Project Done by:

Manibalan Amaranathan

Srivastava Chinnasamy Kamaraj

Kousthubhee Krishna Kotte

Shravan Kumar Vamsi Dadi

Automobile Accident Severity Prediction

1. Introduction

Problem Statement

Automobile accidents are an unfortunate yet frequent occurrence on roads worldwide. The ability to predict the likelihood of an accident resulting in injury is crucial for optimizing emergency response, medical assistance, and resource allocation. This report explores how machine learning, specifically the Naïve Bayes classifier, can be leveraged to classify accidents as injured (True) or not injured (False) using only pre-accident conditions.

Objective

The primary goal of this study is to provide emergency responders and policymakers with a data-driven decision-making tool by analyzing weather conditions, road structure, traffic flow, and time of the accident. By understanding these factors, we can enhance accident response efficiency and reduce potential fatalities.

Dataset Overview

The dataset used in this study consists of **42,183 accident records from the United States (2001)**, each containing relevant environmental, road, and traffic-related data. Accident severity is classified into:

- **MAX_SEV_IR = 0** → No Injury
- **MAX_SEV_IR = 1** → Non-Fatal Injury
- **MAX_SEV_IR = 2** → Fatal Injury

2.Data Preprocessing

To ensure that our model works with relevant features, we performed extensive data preprocessing in RapidMiner before running the Naïve Bayes classifier. Below are the preprocessing steps taken:

2.1 Attribute Renaming

Renamed "ï»¿HOUR_I_R" to Hour for better readability.

2.2 Creating the Injury Column

A new Injury column was generated from MAX_SEV_IR:

MAX_SEV_IR = 1 or 2 → Injury = True (Injured)

MAX_SEV_IR = 0 → Injury = False (Not Injured)

2.3 Including Relevant Attributes

To focus only on pre-accident factors, we included 14 relevant attributes, as they provide useful insights for prediction:

Time of Accident: HOUR_I_R

Weather Conditions: WEATHER_R, LGTCON_I_R, SUR_CON

Road and Traffic Conditions: ALIGN_I, TRAF_CON_R, TRAF_WAY, SPD_LIM

Crash Circumstances: REL_RWY_R, RELJCT_I_R, VEH_INVL, WKDY_I_R, WRK_ZONE, INT_HWY

2.4 Filtering Data

Removed records where INT_HWY = 9 (as they were a small fraction, ~200 records).

2.5 Converting Numerical Variables to Binomial and Polynomial

Binomial Conversion (0/1 → Categorical):

HOUR_I_R, WRK_ZONE, WKDY_I_R, INT_HWY, RELJCT_I_R, REL_RWY_R, Injury

Polynomial Conversion (Categorical with Multiple Values):

ALIGN_I, LGTCON_I_R, PROFIL_I_R, SPD_LIM, SUR_CON, TRAF_CON_R, TRAF_WAY, WEATHER_R

2.6 Setting Target Label

Used Set Role Operator to define Injury as the target variable.

3. Initial Prediction Without Additional Information

Question: If an accident has just been reported and no additional details are available, should we predict INJURY = True or False?

Analysis:

- The safest assumption is to classify new accidents as **INJURY = True** in the absence of additional information.
- **50.9%** of accidents in the dataset resulted in an injury, making it the more frequent outcome.
- A **safety-first approach** ensures emergency responders prioritize medical assistance.
- False negatives (undetected injuries) are far riskier than false positives (overestimating injuries).

Business Context:

Predicting injury cases correctly allows **emergency responders to allocate resources efficiently**. Misclassifying an actual injury as non-injury can lead to delays in medical assistance, increasing the risk of fatalities.

4. Naïve Bayes Classifier Implementation

Model Training

The **Naïve Bayes classifier** was selected because:

- It is computationally efficient for large datasets.
- It provides probabilistic interpretations for predictions.

Results & Evaluation

Confusion Matrix (Threshold = 0.5, Balanced Approach)

Actual	Predicted Not Injured	Predicted Injured
Not Injured	46.00%	54.00%
Injured	39.74%	60.26%

accuracy: 53.26%

	true Injured	true NotInjured	class precision
pred. Injured	5169	4473	53.61%
pred. NotInjured	3409	3811	52.78%
class recall	60.26%	46.00%	

Confusion Matrix (Threshold = 0.7, Prioritizing Injury Detection)

Actual	Predicted Not Injured	Predicted Injured
Not Injured	2.37%	97.63%
Injured	1.18%	98.82%

accuracy: 51.44%

	true Injured	true NotInjured	class precision
pred. Injured	8477	8088	51.17%
pred. NotInjured	101	196	65.99%
class recall	98.82%	2.37%	

The model correctly detects most injuries (98.82%) but misclassifies many non-injuries as injuries.

Model Performance Insights

- **Accuracy:** 51.44%
- **Precision (Injured):** 51.17%
- **Recall (Injured):** 98.82%
- **Specificity:** 2.37%

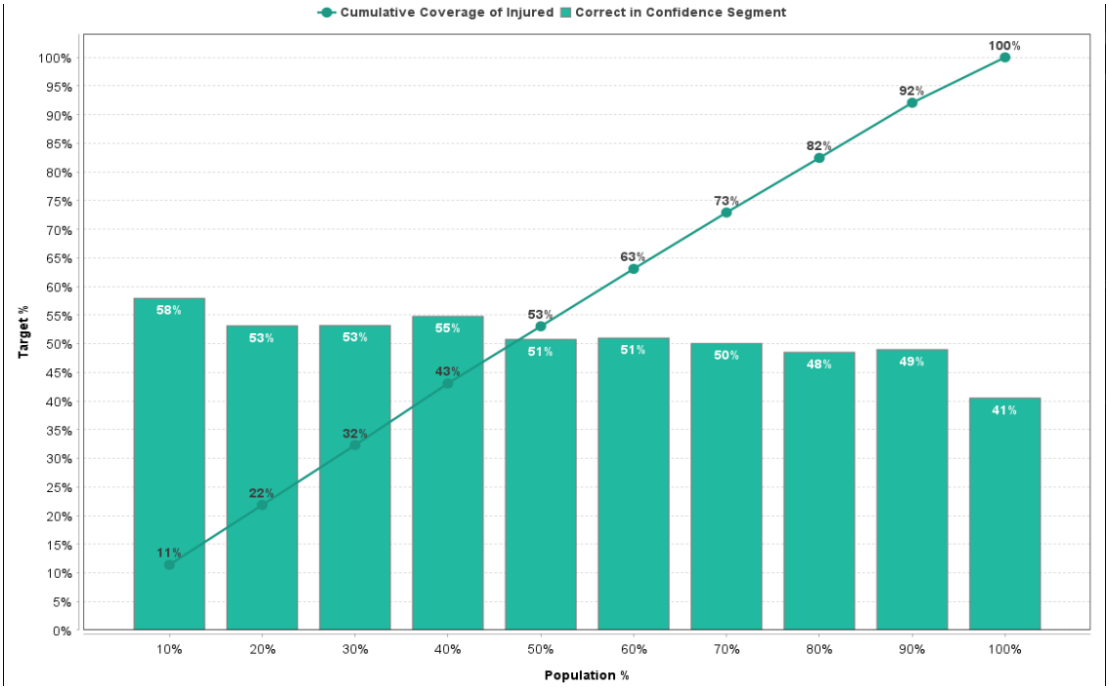
Final Conclusion:

- Injury detection is the priority → Threshold = 0.7 is better.
- For paramedic response, a high recall for injuries (0.7 threshold) is preferred, even if it means more false positives.

Lift Chart Analysis

The lift chart below demonstrates the cumulative percentage of injured cases correctly identified at different population percentiles:

- **Top 10% correctly classifies 58% of injuries**.
- **Model maintains above 50% accuracy up to 60% of the population**.
- **Beyond 70%, classification confidence drops below 50%**.



Error Rate Analysis

Threshold	Error Rate
0.5	46.74%
0.7	48.56%

- **Higher recall (injury detection) comes at the cost of increased false positives.**

Performance vs. Naïve Rule

Model	Accuracy
Naïve Rule (All Injured)	50.90%
Naïve Bayes (Threshold 0.5)	53.26%
Naïve Bayes (Threshold 0.7)	51.44%

- Threshold = 0.5 → 4.64% improvement over naïve rule
- Threshold = 0.7 → 1.06% improvement over naïve rule
- At 0.5, the model improves predictions by 4.64%, making it the better improvement choice over naïve classification.
- **At 0.7, the improvement is small (1.06%), but the model ensures almost all injuries are detected (98.82% recall).**

Conditional Probability Issue (P(INJURY = false | SPD_LIM = 5) = 0)

- The dataset **lacks "Not Injured" cases at SPD_LIM = 5**.
- **Naïve Bayes assigns zero probability** when no training samples exist.
- **Solution**: Apply **Laplace Smoothing** to prevent zero probabilities.

6. Conclusion

- Predict "**Injured**" as the default classification if no data is available.
- **The model effectively predicts injury severity** but can be improved.
- **Threshold 0.5** balances accuracy, while **0.7 prioritizes emergency response**.
- **Threshold 0.7 captures almost all injuries (98.82%),** making it more useful for paramedic dispatching.
- **Laplace Smoothing** addresses the zero probability issue.