GROUP REPORT

W07G3

INTRODUCTION: OVERALL PROJECT OBJECTIVE

"CAN WE PREDICT ACCIDENT SEVERITY USING MACHINE LEARNING MODELS?"

WHY DOES IT MATTER

- Helps identify high-risk conditions linked to severe crashes
- Provide insights for road authorities to improve high-risk zones
- Support data-driven design of targeted safety policies

METHODOLOGY: PROCEDURES AND TOOLS

ACCIDENT SEVERITY INDEX (ASI)

Formula of ASI:

$$\text{(TOTAL DEATHS} + 0.5 \times \text{SEVERE INJURIES} + 0.25 \times \text{OTHER INJURIES) / TOTAL PERSONS}$$

- ASI is a custom index we use to quantify how severe the accident was
- Normalized range from 0 to 1
- The higher the score the more severe the accident was

- Makes the prediction more accurate
- Provides a continuous measure of accident severity that we can categorize it later
- Allows machine learning models to compare severity across accidents of different sizes

ACCIDENT SEVERITY INDEX (ASI)

Distribution and Discretization

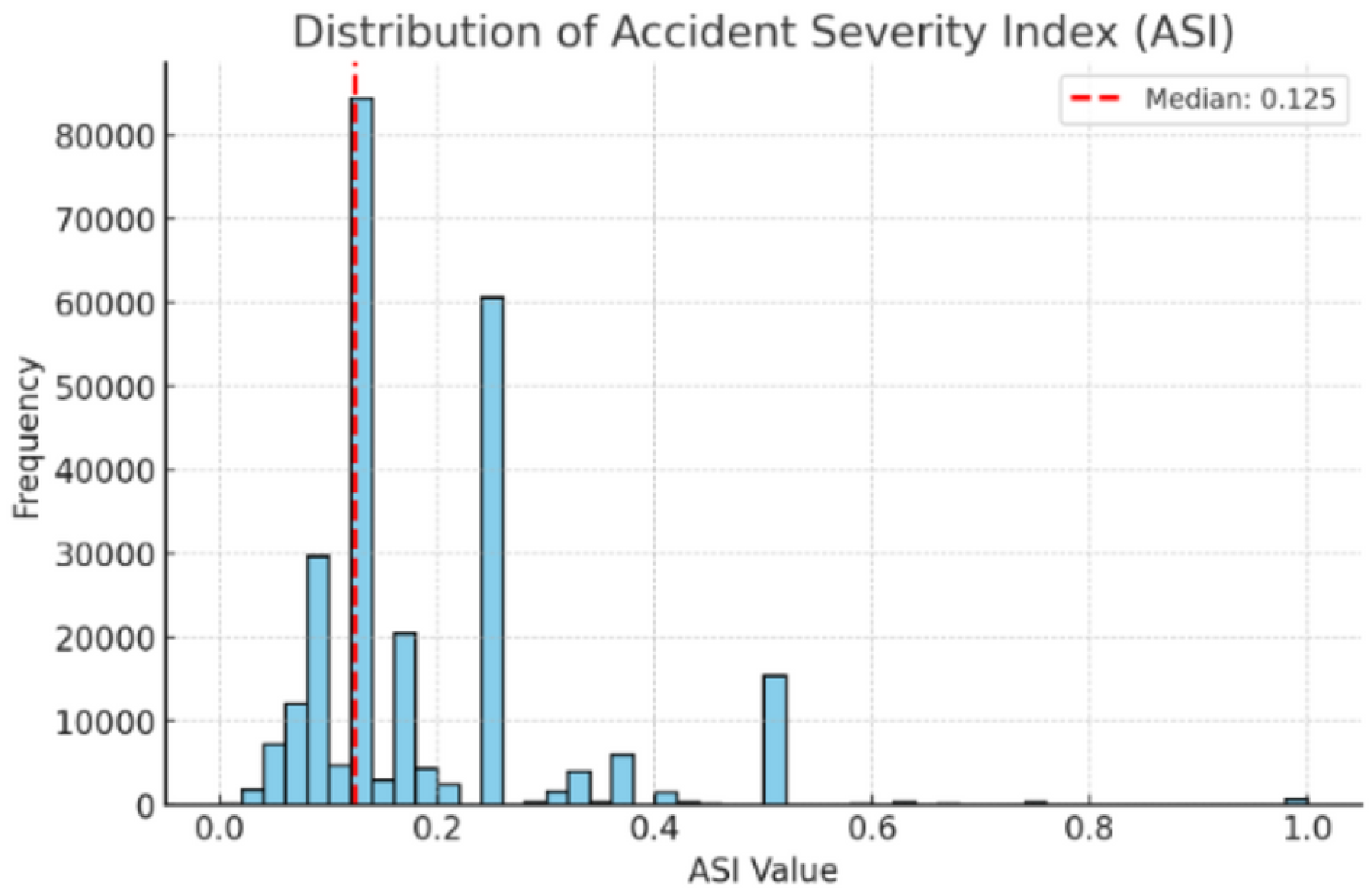


Figure 1: Distribution of Accident Severity Index (ASI)

Bin Label	Range	Description
Low	ASI <= 0.125	Majority of data, low severity
Medium	0.126 <= ASI <= 0.250	Intermediate severity
High	ASI > 0.250	Severe cases, less frequent

Table 3: ASI bin custom standard

DATA PRE-PROCESSING

- Encode categorical variables with numeric indices
- Converted time of day to seconds since midnight
- Normalized continuous variables for uniform scale
- Handled missing values

- Merged vehicle, driver and accident datasets into a unified dataframe
- Ensured each row contained a complete picture of one accident scenario

PEARSON CORRELATION

Feature	Description
NO_OF_CYLINDERS	Number of engine cylinders in the vehicle
NO_OF_WHEELS	Number of wheels on the vehicle
TARE_WEIGHT	Unloaded weight of the vehicle
SEATING_CAPACITY	Maximum seating capacity of the vehicle
ACCIDENT_TIME	Time at which the accident occurred (in seconds from midnight)
TOTAL_NO_OCCUPANTS	Number of occupants in the vehicle

MUTUAL INFORMATION ON SINGLE AND COMPOSITE FEATURES

SINGLE FEATURE

Feature	Description
ASI	Accident Severity Index
HELMET_BELT_WORN	Indicator of helmet or seatbelt usage
LICENCE_STATE	State where the license was issued
AGE_GROUP	Categorical representation of the age of individuals involved
SEX	Gender of the individual
ROAD_GEOMETRY	Configuration of the road where the accident occurred
LIGHT_CONDITION	Lighting conditions during the accident
VEHICLE_YEAR_MANUF	Year of manufacture of the vehicle
VEHICLE_BODY_STYLE	Style of the vehicle's body

COMPOSITE FEATURE

Composite Feature	Description
Speed_Road_Combo	Combination of speed zone and road geometry to capture joint road risk factors
Speed_Road_Combo_Encoded	Numerically encoded version of Speed_Road_Combo for model training
Light_Road_Combo	Combination of light condition and road geometry to capture visibility risks
Light_Road_Combo_Encoded	Encoded version of Light_Road_Combo for numerical model input
Road_Light_Index	Sum of road geometry and light condition as a simplified combined metric

MACHINE LEARNING MODEL

Machine Learning Model used :

Decision Tree , K-Nearest Neighbours (KNN)

Aim to highlight which factores are most predictive of serious outcomes in low, medium, and high severity categories

RESULTS

PEARSON CORRELATION RESULT

HEATMAP

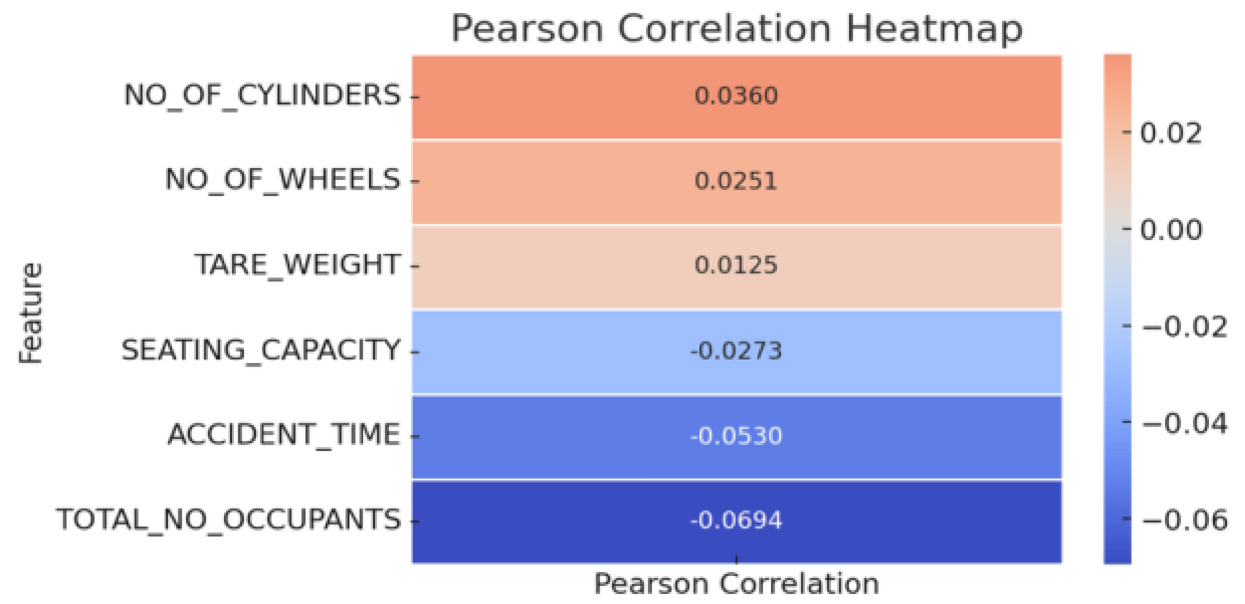


Figure 2: Pearson correlation heatmap

SCATTER PLOT

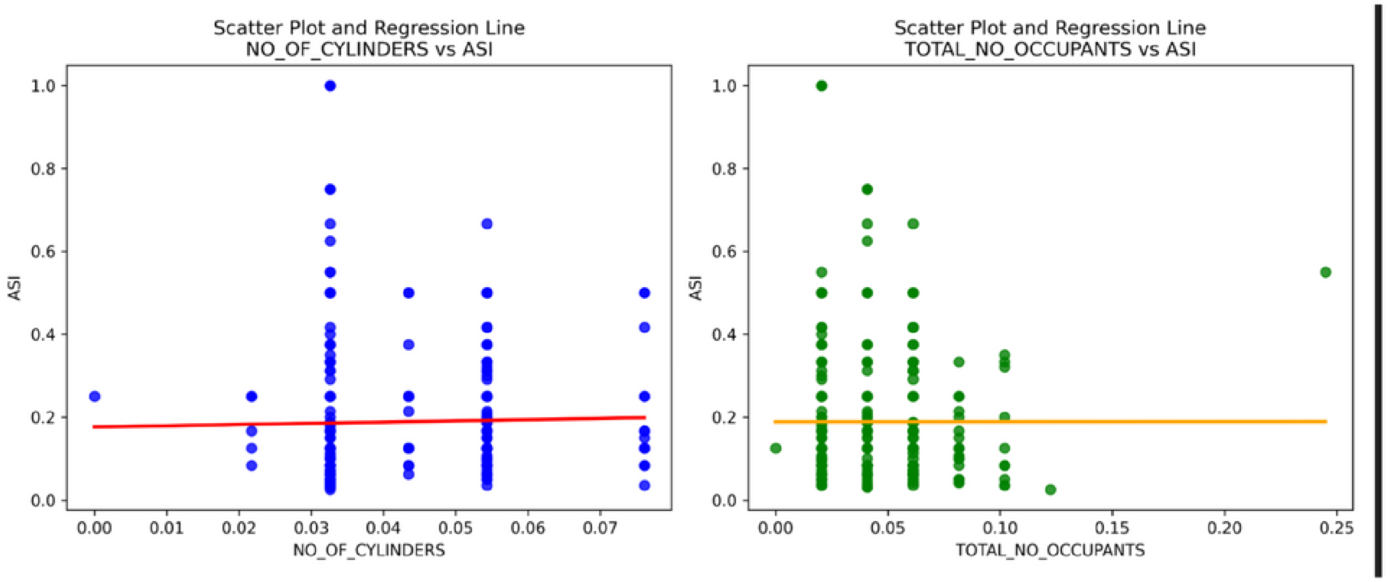
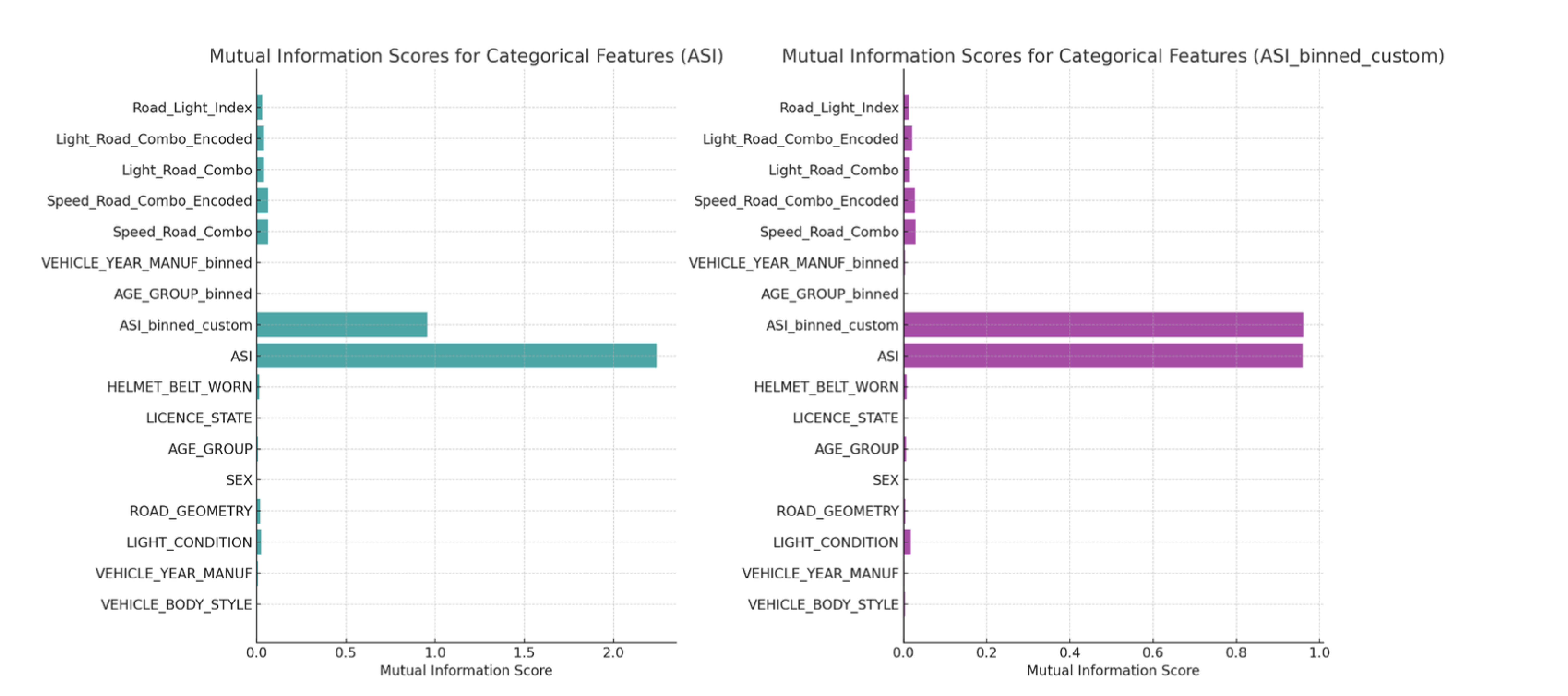


Figure 3: Scatter plot with Regression Line for two most correlated feature (No_of_Cylinders & Total_No_Occopants)

SINGLE FEATURE VS COMPOSITE FEATURE MI RESULT



MODEL PERFORMANCE OVERVIEW

Model	Accuracy	Weighted F1-score
Decision Tree	0.597	0.694
KNN	0.874	0.838

MODEL PERFORMANCE OVERVIEW

Table 3: Decision Tree - Classification Report

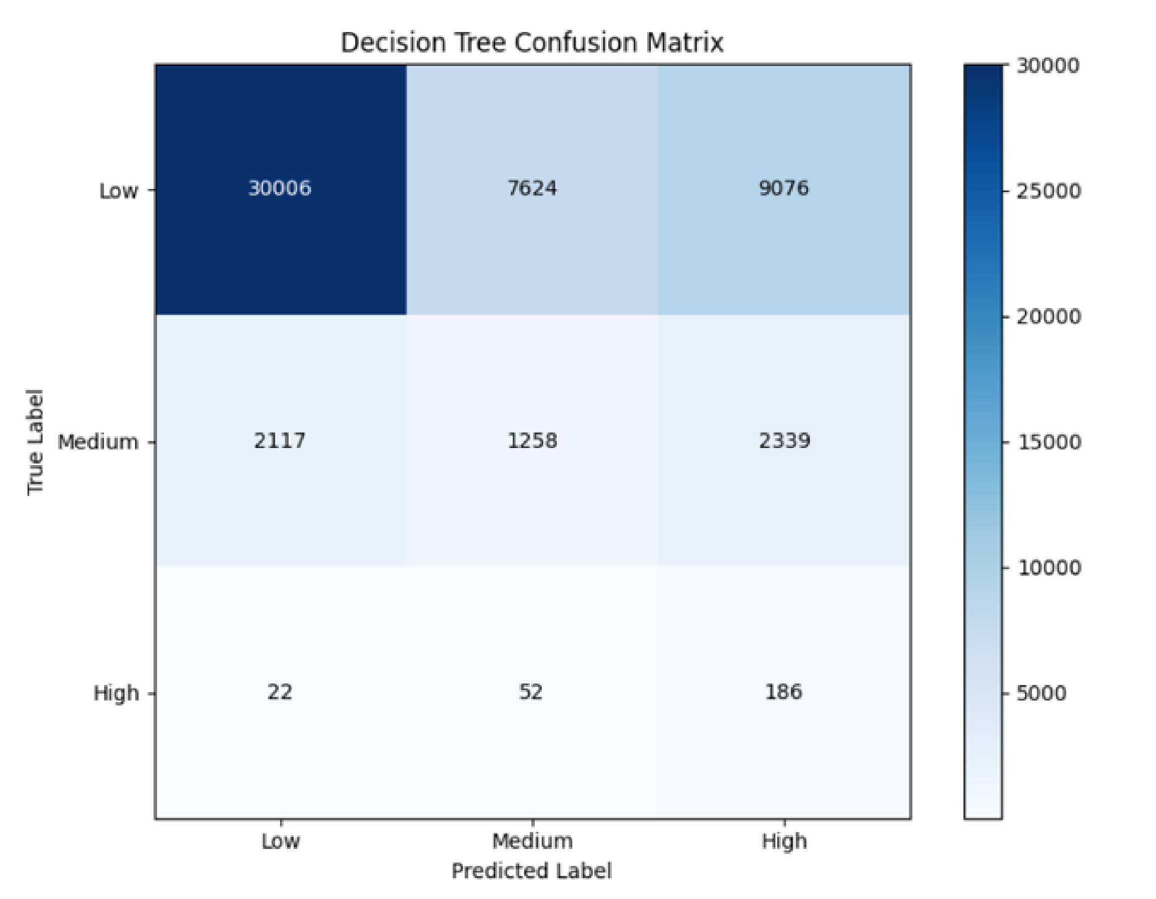
Class	Precision	Recall	F1-Score
Low	0.93	0.64	0.76
Medium	0.14	0.22	0.17
High	0.02	0.72	0.03

Table 4: KNN - Classification Report

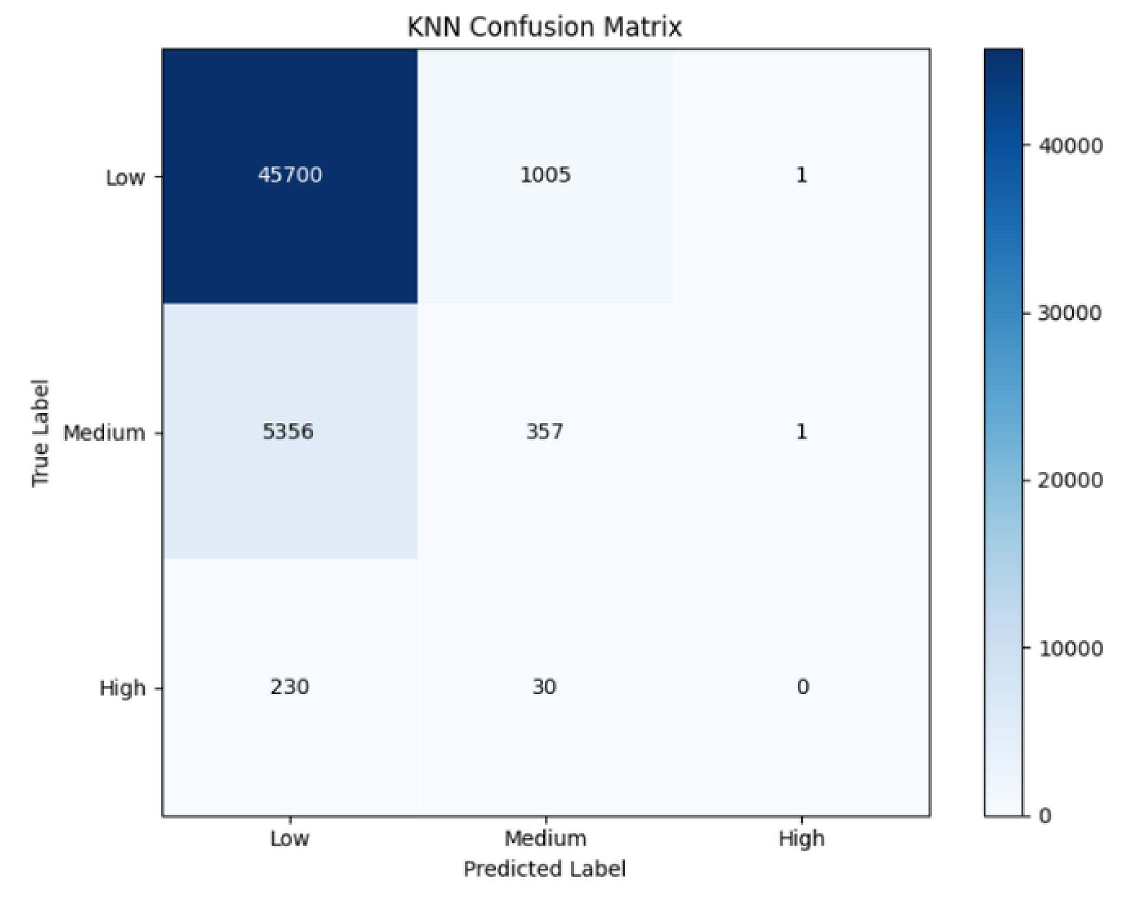
Class	Precision	Recall	F1-Score
Low	0.89	0.98	0.93
Medium	0.26	0.06	0.10
High	0.00	0.00	0.00

CONFUSION MATRIX COMPARISON

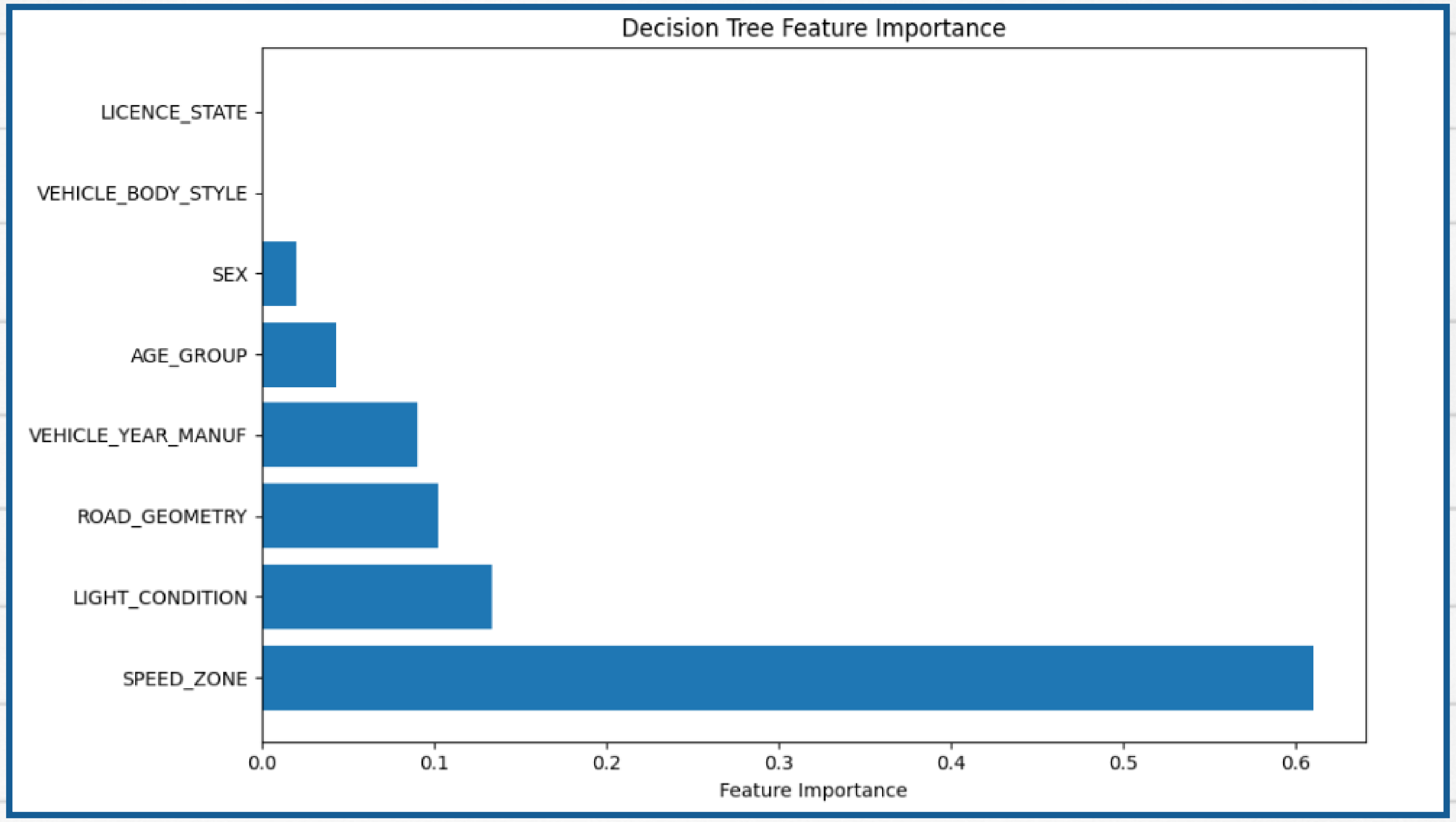
DECISION TREE MODEL



KNN MODEL



WHAT INFLUENCES SEVERITY?



Top Features

- Speed Zone
- Light Condition
- Road Geometry
- Vehicle Year of Manufacture
- Age Group

LIMITATIONS AND FUTURE IMPROVEMENTS

- Limitations:
  - Severe class imbalance led to biased predictions
  - KNN failed to predict high severity cases
  - Decision Tree's high recall for severe cases came with low precision
  - Simplified encoding missed nuanced patterns
- Improvements:
  - Use oversampling (e.g., SMOTE) to address imbalance
  - Try advanced models (XGBoost, Random Forest)
  - Include contextual data like weather and traffic
  - Engineer more composite/interaction features
  - Improve categorical variable handling

DISCUSSION: WHAT HAVE WE LEARNED ABOUT USING AI TO PREDICT ACCIDENT SEVERITY?

Table 3: Decision Tree - Classification Report

Class	Precision	Recall	F1-Score
Low	0.93	0.64	0.76
Medium	0.14	0.22	0.17
High	0.02	0.72	0.03

-KNN’s results are largely uninformative, suffering from mode collapse.

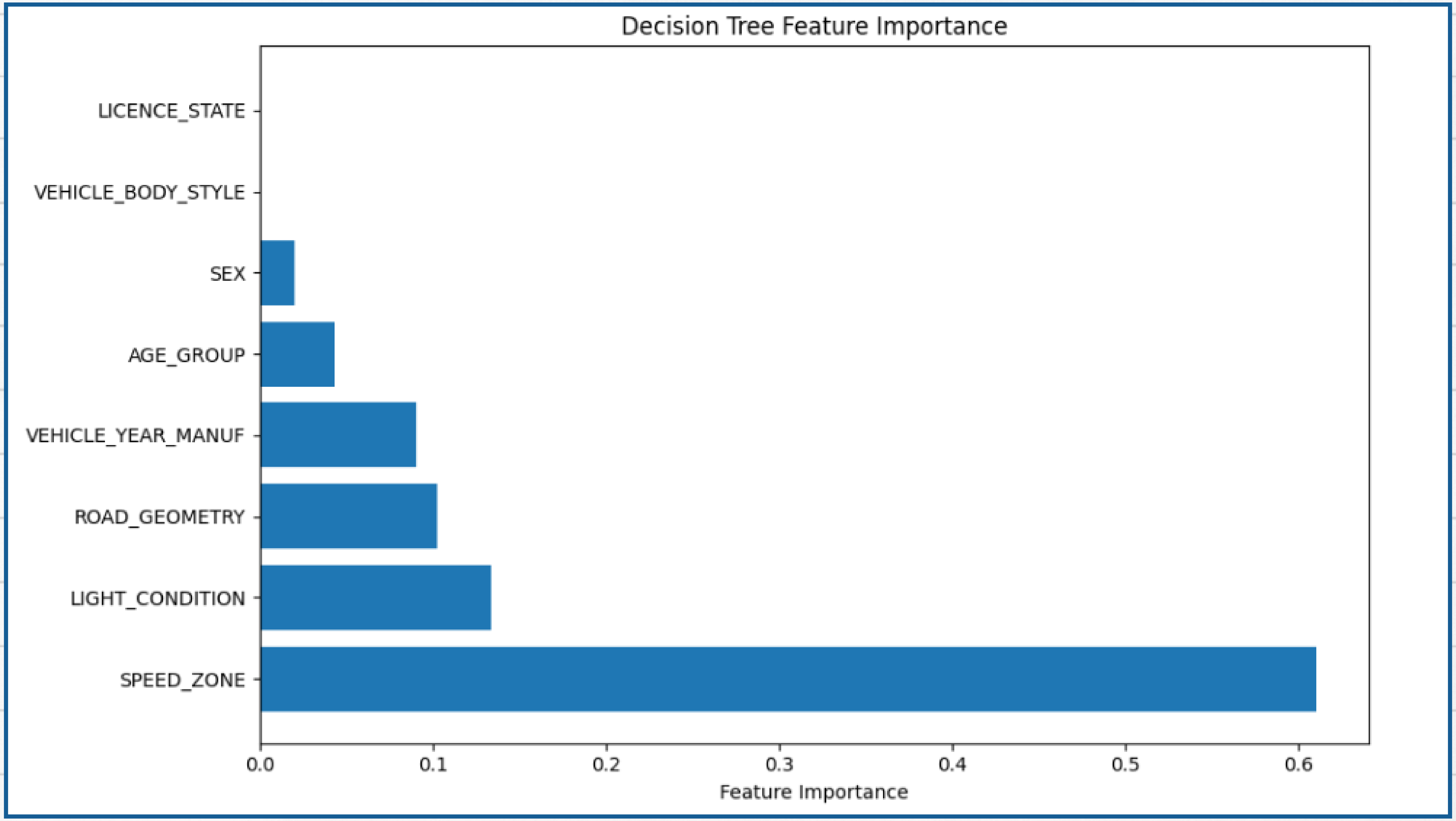
-The decision tree showed high recall but low accuracy for high severity accidents, showing a bias towards high severity (Which might be a good thing!)

-The decision tree model was able to “Catch” high severity accidents more often than not, showing it had (at least some) genuine understanding of the factors which cause high-severity accidents.

-The decision tree struggled with medium cases

-The decision tree model provides a “Proof of concept” which shows the potential for more sophisticated models in the future.

WHAT CAN WE LEARN FROM THE DECISION TREE'S FEATURE IMPORTANCE?



CONCLUSION

We have successfully provided a “Proof of concept”. Even our relatively rudimentary supervised learning models could predict accident severity to some degree

By improving on the model by providing better datasets and incorporating new variables, it is likely that future models could predict accident severity more consistently

We’ve also shown how these models can be used to provide clearer insights into the factors which influence accident severity.

THANK YOU VERY MUCH!