PREDICTING VEHICLE DAMAGE EXTENT IN CAR CRASHES

By: Ryan Ghimire & Aniketh Luchmapurkar

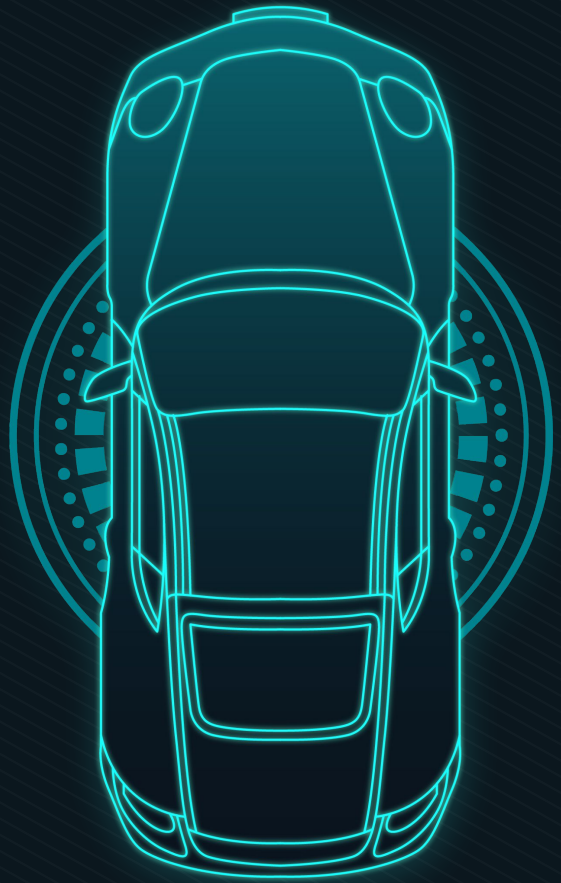
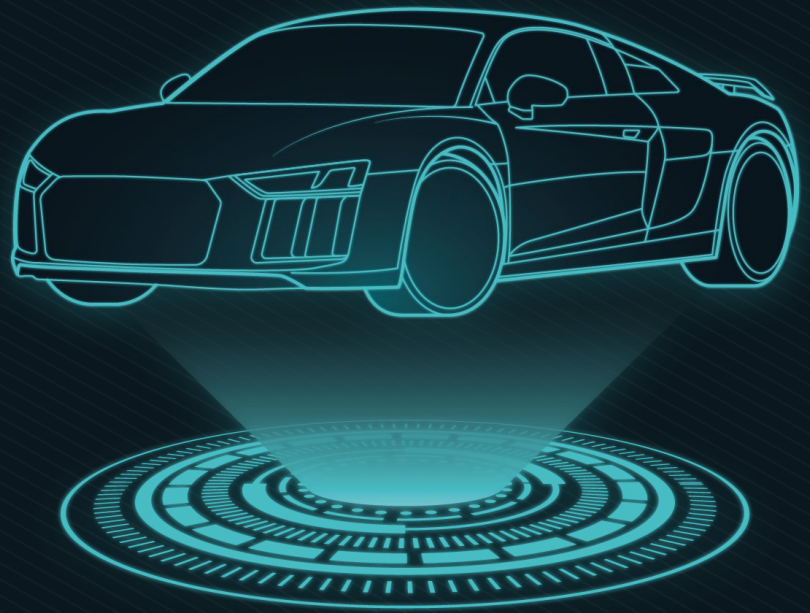


TABLE OF CONTENTS

01

PROJECT STATEMENT



GOAL

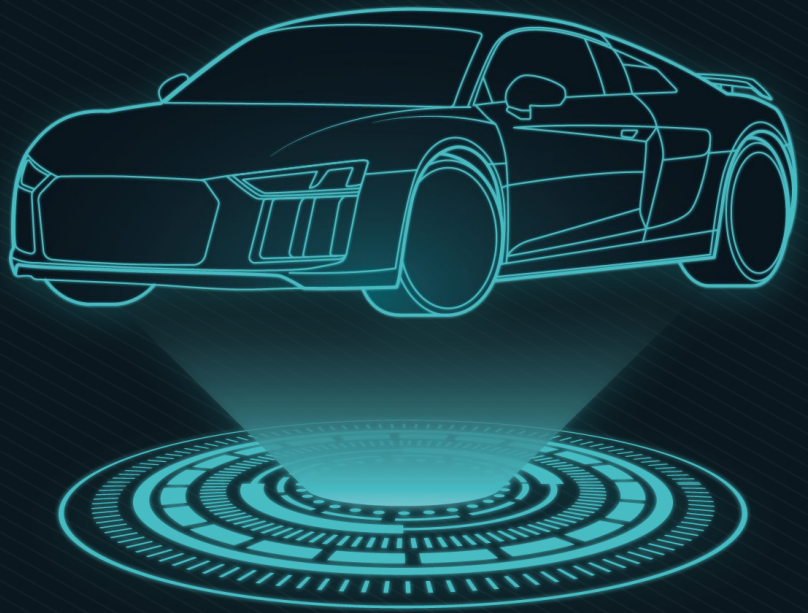
Predicting vehicle damage severity from traffic collision data

REAL-WORLD APPLICATIONS

- Helping police prioritize severe accidents
- Identifying accident-prone areas
  - Improving road and traffic safety

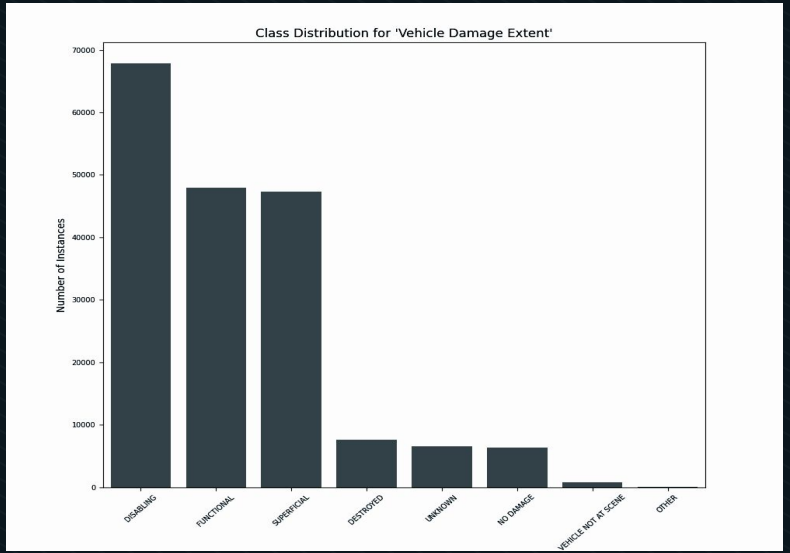
02

DATASET DESCRIPTION



DATASET

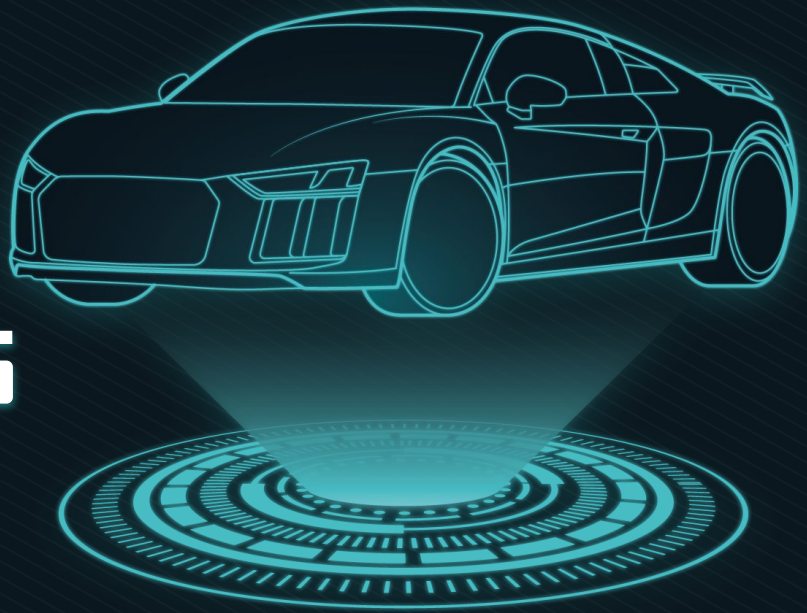
- **Source:** Montgomery County Police Department (Crash Reporting Data)
- **Size:** 184,898 instances, 38 features, 1 class attribute
- **Class Attribute:** Vehicle Damage Extent
    - **Labels:** Superficial, Functional, Disabling, Destroyed, No Damage
- **Skewness:** 0.826, skewed towards right



Attribute	Number of Missing Values	Attribute	Number of Missing Values
Report Number	184897	Traffic Control	26879
Local Case Number	184829	Driver Substance Abuse	44756
Agency Name	0	Non-Motorist Substance Abuse	180420
ACRS Report Type	0	Person ID	158038
Crash Date/Time	184897	Driver At Fault	4686
Route Type	18161	Injury Severity	789
Road Name	19364	Circumstance	165594
Cross-Street Name	28452	Driver Distracted By	36683
Off-Road Description	178336	Drivers License State	11432
Municipality	165771	Vehicle ID	158050
Related Non-Motorist	179002	Vehicle First Impact Location	3238
Collision Type	1388	Vehicle Body Type	9139
Weather	14133	Light	3826
Surface Condition	21848	Vehicle Movement	4133

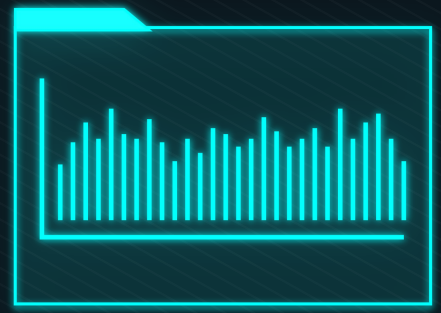
03

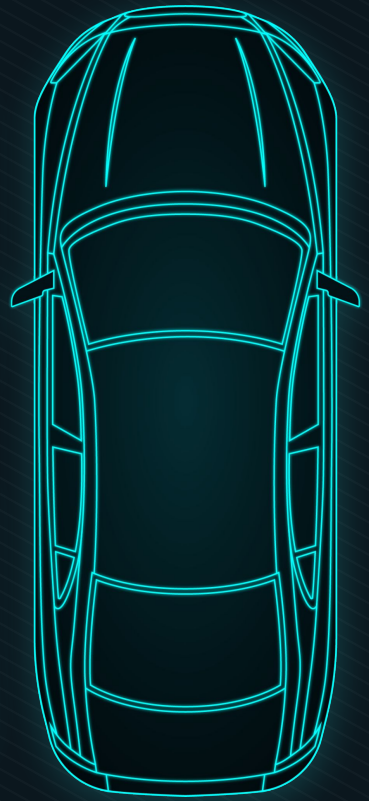
PREPROCESSING



PREPROCESSING

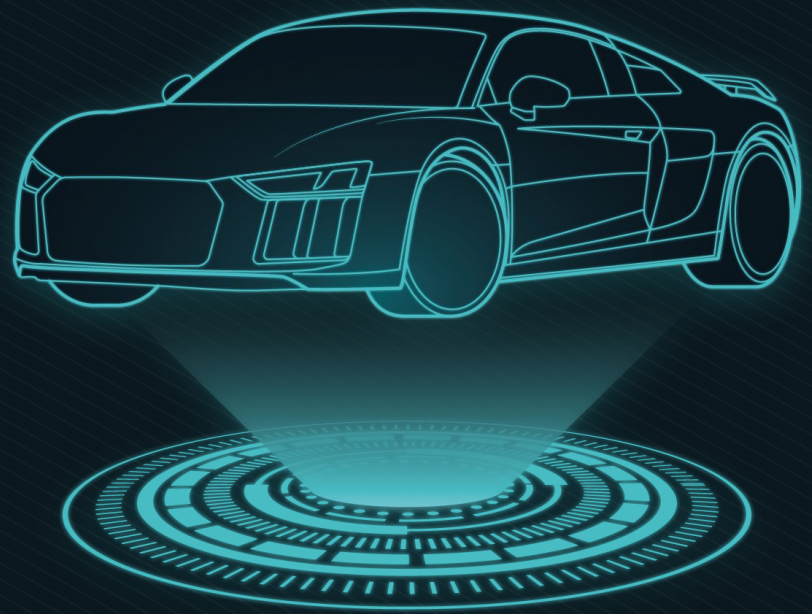
- Cleaned up dataset (“\n”, “ ‘ “, etc)
- Removed attributes (before attribute selection)
  - IDs, agency names
  - Derivable attributes like crash date/time, location
- Removed missing values
  - Blank data, other, N/A, unknown, vehicle not at scene
- Label encoding
- Z-score normalization





04

ATTRIBUTE SELECTION

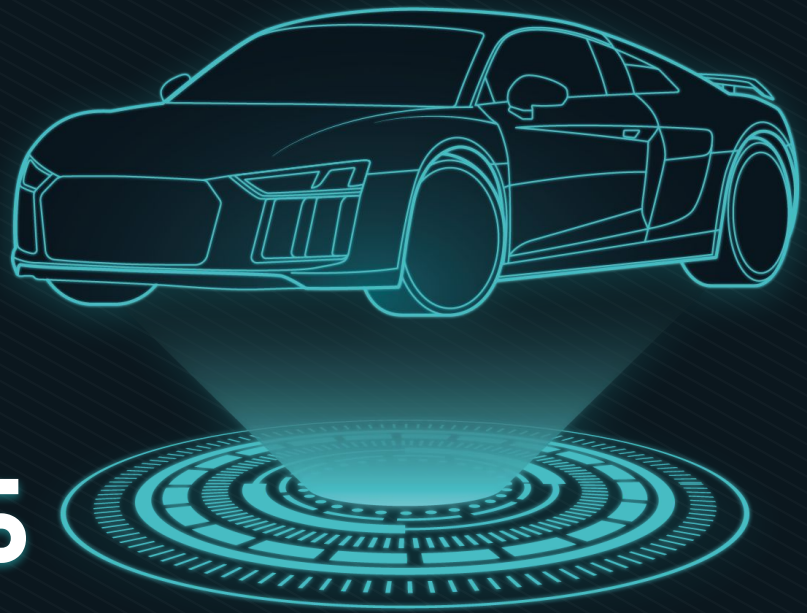


ATTRIBUTE SELECTION METHODS

- Non-Weka Manual Analysis
  - Removed “Driverless Vehicle” due to no data
- Information Gain
  - Used entropy to measure attribute importance
  - Cutoff: 0.008
- Gain Ratio
  - Adjusted Information Gain for bias
  - Cutoff: 0.005
- Correlation
  - Measured pearson correlation with class label
  - Cutoff: 0.0255
- CFS (Correlation-Based Feature Selection)
  - Selected features with high class correlation but low inter-feature correlation

05

TRAINING AND ANALYSIS



CLASSIFICATION METHODS

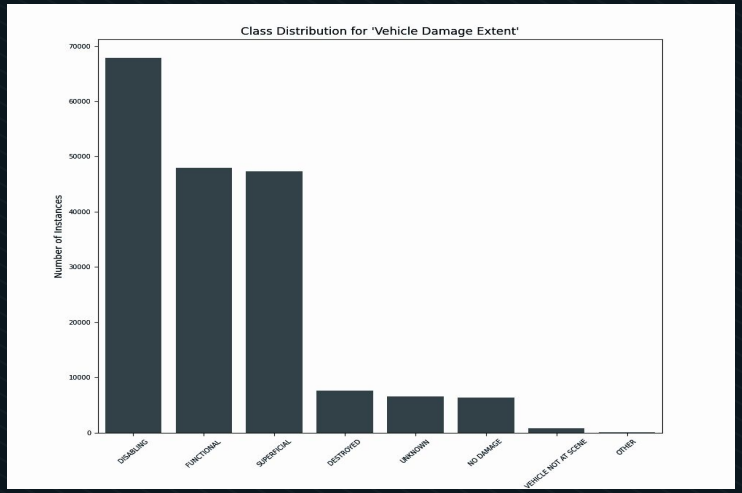
- J48 (Decision Tree)
  - Good for qualitative or discrete data
- Decision Table
  - Handles non-linear relationships well
- Random Forest
  - Good with high-dimensional data and different data types
- Naïve Bayes
  - Simple, fast, good for independent attributes

MODEL ACCURACIES

Attribute Selection

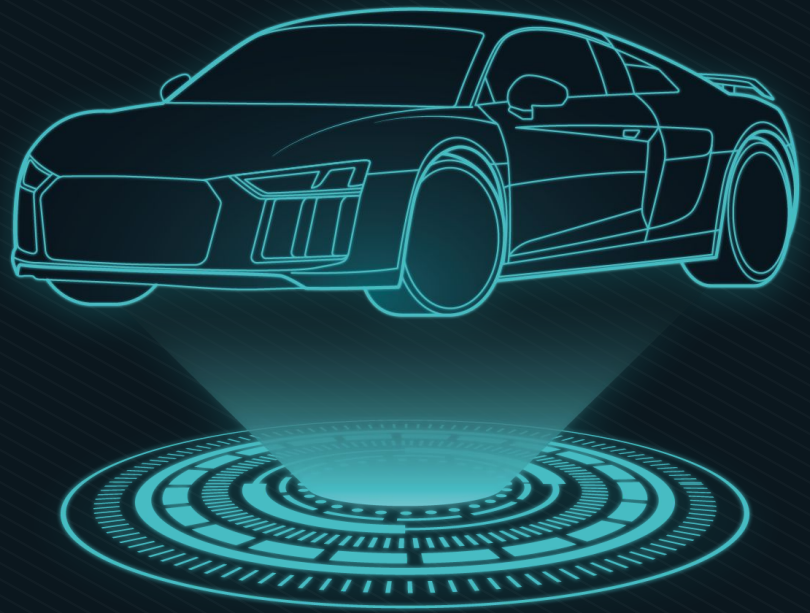
Model	Non-Weka	Info Gain	Gain Ratio	Correlation	CFS
J48	44.9063%	44.4403%	47.2168%	45.7673%	46.3451%
Decision Table	51.8493%	51.8493%	51.8493%	50.2315%	51.6947%
Naive-Bayes	43.4918%	43.4792%	44.417%	44.2477%	46.1369%
Random Forest	**53.616%**	53.5926%	53.4486%	50.3745%	48.8958%

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.817    0.384    0.589      0.817    0.685      0.428    0.783     0.666     DISABLING
                 0.483    0.162    0.501      0.483    0.492      0.325    0.772     0.509     SUPERFICIAL
                 0.298    0.156    0.428      0.298    0.352      0.161    0.646     0.397     FUNCTIONAL
                 0.006    0.000    0.532      0.006    0.012      0.055    0.801     0.171     DESTROYED
                 0.066    0.001    0.604      0.066    0.120      0.194    0.814     0.205     NO DAMAGE
Weighted Avg.    0.536    0.239    0.520      0.536    0.502      0.307    0.743     0.520

=== Confusion Matrix ===

     a     b     c     d     e   <-- classified as
 33785  3242  4267    22    19 |     a = DISABLING
  6756 12496  6575     0    69 |     b = SUPERFICIAL
 12660  7625  8631     0    25 |     c = FUNCTIONAL
  3566   124   237    25     3 |     d = DESTROYED
   610  1445   435     0   177 |     e = NO DAMAGE
```



06

CONCLUSION



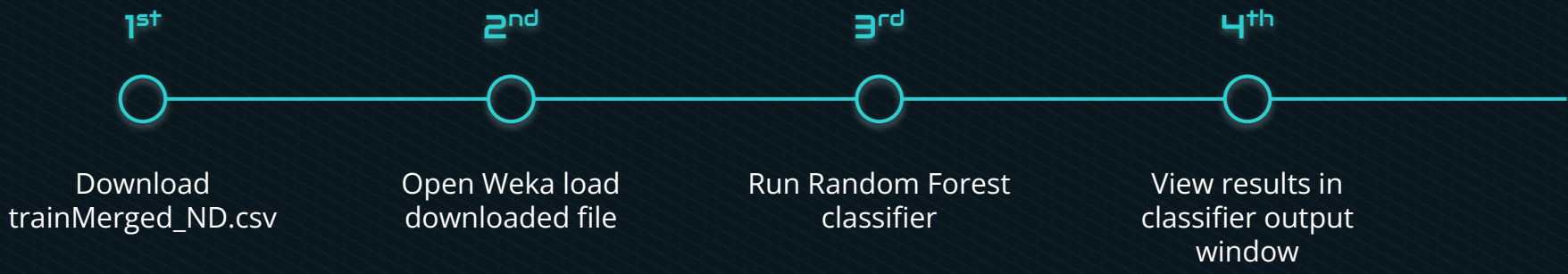
CONCLUSION

- Low accuracies attributed to multiple factors
  - Make and model have significant impact
    - Not represented well in dataset
  - Minor changes in angle and speed make major differences
  - Subjective labels
    - Functioning vs superficial
  - Skewed towards disabling label
  - High amount of missing values

IMPROVEMENTS

- PCA
  - Assumes linear relationship between variables
- Balancing class labels
  - SMOTE
- In the future we will:
  - Obtaining better data
    - More varied
    - More detailed
    - Less missing values

REPRODUCING OUR MODEL



THANK YOU!

ANY QUESTIONS?