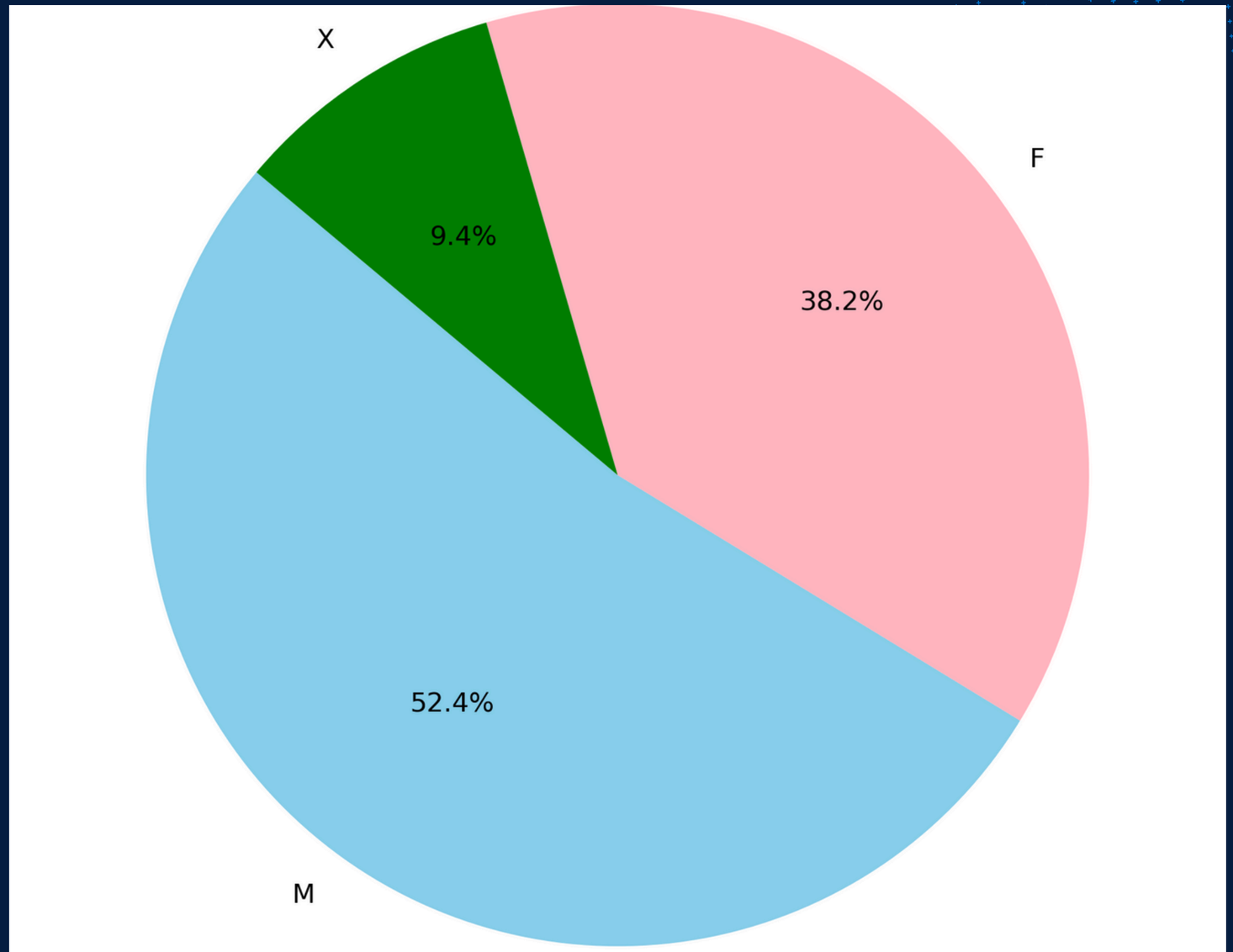
- Two datasets were merged for this project.
- Traffic_Crashes_-_People_20240824.csv from Driver/Passenger Data: This data contains information about people involved in a crash and if any injuries were sustained.
- Traffic_Crashes_-_Vehicles_20240824.csv from Vehicle Data: This dataset contains information about vehicles (or units as they are identified in crash reports) involved in a traffic crash.

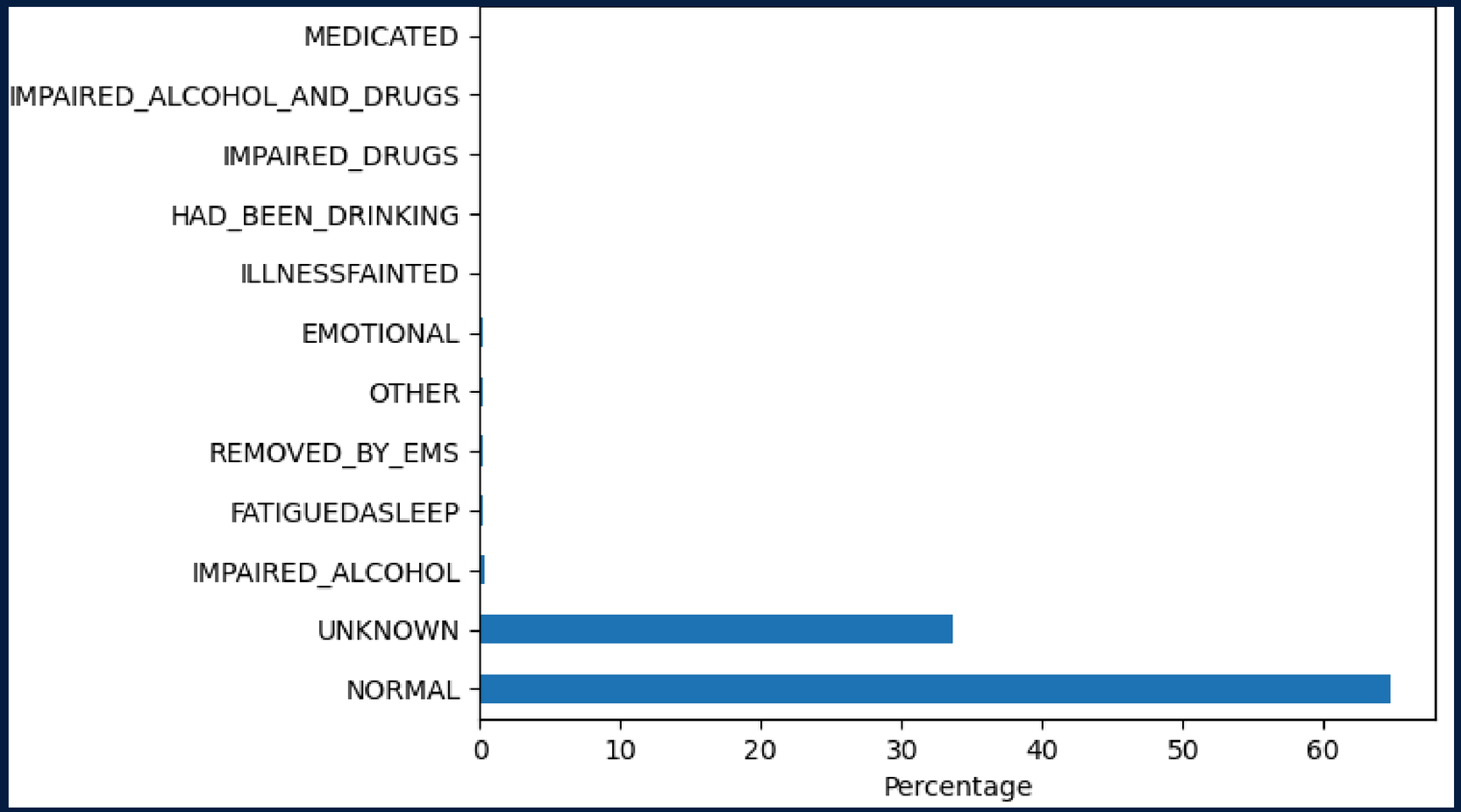


EDA

Gender Distribution

- The dataset had a larger sample of males involved in accidents compared to other gender with a leading percentage of 52.4% which is more than half of the dataset.
- The females had a 38.2% while X a non-binary or gender non-conforming identity had 9.4%.

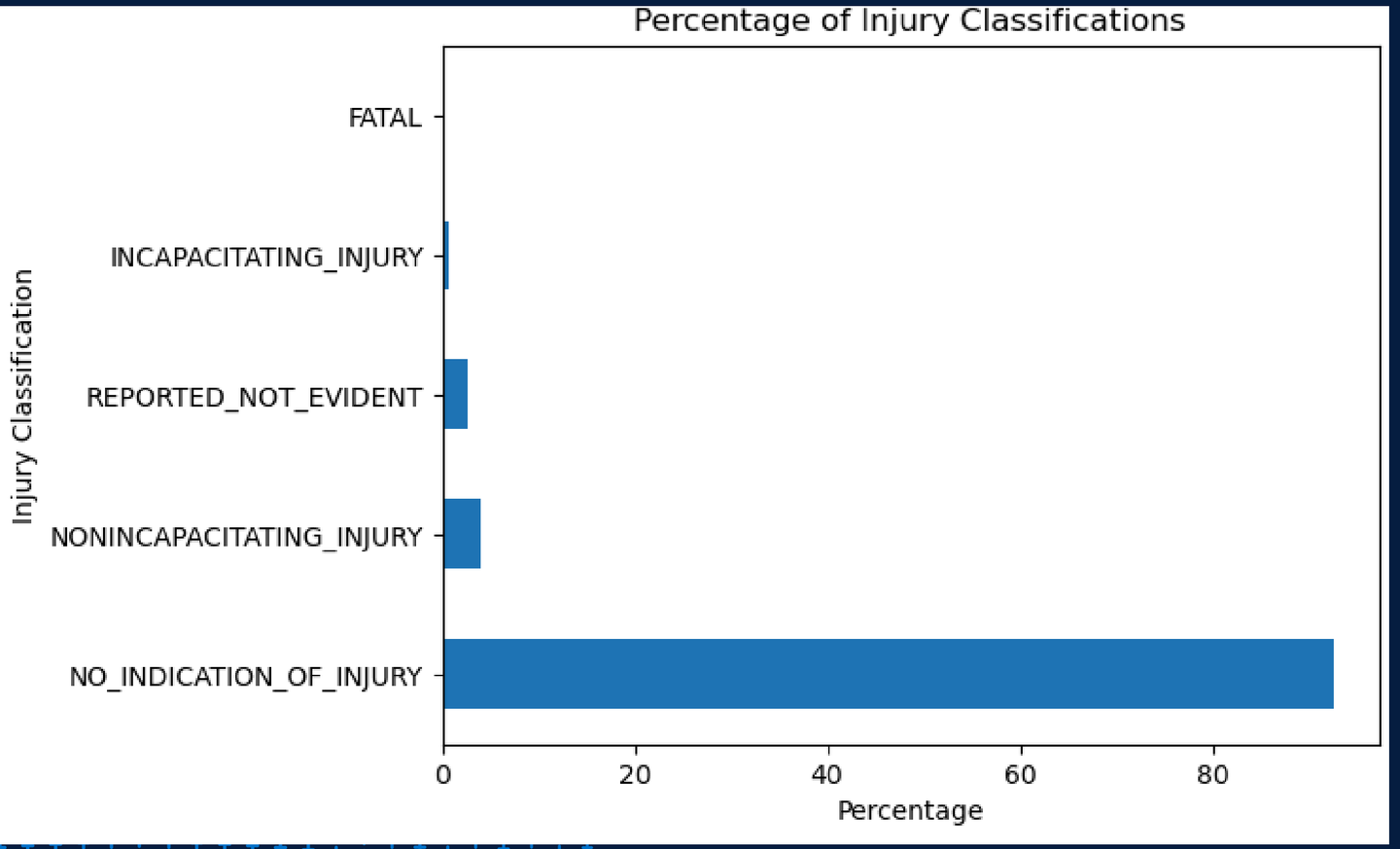




Classification of Physical Condition

- The NORMAL classification had the highest percentage with over 65% which shows that 65% of people who were involved in a crash were in normal conditions.

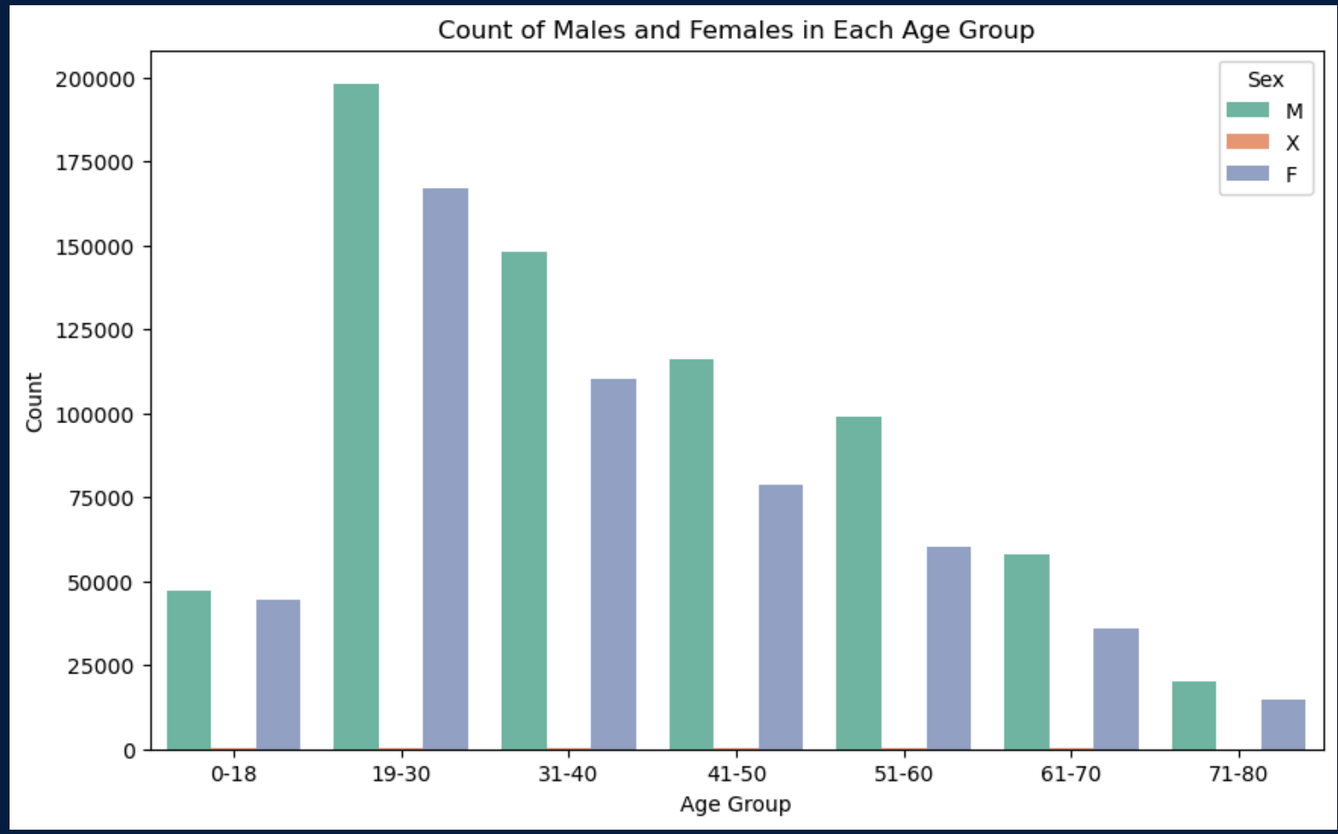
Injury Classification Percentages



- Most of the victims of the crashes had no indication of injuries.
- The crashes were either not severe or the individuals had applied protective measures.

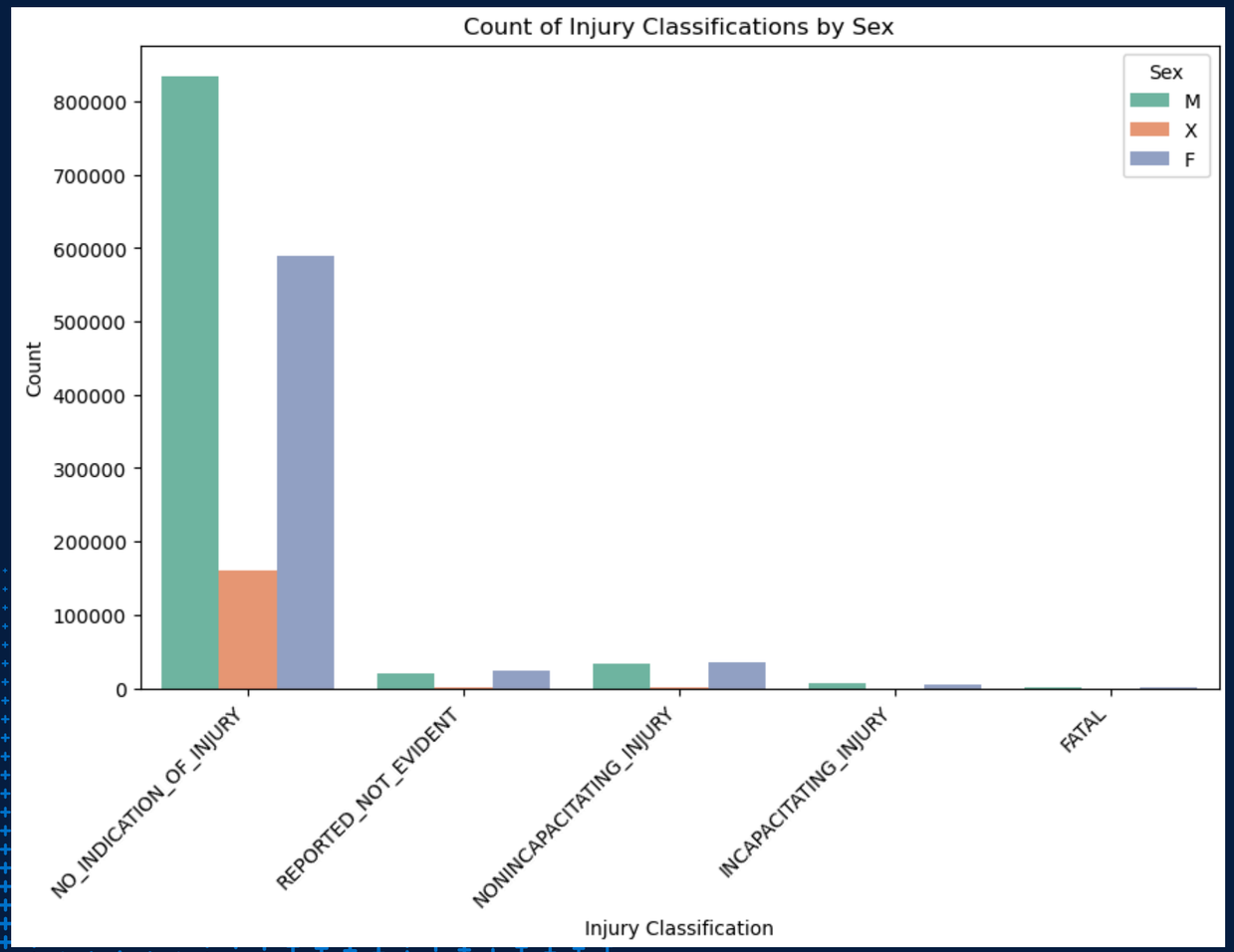
Age_groups With the Most Accident Occurrences.

- Most people involved in the car crashes were between the age of 19-30 years.
- Males in this age group had the highest occurrences with over 180000 crashes while over 170000 Females in the same age group had a crash.



Injury Classifications by Gender

- The graph shows the males leading in most categories.
- Both males and females involved in the accidents recorded no injuries.



Preprocessing

- The null values from the target column INJURY_CLASSIFICATIONS were dropped to avoid data leakage if we imputed the missing values.
- The dropped rows from our target variable were dropped from our features to maintain alignment between the features and the target variable and avoid incorrect predictions.
- Categorical variables were encoded using OneHotEncoder from scikit learn for the training data and label encoding for target variable.

Modelling

- We started by creating a dummy model. After running the cross-validation the dummy classifier results were [0.92605473, 0.92605806, 0.92605806, 0.92605447, 0.92605447]. The model seemed to perform well.
- A DecisionTreeClassifier with max_depth=5 was also trained and fit on our training data. Our decisiontree regressor showed a mean accuracy score of 92% across all folds.
- This means that the model could accurately predict 92% instances in the dataset. The model was either overfitting or biased.

Feature Selection

- Here we focused only on relevant features with the highest percentage of coefficients after fitting a DecisionTreeClassifier to mitigate overfitting and also help improve training time since fewer features mean less data for the model to process.
- F (0.977655) representing the gender for female meaning that the model consider this feature the most significant when making predictions.
- DEPLOYED_COMBINATION (0.404507): This refers to instances where multiple safety systems, such as both front and side airbags, were deployed together during an accident.
- AGE_missing (0.347520): This indicates that missing age data or where the age was not recorded has a moderate influence on the model's predictions.
- REMOVED_BY_EMS (0.299424), DEPLOYED_FRONT (0.095676) and DEPLOYED_SIDE (0.037232) were also significant in the predictions but less than the previous features.

Logistic Regression

- We introduced a logistic regression model and fit to our training set and recorded an accuracy score of 93% and it was evident the model performed well on the majority class NO_INDICATION_OF_INJURY but poorly on minority classes.
- We trained another LogisticRegression model with class_weight='balanced' to cater for the imbalances, random_state = 42, penalty = 'l2', solver='saga' and max_iter = 1000.
- The model improved in detecting minority classes but performed poorly with an accuracy of 31%

Scaling

- After scaling, recall for the minority classes (e.g., "FATAL" and "INCAPACITATING_INJURY") improved significantly but precision for these classes dropped to nearly 0.
- For "NO_INDICATION_OF_INJURY," both precision and recall dropped significantly after scaling.
- Accuracy dropped drastically from 0.31 before scaling to 0.08 after scaling.
- The macro average F1-score decreased from 0.15 to 0.05, and the weighted average F1-score dropped from 0.44 to 0.14 after scaling. This suggests that the model's overall performance across all classes worsened after scaling.

Conclusion

- The LogisticRegression model before scaling performed better.
- The feature "F" (indicating gender category for female) is the most influential factor in the model, suggesting that gender plays a significant role in the prediction of crash outcomes.
- The features related to airbag deployment and involvement of safety systems like airbags is related to crash severity or type.
- The AGE missing factor suggested that missing data on age might correlate with specific injury classification or that age itself is an important but underrepresented factor.
- Need for EMS removal suggested severe or complex crash scenarios, potentially tied to specific causes.

Reccomendations

- The original LogisticRegression model is recommended since it takes into consideration the dataset’s imbalance. It provides the best performance in terms of stability and accuracy.
- Given the high importance of the gender feature, it would be worthwhile to investigate further why gender plays such a significant role. This could involve examining gender-specific behaviors, 1types of vehicles driven, or other socio-demographic factors that may contribute to crash causation.
- The current model suggests that certain socio-demographic factors (like gender) and response indicators (like EMS involvement) are primary drivers in the predictions. To improve accuracy and interpretability, focusing on data completeness and broadening the range of features considered might yield more actionable insights.

Acknowledgement

Special thanks to my Technical Mentors from Moringa School, Mildred Jepkosgei and Faith Rotich for their continued support in this journey.

Connect with me:

patriciangari27@gmail.com
https://github.com/PatriciaNgari
https://www.linkedin.com/in/patricia-ngari/