ANALYZING TRAFFIC ACCIDENT TRENDS

Cofinanciado por la Unión Europea

Traffic Accident Prediction

ESTONIA TEAM

The BIDUSA project (Big Data Unites the Sciences and the Arts) is co-financed by the European Union. The opinions and viewpoints expressed in all publications are solely those of their authors (Ítaca High School) and do not necessarily reflect those of the European Union or the Spanish Service for the Internationalization of Education (SEPIE). Neither the European Union nor the SEPIE National Agency can be held responsible for them.

Abstract

Objective

Analyze the distribution of traffic accidents by driver age and vehicle brand.

Method

J48 decision tree model applied to a traffic accident dataset.

Results

Key risk groups identified by age and brand.

Objective

This dataset was chosen for its comprehensive coverage of real-world traffic accidents in Estonia, offering the ability to analyze demographic patterns, particularly age-based accident risks.

Its relevance for road safety and environmental impact.

Dataset Justification

- Some classes are underrepresented (e.g. ages 0-12), and the ’Age’ attribute appears to be the target instead of accident occurrence, which may affect prediction quality.

- Original dataset contains only positive accident cases – no “non-accident” records for binary classification. No time-of-day or weather context is present. Potential underrepresentation of younger age groups.

Data Limitations

Data Processing and Formats

- The dataset was loaded into Weka in ARFF format.
- Date fields were converted using DateToNumeric where applicable.
- The “Age“ attribute was used as the target class, representing different age categories of drivers involved in accidents
- Preprocessing included attribute selection and 10-fold cross-validation.

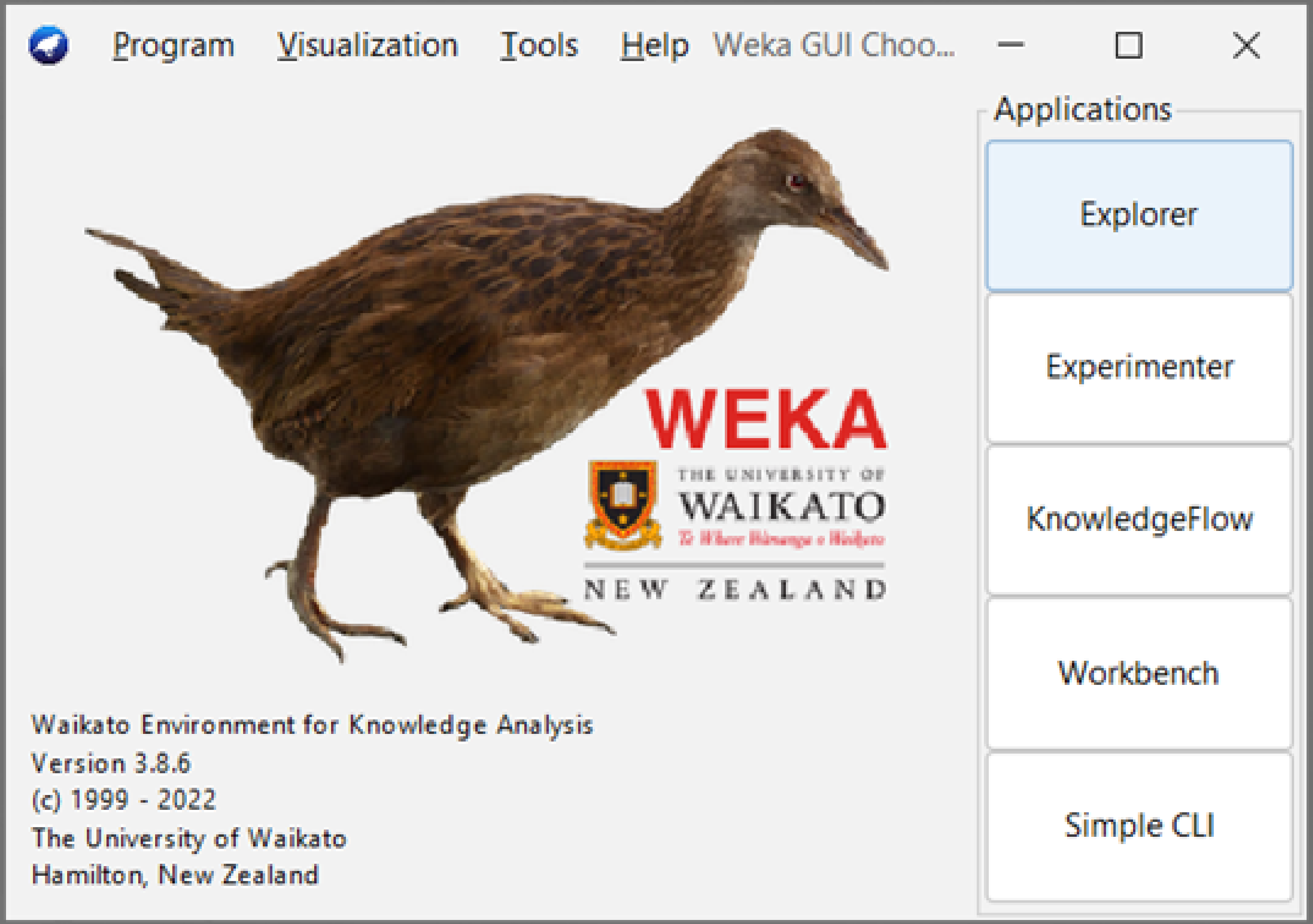
Three classifiers were compared:

- NaiveBayes: classified all into one class
- MultilayerPerceptron
- J48: interpreted patterns across vehicle type, brand, weekday, and city.
- RandomForest: tested on a reduced sample, performance similar to J48.

J48 was selected for its interpretability and logical structure.

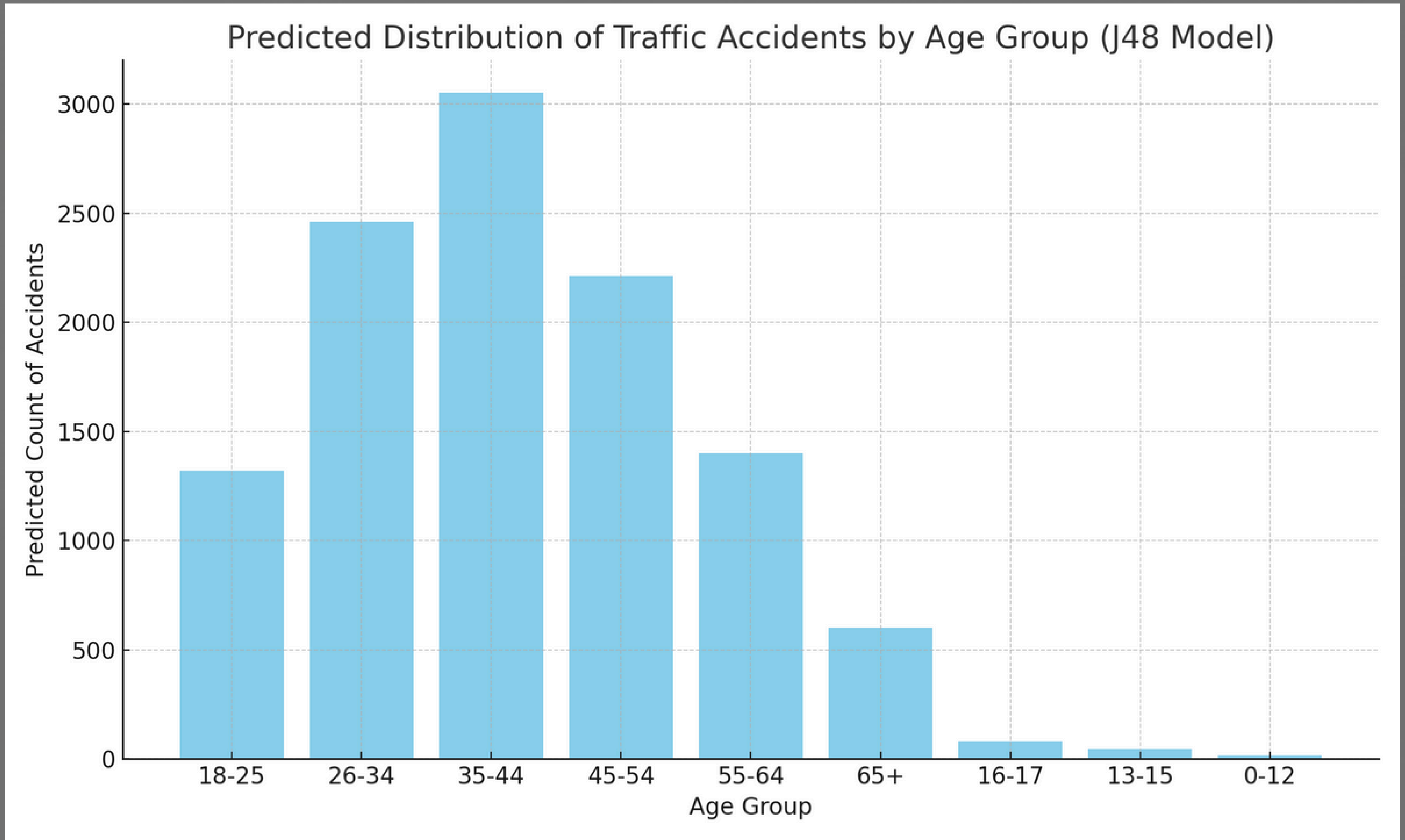
Analytical Methods

- As our main software, we used Weka.



Waikato Environment for Knowledge Analysis
Version 3.8.6
(c) 1999 - 2022
The University of Waikato
Hamilton, New Zealand

Software Used

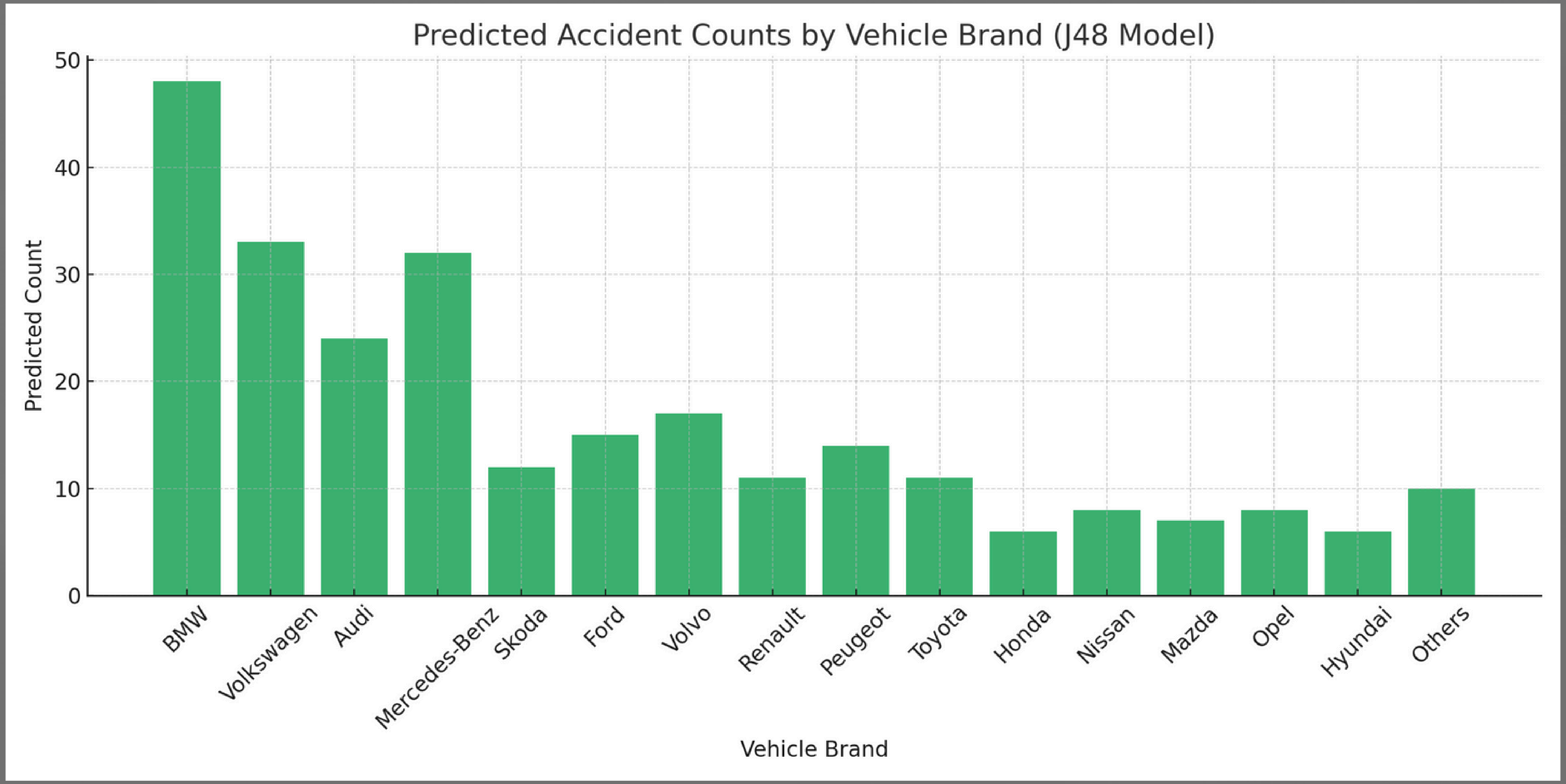


Key Results: Age Distribution

Why these age groups are high-risk.

-driving experience, habits, frequency, middle age crisis

Age Group Analysis



Key Results: Vehicle Brand Distribution

Possible reasons:

- popularity
- driving style
- vehicle power.

Brand Analysis

For insurance companies:

- traffic police
- safety campaigns.

Predictions & Practical Applications

- The study showed that machine learning models can effectively classify and detect trends in accident risk based on age.

- RandomForest achieved the best predictive performance, while J48 was used to interpret critical patterns.

- Future work could expand the dataset to include negative (non-accident) samples and additional context like time of day, weather, or alcohol influence to improve the model's realism and use in real-world risk forecasting.

Conclusion

“Data drives decisions in traffic safety.”

Thank you for your attention!