W207 SUMMER 2025
Oviya Adhan
Nory Arroyo
Caitlin Gainey
Christine Sako


(Wojcicki, 2018)

Ahead of the Impact:

Predicting Injury Severity for Safer Roads

A Machine Learning approach to uncover crash patterns that guide safer, smarter integration of automated driving technologies.

W207 SUMMER 2025

Contents

Oviya Adhan ▪ Nory Arroyo ▪ Caitlin Gainey ▪ Christine Sako

W207 SUMMER 2025

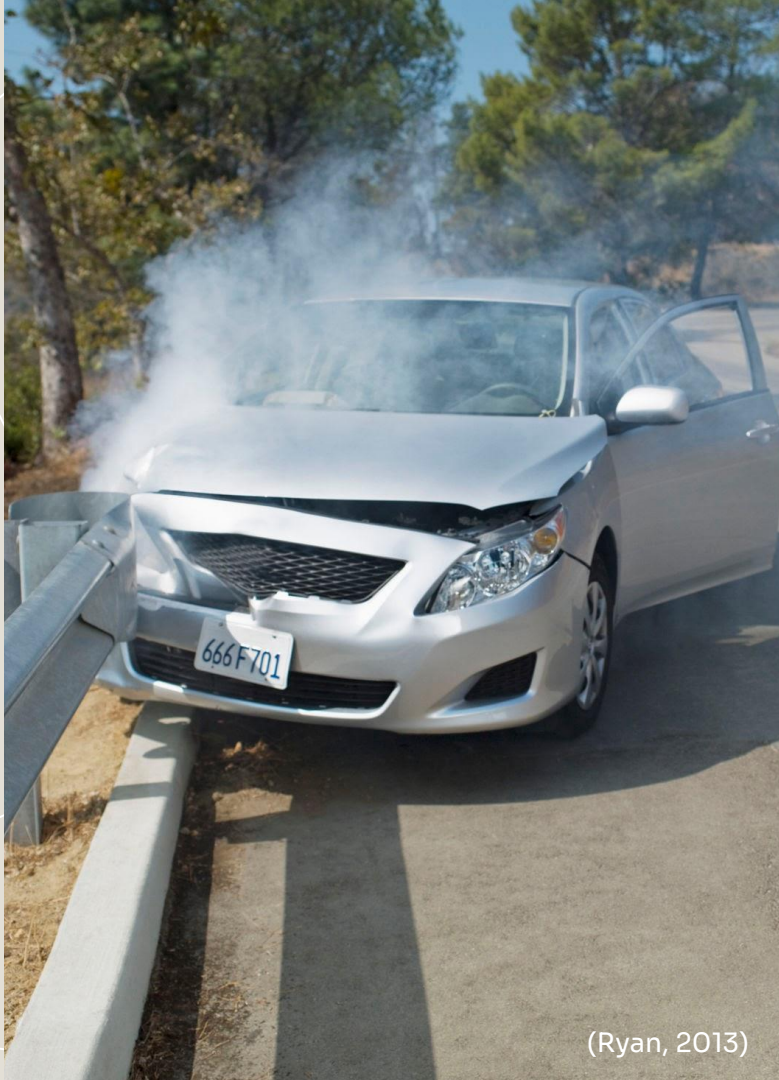
Introduction

Automated Systems Are on the Rise

Crash Data Reveals Critical Patterns

Machine Learning Offers New Tools

Planning for Safety and Equity



(Ryan, 2013)

Oviya Adhan ▪ Nory Arroyo ▪ Caitlin Gainey ▪ Christine Sako

Related Work

Work	Methods	Results
Ardakani et al. (2023)	"*A study on road accident prediction and contributing factors using explainable machine learning models: Analysis and performance*"	→ Decision Trees → Random Forest → Multinomial Logistic Regression → Naive Bayes	Three of the four models produced an "acceptable level of accuracy" for car accident prediction
Pourroostaei Ahmed et al. (2023)	"*Road Car Accident Prediction Using a Machine-Learning-Enabled Data Analysis*"	→ Random Forest → Decision Jungle → Adaptive Boosting → Extreme Gradient Boosting → Light Gradient Boosting → Categorical Boosting	Use of explainable ML (XML) allowed the researchers to retrain and tune the models to make more accurate predictions. Random forest was found to be the best model with accuracy, precision, recall, and F1 all falling within the 81-82% range.

Dataset

California Crash Reporting System (CCRS)

- ➔ California Open Data Portal, California Department of Technology
- ➔ Collected by the California Highway Patrol - Information Technology Division
- ➔ Total Data available from 2016 through to-date 2025

Oviya Adhan ▪ Nory Arroyo ▪ Caitli

W207 SUMMER 2025

Preprocessing

We preprocess 3 source datasets to arrive at 2 time separated datasets for training/validation (2024) and testing (2025)

1

Crash Dataset

Data from individual automobile crashes

`crashes_2024`:
194.5 MB
shape: (410348, 73)

`crashes_2025`:
65.9 MB
shape: (140311, 73)

2

Parties Dataset

Data about parties involved with crashes

`parties_2024`
181.3 MB
shape: (801856, 38)

`parties_2025`
63 MB
shape: (272242, 38)

3

Injured/Witness/ Passengers Dataset

Data about possible injuries from those involved as witnesses or passengers in crashes .

`injuredwitnesspassengers_2024`:
63.1 MB
shape: (485031, 21)

`injuredwitnesspassengers_2025`:
24.1 MB
shape: (164580, 21)

4

Merged Dataset

One sample record per crash, with a single, most extreme outcome from aggregated parties

`final_merged_2024`:
98.5 MB
shape (406874, 13)

`final_merged_2025`:
35.5 MB
shape (139443, 13).

Oviya Adhan ▪ Nory Arroyo ▪ Caitlin Gainey ▪ Christine Sako

W207 SUMMER 2025

Features

We merge 5 features from crash data,
6 from parties data,
and 1 single outcome variable from our injured/witness/passengers data

1

Crash Features

`CollisionTypeDescription`, category
`IsHighwayRelated`, bool
`Weather1`, category
`RoadCondition1`, category
`LightingDescription`, category

2

Parties Features

`SpeedLimit`, float64
`MovementPrecCollDescription`, category
`AirbagDescription`, category
`SafetyEquipmentDescription`, category
`SobrietyDrugPhysicalDescription1`, category
`SpecialInformation`, category

3

Injured/Witness Passengers Features

`ExtentOfInjuryCode`, category

4

Outcome Variable

`No Injury` - 61.4%
`Minor` - 34.9%
`Serious` - 2.9%
`Fatal` - 0.8%

Oviya Adhan ▪ Nory Arroyo ▪ Caitlin Gainey ▪ Christine Sako

W207 SUMMER 2025

EDA

Step 1

Consolidate and merge 3 datasets

Step 2

Assign worst case injury (outcome variable) per crash

Step 3

Check distribution of injuries (outcome variable)

Step 4

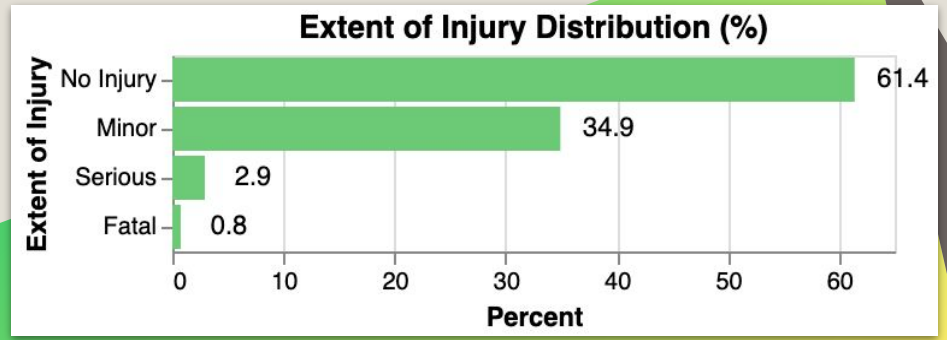
Check distribution of injuries (outcome variable) across features of interest

Step 5

Normalize numerical features and one-hot encode categorical variables

Step 6

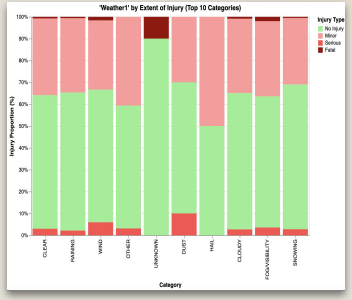
Split into training, validation, and test datasets

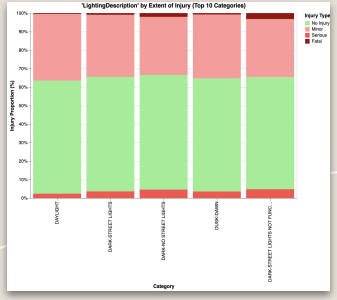


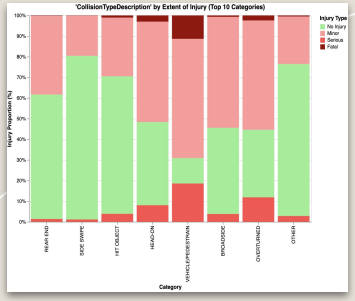
Oviya Adhan ▪ Nory Arroyo ▪ Caitlin Gainey ▪ Christine Sako

EDA

Feature	Takeaways
`Weather1`	▪ Majority of fatal crashes (~10%) occurred in unknown weather
`LightingDescription`	▪ Daylight most common for crashes overall. ▪ Highest percentage of fatal crashes (~3.2%) occurred on dark streets without functioning lights
`CollisionTypeDescription`	▪ Vehicle-pedestrian collisions had the highest percentage for actual injuries: Minor: 57.7% Serious: 8.6% Fatal: 11.4%







Oviya Adhan ▪ Nory Arroyo ▪ Caitlin Gainey ▪ Christine Sako

W207 SUMMER 2025

Methods

We use 4 different models to see which performs best at predicting our outcome variable

1

Logistic Regression (Baseline)

▪ Simple, interpretable, and efficient, this classifier can generate class probabilities and outputs the one with the highest score.

2

Neural Network

▪ Capturing complex relationships in the data, this model uses hidden layers to uncover patterns in the input features.

3

Random Forest

▪ Providing relatively interpretable outputs, as compared to Neural Networks, this model uses ensembles of decision trees to predict injury severity in car crashes.

4

XGBoost

▪ Working as an improvement over the Random Forest algorithm, this model sequentially produces new trees from the errors in the previous ensemble to learn patterns in the output features.

▪ Inclusion of regularization features and hyperparameters that make it more robust to overfitting.

Oviya Adhan ▪ Nory Arroyo ▪ Caitlin Gainey ▪ Christine Sako

Experiments

	Logistic Regression (Baseline)	Neural Network	Random Forest	XGBoost
Details	Scikit-learn Multiclass Logistic Regression classifier with balanced class weights	TensorFlow Keras sequential model with hidden layers and balanced class weights	Random Forest classifier with PCA, SMOTE (Synthetic Minority Oversampling Technique), 3-fold Stratified Cross, and hyperparameter tuning	XGBoost Classifier with balanced class weights and hyperparameter tuning
Parameters & Tuning	→ None	→ Learning Rate → Number of Epochs → Activations (ReLU / Tanh) → Optimizers (SGD / Adam) → Number of Layers (1 / 2) → Layer Sizes (256 / 128)	→ Randomized Grid Search (for number of estimators, splits, leaves, and tree depth)	→ Randomized Grid Search (for learning rate, number of trees, L1 & L2 Regularization)
Evaluation Metrics	→ Macro F1-score → 'Fatal' F1-score → Accuracy	→ Macro F1-score → 'Fatal' F1-score → Accuracy	→ Macro F1-score → 'Fatal' F1-score → Accuracy	→ Macro F1-score → 'Fatal' F1-score → Accuracy

Oviya Adhan ▪ Nory Arroyo ▪ Caitlin Gainey ▪ Christine Sako

W207 SUMMER 2025

Results

		1 - Multiclass Logistic Regression	2 - Neural Network	3 - Random Forest	4 - XGBoost
Accuracy		0.61	0.60	0.66 *	0.63
Macro-Averaged F1		0.37	0.37	0.42 *	0.40
	F1 - No Injury	0.75	0.73	0.76 *	0.75
	F1 - Minor	0.48	0.49	0.55 *	0.52
	F1 - Serious	0.17	0.17	0.19 *	0.19 *
	F1 - Fatal	0.09	0.08	0.17 *	0.13

Logistic Regression (Baseline)

Moderate overall accuracy

Accuracy: 61%

Moderate F1-Score overall for Fatal class

Macro F1-Score: 37%

Neural Network

Lowest overall accuracy

Accuracy: 60%

Lowest F1-Score overall for Fatal class

Macro F1-Score: 37%

Random Forest

Highest overall accuracy

Accuracy: 66%

Highest F1-Score overall for Fatal class

Macro F1-Score: 42%

XGBoost

Moderate overall accuracy

Accuracy: 63%

Moderate F1-Score overall for Fatal class

Macro F1-Score: 40%

Oviya Adhan ▪ Nory Arroyo ▪ Caitlin Gainey ▪ Christine Sako

* Highest score across models in metric category

Discussion

Neural Network | **Logistic Regression (Baseline)** | **XGBoost** | **Random Forest**

(Overall performance)

- Models complexity of noise in data
- Misses overarching patterns

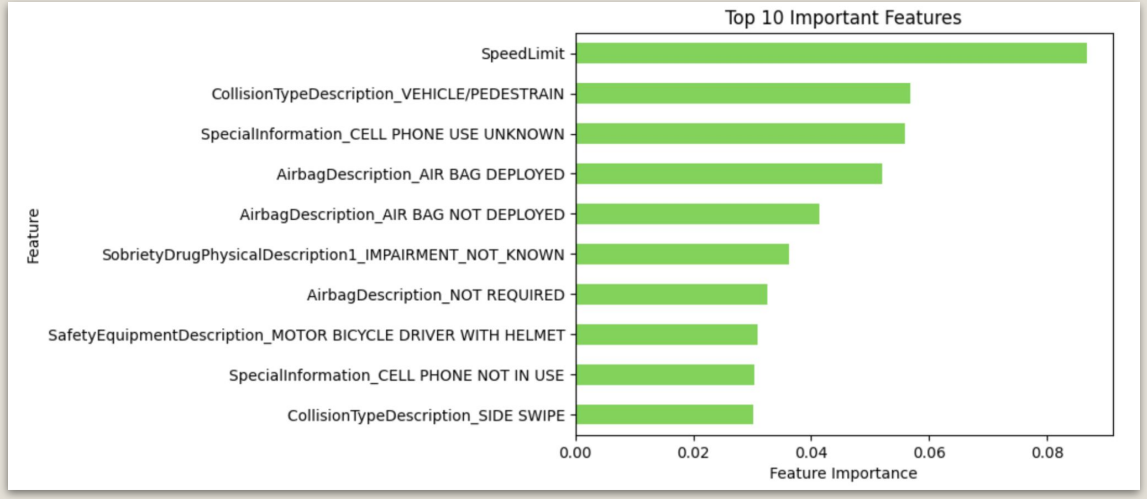
- Robust handling of collinearity compared to NN
- Assumes linearity

- Gradient boosting through sequential decision trees
- Prone to overfitting on noise

- Bagging - Parallel decision trees
- Averaging reduces overfitting risk

Oviya Adhan ▪ Nory Arroyo ▪ Caitlin Gainey ▪ Christine Sako

Discussion



Oviya Adhan ▪ Nory Arroyo ▪ Caitlin Gainey ▪ Christine Sako

Conclusion

Best Performing Model

- Random Forest Classifier

Challenges

- Merging Datasets
- Class Imbalance

Ideas for Improvement

- Binary Classification
- Reducing features
- More iterations of all models

Oviya Adhan ▪ Nory Arroyo ▪ Caitlin Gainey ▪ Christine Sako

Resources

1. Ahmed, S., Hossain, M. A., Ray, S. K., Bhuiyan, M. M. I., & Sabuj, S. R. (2023). A study on road accident prediction and contributing factors using explainable machine learning models: Analysis and performance. Transportation Research Interdisciplinary Perspectives, 19, Article 100814. https://doi.org/10.1016/j.trip.2023.100814

2. California Department of Technology, California Department of Motor Vehicles, & California Highway Patrol. (2025). California Crash Report System (CCRS). Data.ca.gov. Retrieved June 16, 2025, from https://data.ca.gov/dataset/ccrs

3. Pourroostaei Ardakani, S., Liang, X., Mengistu, K. T., So, R. S., Wei, X., He, B., & Cheshmehzangi, A. (2023). Road Car Accident Prediction Using a Machine-Learning-Enabled Data Analysis. Sustainability (Switzerland), 15(7), Article 5939. https://doi.org/10.3390/su15075939

4. Ryan, C. (2013). [Photograph of a car]. OJO Images. Getty Images. https://www.gettyimages.com/detail/photo/car-royalty-free-image/90201033

5. Wojcicki, A. (2018). [Illustration of a driverless car]. Science Photo Library. Getty Images. https://www.gettyimages.com/detail/illustration/driverless-car-royalty-free-illustration/956353340