MSIS 5633 - Predictive Analytics Technologies
Section - In class

Term Project: Team 15
Injury Severity Risk Factors in Automobile Crashes

Due Date
5th May 2024

By

Karthik Suman Bathini (A20399645)

Sunanda Sen (A20400681)

Chinmayi Rane (A20386634)

Dr. Delen Dursun, Course Instructor

Team Members:

Karthik Bathini
Graduate Student in Management Information Systems

Sunanda Sen
Graduate Student in Business Analytics & Data Science

Chinmayi Rane
Graduate Student in Management Information Systems

Table of Contents

1 Executive Summary 4

2 Business Understanding 5

3 Data Understanding

3.1 Overview 5

3.2 Initial Data Collection 6

3.3 Describing Data 6

3.4 Exploring Data 6

4 Data Preparation & Preprocessing

4.1 Overview 7

4.2 Data homogenization & constructing the dependent variable 7

4.3 Data Transformation 8

4.4 Missing Value Imputation & Variable Selection 9

5 Data Modelling

5.1 Number Based Models 11

5.1.1 Neural Networks 12

5.1.2 Logistic Regression 13

5.2 Set Based Models 14

5.2.1 Naïve Bayes 14

5.2.2 Random Forest 15

5.2.3 Gradient Boosted Trees 16

5.2.4 Decision Tree 17

6 Testing & Evaluation

6.1 Variable Importance Analysis 19

6.2 Model Evaluation - ROC Curve 20

6.3 Model Summary 21

7 Deployment

8 Conclusion & Future Scope

8.1 Conclusion 22

8.2 Future Scope 23

9 References 24

1 Executive Summary

The predictive modeling project by Team 15 focuses on identifying risk factors influencing injury severity in automobile crashes to enhance road safety and policy-making. Our comprehensive analysis of a dataset from the Crash Report Sampling System (CRSS) facilitated by predictive modeling tools in KNIME aims to pinpoint significant factors contributing to crash injuries. This insight enables stakeholders such as government agencies, automobile manufacturers, and public health organizations to implement effective safety interventions.

Our approach includes multiple predictive modeling techniques, such as decision trees, random forests, neural networks, and logistic regression, to evaluate and forecast the severity of injuries. The Random Forest model exhibited the most robust performance, providing valuable insights into how various factors like vehicle age, accident location, and safety features influence injury outcomes. The models underline the importance of enhancements in vehicle safety and the promotion of safer driving behaviors as significant means to mitigate crash severity. Key findings suggest that factors such as the ejection of passengers, vehicle towing status, and the activation of airbags play crucial roles in the severity of injuries.

As we advance, it's crucial to extend our data sources to include emerging technologies in vehicle safety, like autonomous and semi-autonomous systems, and to incorporate broader demographic and geographical diversity. Expanding the dataset will allow for a better understanding of different variables that contribute to injury severity in diverse environments and scenarios. This comprehensive data integration will help tailor safety interventions more accurately to specific contexts and populations, enhancing the effectiveness of predictive models in real-world applications.

Moreover, there is substantial potential for these models to inform national policy and industry standards. By closely working with policymakers and regulatory bodies, the predictive insights can be translated into legislative actions and safety standards that preemptively address risk factors identified through our analysis. This proactive approach not only enhances vehicular safety but also fosters a culture of safety that permeates all levels of driving and road use. Continued collaboration with stakeholders will ensure that the recommendations keep pace with technological advancements and changes in driving behavior, thereby sustaining the relevance and impact of our initiatives on road safety.

2 Business Understanding

Cars are integral to American transportation, yet they pose substantial risks, given the underdeveloped public transportation infrastructure. The prevalence of personal automobiles contributes to a high incidence of car crashes, leading to numerous injuries and fatalities annually. According to the National Highway Traffic Safety Administration (NHTSA), 42,915 fatalities were reported in car crashes in 2021. Despite this, complete elimination of car crashes remains improbable as long as driving persists. However, advancements in automobile safety features have been significant, with innovations such as tempered glass and automated braking systems aiming to mitigate injury severity. Yet, there remains a pressing need for further enhancements to ensure road safety.

The NHTSA collects extensive data on U.S. car crashes, providing invaluable insights into the determinants of crashes and the severity of resulting injuries. This study seeks to address the crucial aspect of injury severity prediction, recognizing that car crashes can range from minor incidents to catastrophic events resulting in coma, paralysis, or death. Employing various prediction algorithms like decision trees, naive Bayes, artificial neural networks, random forest, and logistic regression, this research aims to identify the primary determinants influencing injury severity. Input parameters mainly encompass attributes related to the vehicle, the driver, and the road. By discerning the variables with the most significant impact on injury severity, stakeholders such as civil engineers, governmental bodies, and car manufacturers can formulate targeted interventions to improve road safety and minimize injury severity in car crashes.

3 Data Understanding

3.1 Overview

The National Highway Traffic Safety Administration (NHTSA) utilizes data from various sources for its critical work. This analysis will focus specifically on a dataset obtained from the Crash Report Sampling System (CRSS). A thorough understanding of the data is paramount, as proper analysis at this stage minimizes potential issues in subsequent phases. CRSS leverages police reports on crashes involving automobiles, road users, and cyclists to create a representative sample of these incidents. Data for this system is culled from a comprehensive examination of over six million crashes.

The CRSS offers a valuable resource for more than just estimating overall crash statistics. It also plays a crucial role in identifying traffic safety hazards, gathering information on driver behavior, and forming the foundation for highway safety regulations and public awareness campaigns. However, it's important to acknowledge the limitation of the CRSS data. Since it solely relies on police-reported incidents, there is a possibility of missing data, particularly for crashes with minimal property damage that may not be reported to law enforcement.

3.2 Initial Data Collection

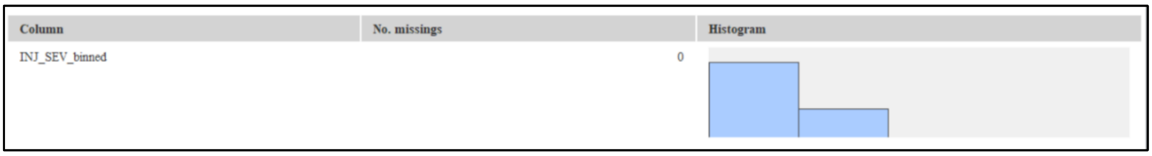
The initial data collection for the CRSS involves gathering detailed reports from police-reported motor vehicle crashes across the United States. This collection captures a broad spectrum of accidents, from minor property damage incidents to fatal crashes. Each report includes specific information on the vehicles involved, the crash environment, and the demographics of the individuals involved. This stage ensures that the data represents a nationally significant sample, reflecting varied geographic, demographic, and incident-related factors. The process involves trained personnel systematically sampling and coding tens of thousands of crash reports each year from selected sites across the country.

3.3 Describing Data

For our analysis, we were provided with four distinct SAS data files: the accident dataset, vehicle dataset, person dataset, and distract dataset. The accident dataset offers insights into crash characteristics and environmental conditions prevailing at the time of the incident, featuring variables such as Alcohol Involved in Crash, Atmospheric Conditions, and Manner of Collision. Conversely, the vehicle dataset provides information on motor vehicles and their drivers involved in the crash, including variables like Vehicle Model Year, Movement Prior to Critical Event, and Driver Drinking in Vehicle. The person dataset encompasses details concerning all individuals involved in the crash, spanning motorists and non-motorists alike. Lastly, the distract dataset furnishes data pertaining to driver distractions, contributing valuable insights into factors influencing road safety and accident occurrences.

3.4 Exploring Data

After completing data preprocessing, the final dataset was examined using the Data Explorer node in the KNIME tool. Of particular interest is the "Dependent Variable," which is the primary attribute requiring careful scrutiny. The output generated by the Data Explorer node offers valuable insights into this attribute.



Column	No. missings	Histogram
INJ_SEV_binned	0	

Figure 1: Frequency Distribution of our target variable

From the above diagram we can see the distribution of injury severity, as binned. The dataset comprises 4,423 records categorized under high injury, contrasted with 18,936 records classified as low injury, equating to approximately a 1 to 5 ratio for low and high injury instances. This disparity highlights a significant imbalance in the data, necessitating corrective measures before model implementation to ensure balanced and reliable predictions.

4 Data Preparation & Preprocessing

4.1 Overview

The data preprocessing plan involves several key steps to ensure the dataset is adequately prepared for predictive modeling and analysis using KNIME. Firstly, Data Exploration is conducted utilizing KNIME's Data Explorer node to comprehensively visualize and understand the dataset. Summary statistics, histograms, box plots, and correlation matrices are generated to gain insights into data distributions and relationships.

4.2 Data homogenization & constructing the dependent variable

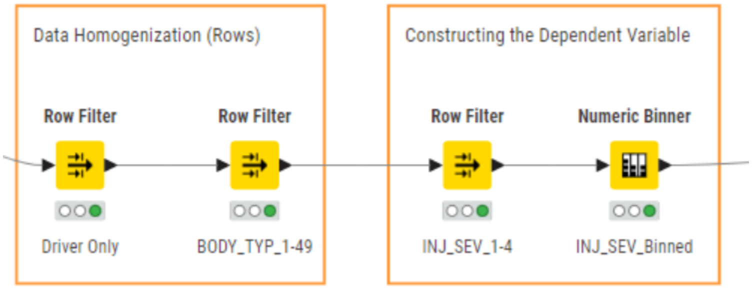


Figure 2: Data homogenization & constructing the dependent variable.

The data preprocessing began by importing four datasets – Accidents, Vehicle, Persons, and Distract utilizing the SAS7BDAT Reader node within KNIME. These datasets were subsequently merged using the joiner node in KNIME. The accidents and vehicle datasets were merged using an inner join. Subsequently, the person and distract datasets were each merged with the resulting dataset using left outer joins. Following the merge, data homogenization was performed through row filter nodes for two primary objectives. Firstly, to narrow the analysis scope to driver-centric factors, the SEAT_POS variable was filtered to encompass solely the driver's seat positions. Secondly, for a focused examination on lighter vehicles, the Body_TYP variable was refined to exclude buses, medium, and heavy trucks.

In examining our dependent variable, Injury Severity, we identified eight attributes delineating the severity of injuries sustained during crashes, with attribute 0 representing no injury and attributes 5, 6, and 9 signifying cases of unknown severity. Focusing on predicting injury severity, we filtered out attributes 0, 5, 6, and 9. Subsequently, utilizing KNIME's numeric binner node, we binned injury severity into two categories: low (attributes 1-2.5) and high (attributes 2.5-4). This streamlined approach facilitates a targeted analysis, aiding in the identification of patterns and insights pertinent to injury severity within our dataset.

4.3 Data Transformation

All the variables used in our analysis was feature engineered to make them more suitable for modeling and to improve the model's predictive power. Nodes such as the Rule Engine, Math Formula and Numeric Binner were used for this purpose**.** In our project, the Rule Engine node was crucial for transforming and categorizing data into formats better suited for analysis, enhancing both the interpretability and effectiveness of the predictive models. For example, it was used to convert the RELJCT1 data into a binary format, simplifying the input to indicate presence or absence within a junction, which is vital for assessing the impact of junction proximity on injury severity. Similarly, for the TOWED variable, the Rule Engine differentiated vehicles towed due to disabling damage from those towed for other reasons, a distinction that directly correlates with the severity of the crash and potential injuries.

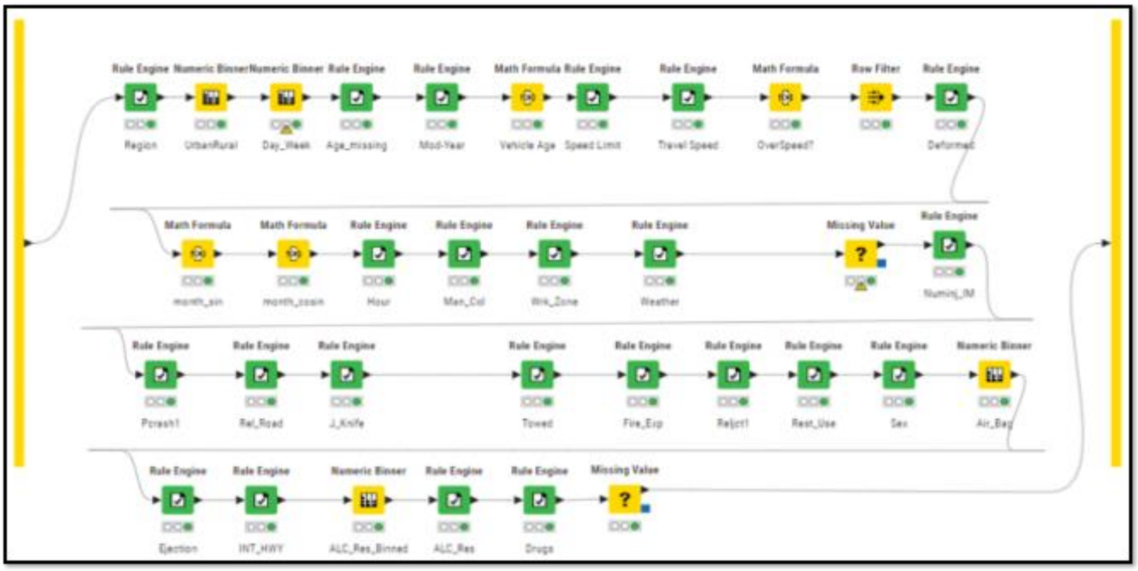


Figure 3: Data Transformation using Feature Engineering

The node also facilitated the grouping of light conditions (LGT_COND) into broader categories such as "Daylight" and "Dark - Lighted," which helps the models more accurately evaluate the influence of lighting on crash outcomes. Additionally, it categorized vehicle age into "New" and

"Old" to reflect differences in safety features that affect injury severity, and it marked ejection statuses as "Ejected" or "Not Ejected," which are critical for predicting injury severity. Through these transformations, the Rule Engine node ensured that each variable was optimally prepared to enhance the predictive models' accuracy and provide actionable insights.

The Math Formula node was crucial for applying numerical transformations and calculations to refine the dataset for predictive modeling. It was used to calculate vehicle age from the manufacturing year. Additionally, it was capable of creating interaction terms, such as between Overspeed to capture complex interactions and enhance model accuracy. This node was also used to transform the "Month" variable to capture seasonal effects effectively, particularly useful for analyzing time-sensitive data like traffic accidents. A common approach is using sinusoidal transformations, such as sin(2*pi*$Month$/12) and cos(2*pi*$Month$/12), which help model the cyclical nature of months, ensuring that December and January are recognized as sequentially close, unlike linear numerical encoding. This transformation is particularly beneficial for datasets where seasonal patterns, such as weather variations, significantly impact the outcomes, thereby allowing the models to accommodate the recurring annual trends.

In the project, the Numeric Binner node in KNIME was instrumental in transforming numerical data into categorized bins for enhanced interpretability and analysis. For instance, the "ALC_RES" variable, indicating blood alcohol concentration, was binned into categories such as 'No Alcohol', 'Low', 'Medium', and 'High' using the Numeric Binner. This categorization allowed the models to more effectively assess the impact of varying alcohol levels on injury severity. Additionally, the "DAY_WEEK" variable, representing days of the week, could be categorized into 'Weekday' and 'Weekend', enabling the analysis to discern patterns in crash occurrences and severities based on the day of the week. These transformations help in simplifying the analysis and improving the accuracy of predictive modeling by addressing the nuances in data distribution.

4.4 Missing Value Imputation & Variable Selection

During preprocessing, missing values were handled by imputing numeric variables with the median and categorical variables with "Missing", while specific variables such as WRK_ZONE, Travel Speed, and Speed Limit were imputed with rounded means or medians as appropriate. Variable selection involved utilizing the column filter node to retain 32 pertinent variables for predictive modeling. Additionally, the color manager node was utilized to assign colors to the target variable for enhanced visualization. To maintain data consistency without compromising information integrity, a normalizer node was employed to adjust values to a common scale. These preprocessing steps ensure the dataset is optimized for subsequent analysis and model development. The 32 variables in our final prediction model are:

- Hour
- MAN_COLL
- RELJCT1
- REL_ROAD
- WRK_ZONE
- LGT_COND
- WEATHER
- INT_HWY
- NUM_INJ
- BODY_TYP
- J_KNIFE
- TOWED
- FIRE_EXP
- PCRASH1
- SEX
- SEAT_POS
- REST_USE
- EJECTION
- ALC_RES
- DRUGS
- Region
- Urban city
- DAY_WEEK
- Age
- MOD_Year
- VEH_Age
- VSPD_Lim
- TRAV_SP
- Overspeed
- Deformed
- Month
- AIR_BAG

5 Data Modelling

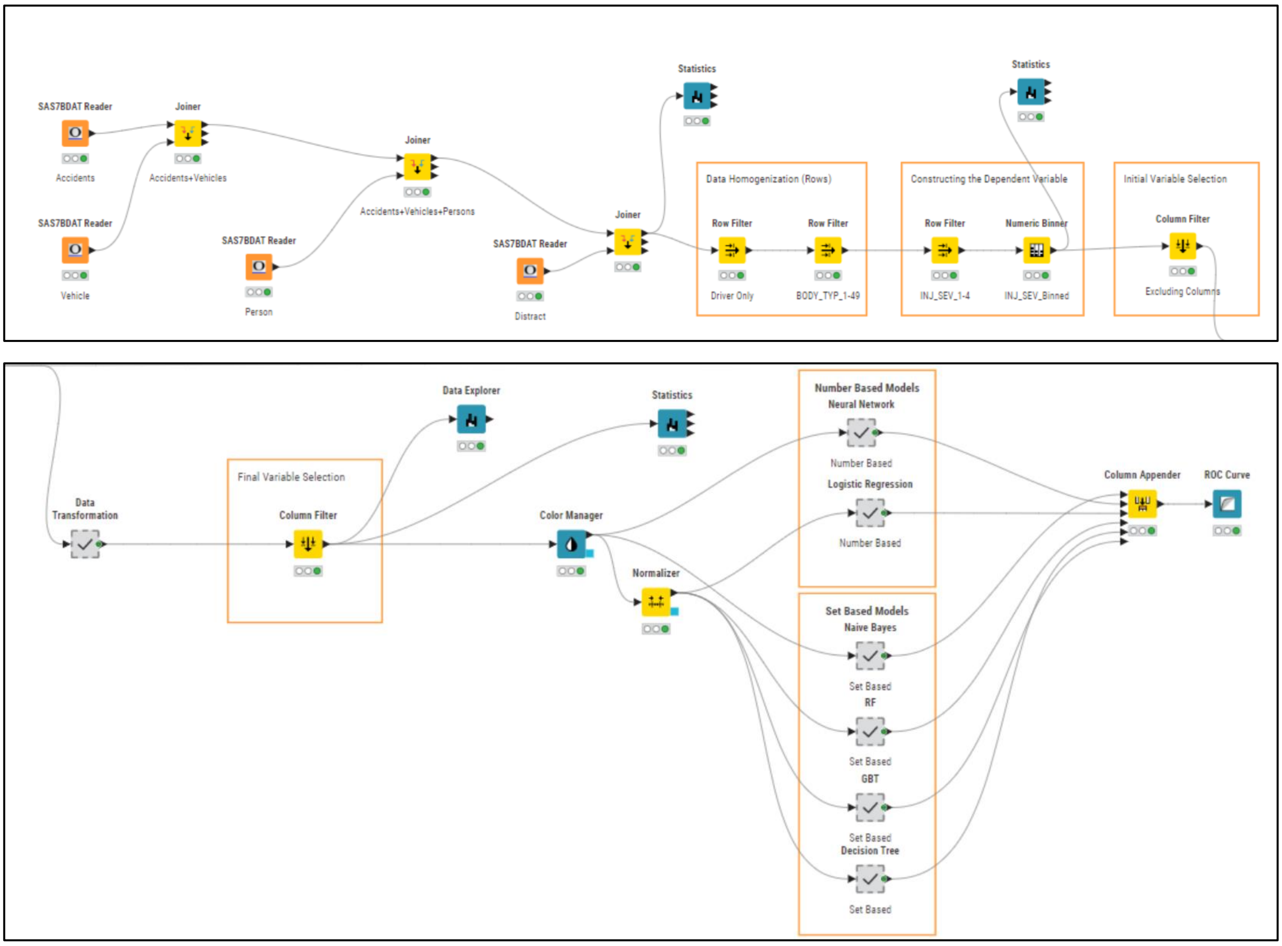


Figure 4: KNIME workflow

5.1 Number Based Models

Number-based prediction models are statistical or computational models that use numerical data to make predictions or forecasts about future events or outcomes. These models rely on mathematical algorithms to analyze historical data, identify patterns or trends, and generate predictions based on these patterns.

5.1.1 Neural Networks

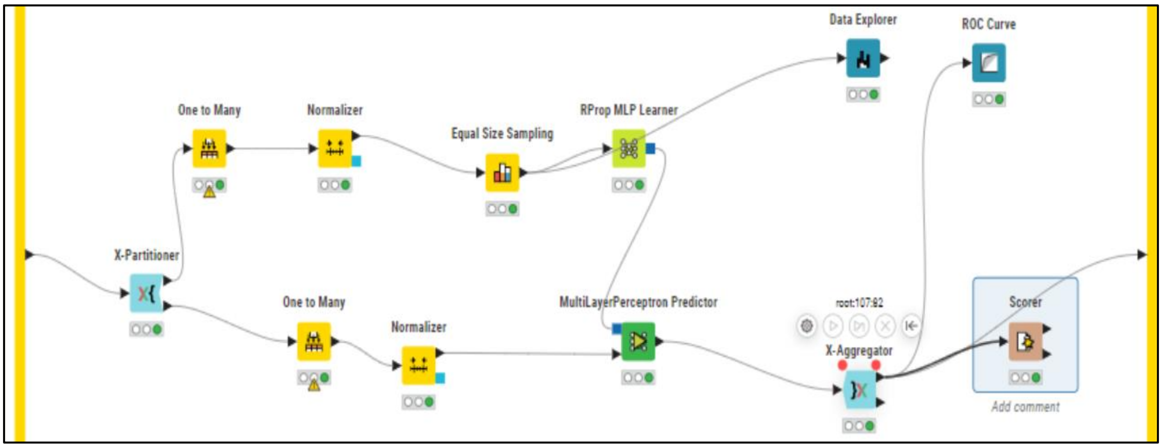
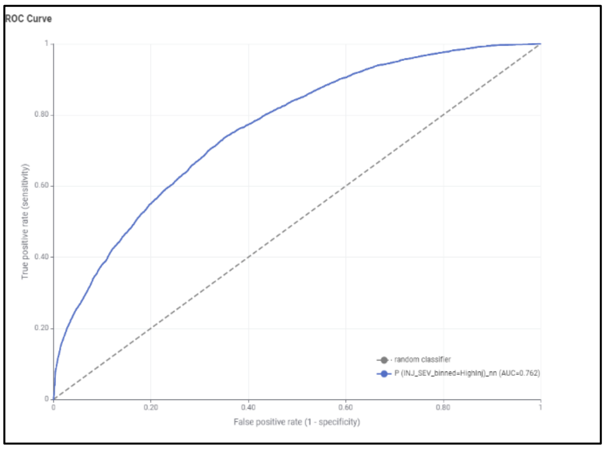
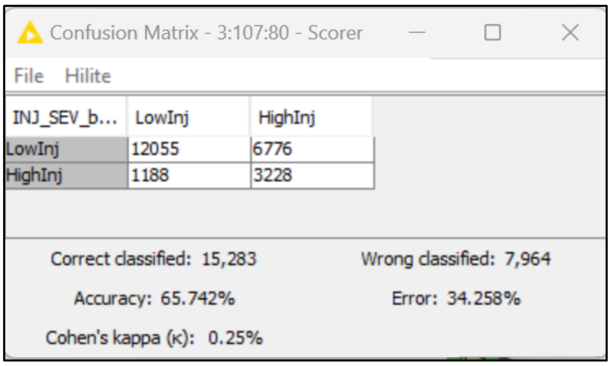


Figure 5: Neural Networks

The Artificial Neural Network (ANN) is a computational model inspired by biological neural networks, aimed at replicating learning and decision-making processes. During training, categorical variables are transformed using one-to-many conversion techniques, while numerical features are scaled using normalizers to ensure uniformity across the dataset. The training process includes employing k-fold cross-validation, which aids in evaluating model performance and mitigating overfitting by partitioning the dataset into training and validation sets multiple times. Evaluation metrics such as accuracy, sensitivity, and specificity are utilized to assess the ANN's predictive capabilities, with k-fold cross-validation providing reliable estimates of its performance across different data subsets. These methods collectively enhance the ANN's ability to analyze complex datasets and make accurate predictions in various domains.





INJ_SEV_b...	LowInj	HighInj
LowInj	12055	6776
HighInj	1188	3228

Correct classified: 15,283 Wrong classified: 7,964

Accuracy: 65.742% Error: 34.258%

Cohen's kappa (κ): 0.25%

Figure 6: ROC Curve for Neural Networks

Figure 7: Confusion Matrix

5.1.2 Logistic Regression

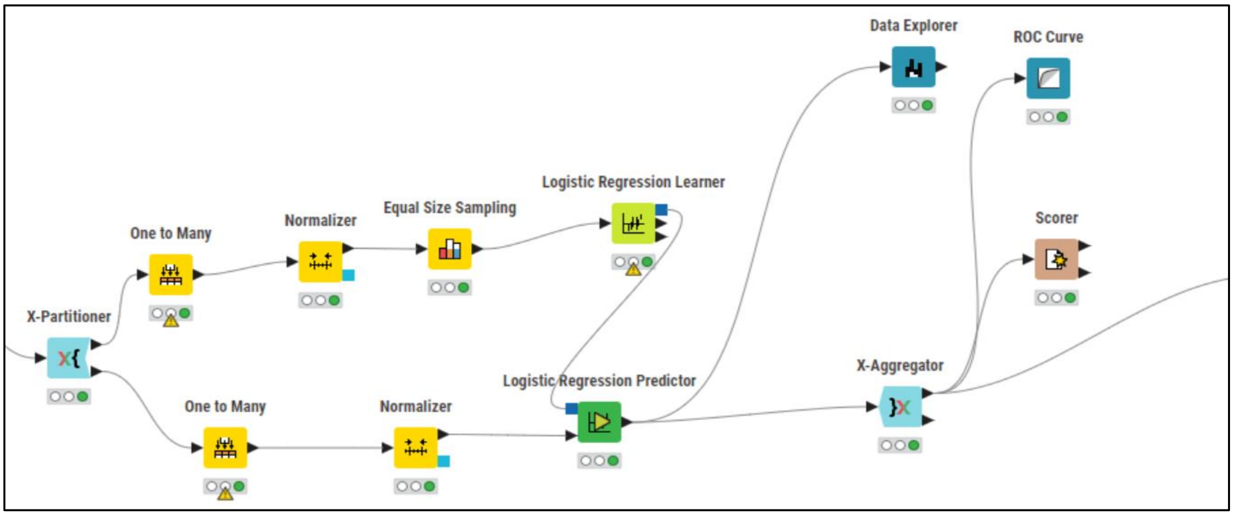
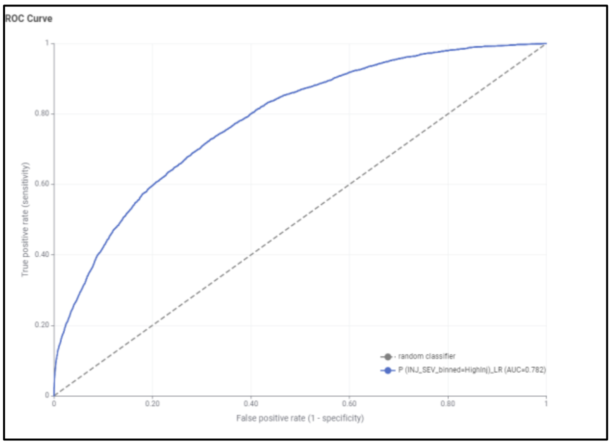
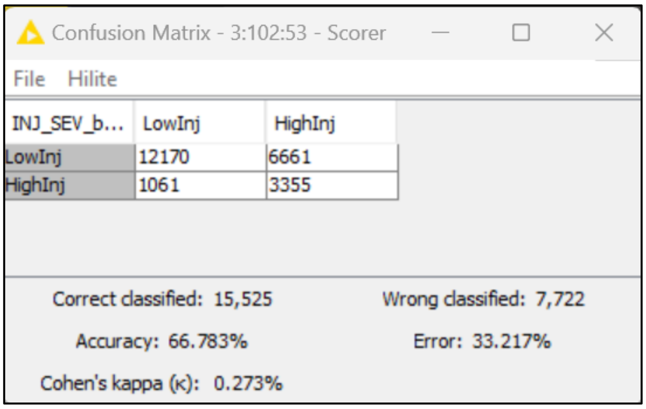


Figure 8: Logistic Regression

Logistic regression, a statistical method for predicting binary outcomes based on independent variables, was employed in this study due to its suitability for binary-dependent-variable scenarios. Data preprocessing included normalization using the normalizer node for consistent feature scaling. The x-partitioner node aided in data partitioning, with categorical variables handled by the one-to-many node. Class representation within the training data was balanced using equal size sampling. The logistic regression learner node facilitated model training, and predictions were generated using the predictor node, applying the same preprocessing steps as the training data. Results were aggregated using the x-aggregator node for subsequent analysis and evaluation purposes.





INJ_SEV_b...	LowInj	HighInj
LowInj	12170	6661
HighInj	1061	3355

Correct classified: 15,525 Wrong classified: 7,722

Accuracy: 66.783% Error: 33.217%

Cohen's kappa (κ): 0.273%

Figure 9: ROC Curve for Logistic Regression

Figure 10: Confusion Matrix

5.2 Set Based Models

Set-based models are a type of computational or mathematical model that operates on sets of data rather than individual data points. Instead of focusing on individual data instances, set-based models consider groups or collections of data elements as a whole. These models are particularly useful when dealing with data that naturally exists in groups or sets.

5.2.1 Naïve Bayes

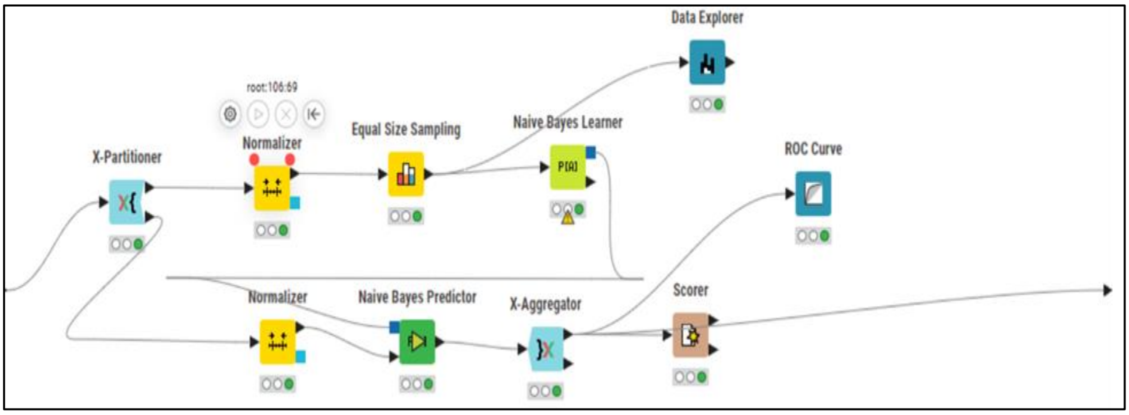
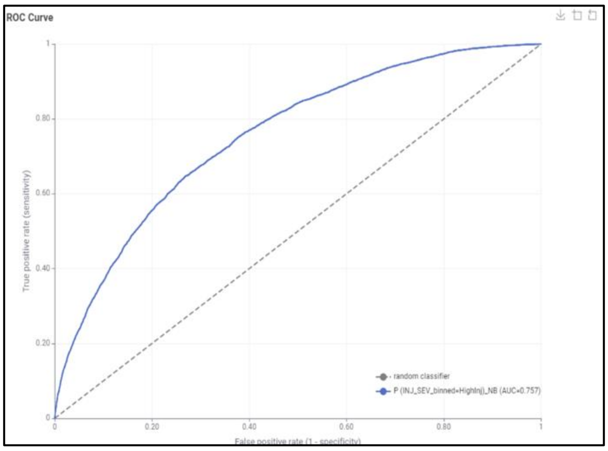
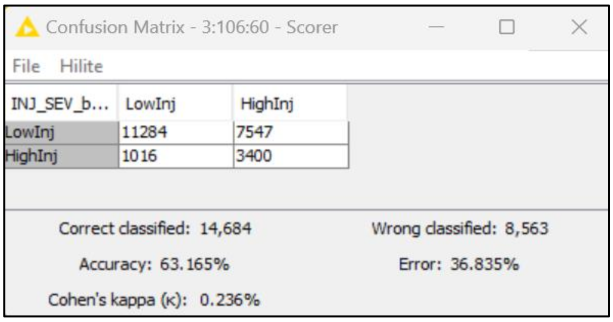


Figure 11: Naive Bayes

Naive Bayes is a straightforward and efficient probabilistic classifier based on Bayes' theorem, assuming feature independence for classification tasks. Data preprocessing, including categorical variable handling via the color manager node, ensured consistency. Robust model evaluation was achieved through k-fold cross-validation using the x-partitioner and x-aggregator nodes. Normalization via the normalizer node maintained uniform feature scaling. Additionally, equal size sampling balanced class representation in the training data, improving model generalization. These key features collectively optimized the Naive Bayes model's performance for accurate classification outcomes.





INJ_SEV_b...	LowInj	HighInj
LowInj	11284	7547
HighInj	1016	3400

Correct classified: 14,684

Wrong classified: 8,563

Accuracy: 63.165%

Error: 36.835%

Cohen's kappa (κ): 0.236%

Figure 12: ROC Curve for Naive Bayes

Figure13: Confusion Matrix

5.2.2 Random Forest

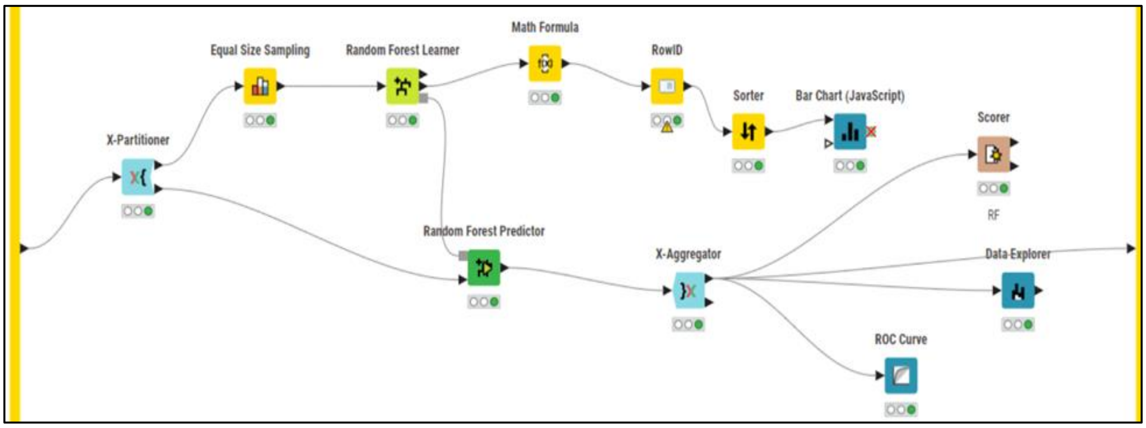
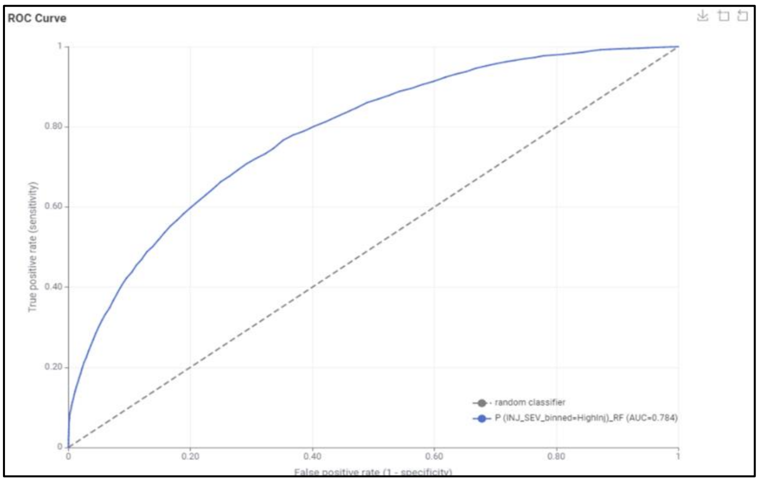
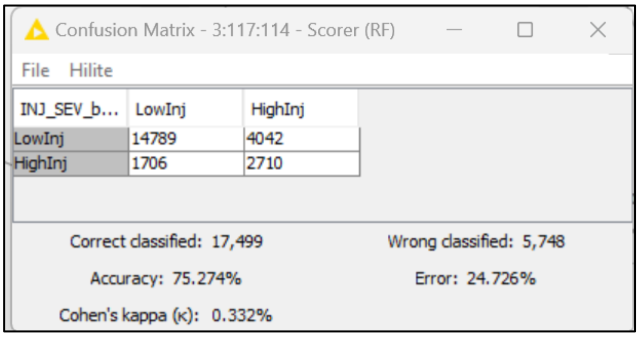


Figure 14: Random Forest

Random forest is an ensemble learning method that enhances predictive accuracy and reduces overfitting by combining predictions from multiple decision trees. Key features include data normalization for consistent numerical feature scaling, k-fold cross-validation with stratified sampling for robust performance evaluation, and equal size sampling with a random seed for balanced class representation during training. Efficiency in model training was improved using the "use exact sampling" option. Additionally, variable importance analysis via the random forest bar chart node provided insights into feature significance for injury severity prediction. These features collectively optimize the random forest model's performance and interpretability.




Confusion Matrix - 3:117:114 - Scorer (RF)

INJ_SEV_b...	LowInj	HighInj
LowInj	14789	4042
HighInj	1706	2710

Correct classified: 17,499 Wrong classified: 5,748

Accuracy: 75.274% Error: 24.726%

Cohen's kappa (κ): 0.332%

Figure 15: ROC Curve for Random Forest

Figure 16: Confusion Matrix

5.2.3 Gradient Boosted Trees

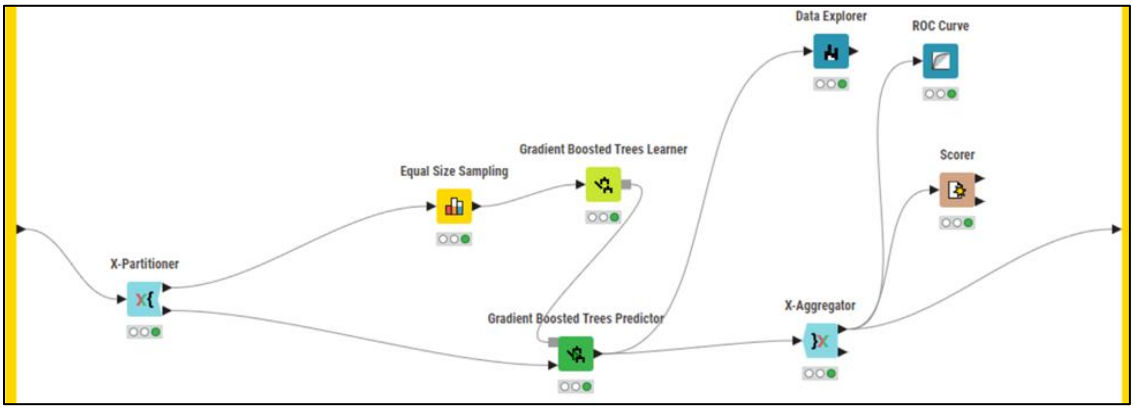
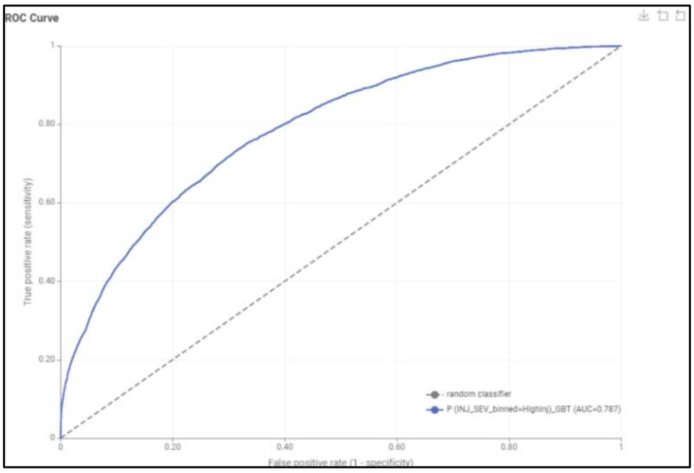
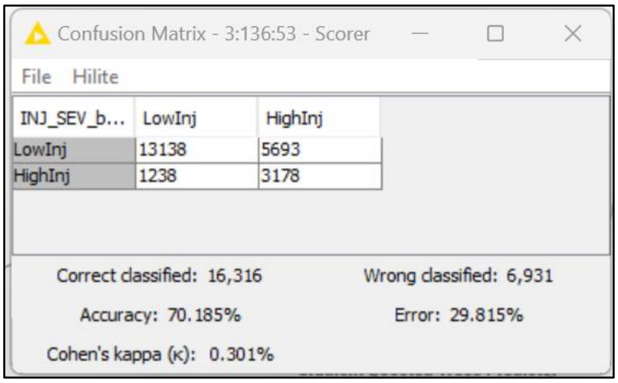


Figure 17: Gradient Boosted Tree

Gradient Boosting Trees (GBT) is an ensemble learning technique for classification and regression tasks, building an ensemble of decision trees sequentially to learn from predecessors' errors. Key features include sequential tree construction focusing on error reduction, using an equal sampling node with a static seed for result consistency, and employing k-fold cross-validation with INJ_SEV_BINNED as the target column for robust model performance assessment across varied data subsets. These strategies collectively optimize GBT's predictive power and generalization ability in complex data scenarios.





INJ_SEV_b...	LowInj	HighInj
LowInj	13138	5693
HighInj	1238	3178

Correct classified: 16,316

Wrong classified: 6,931

Accuracy: 70.185%

Error: 29.815%

Cohen's kappa (κ): 0.301%

Figure 18: ROC Curve for Gradient Boosted Tree

Figure 19: Confusion Matrix

5.2.4 Decision Trees

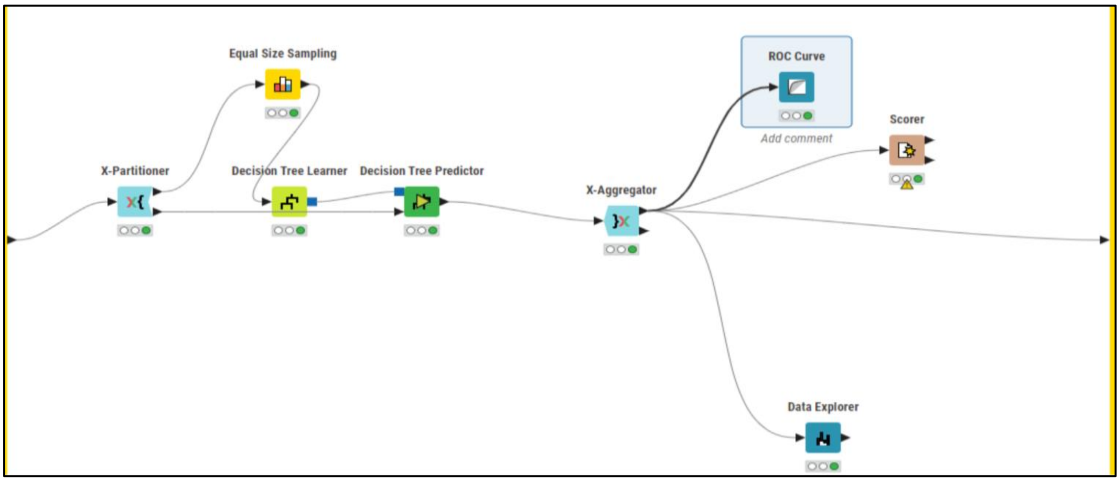
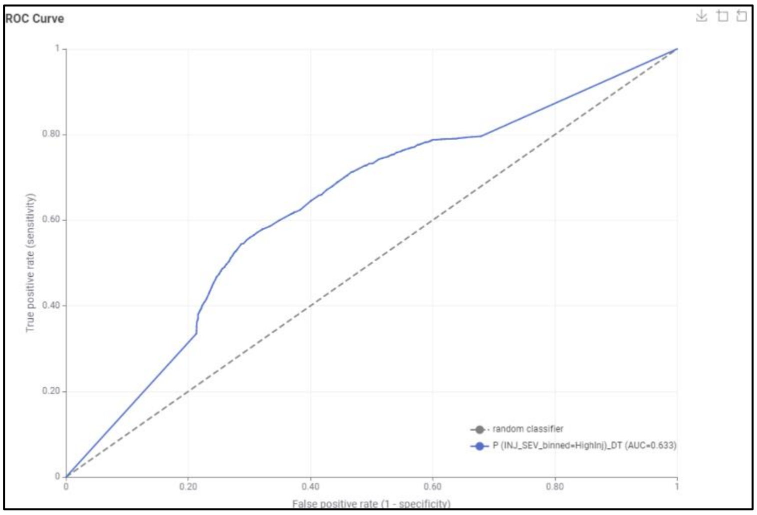
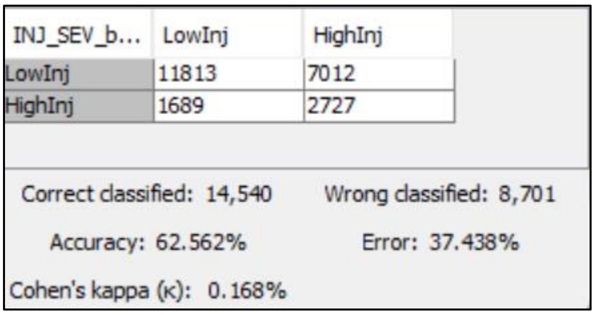


Figure 20: Decision Tree

A decision tree algorithm uses a tree-like structure to classify data based on features, leading to predicted outcomes like injury severity. Key features in our project included sourcing data from the normalizer node for consistent numerical scaling, employing equal sampling for class balance, and using a binary nominal split criterion for efficient categorical classification. To prevent overfitting, we used k-fold cross-validation with a random seed for reproducible results and stratified sampling to maintain class proportions in each fold, ensuring reliable model evaluations. These strategies optimize decision tree modeling for accurate and robust predictions in our dataset.





INJ_SEV_b...	LowInj	HighInj
LowInj	11813	7012
HighInj	1689	2727

Correct classified: 14,540 Wrong classified: 8,701

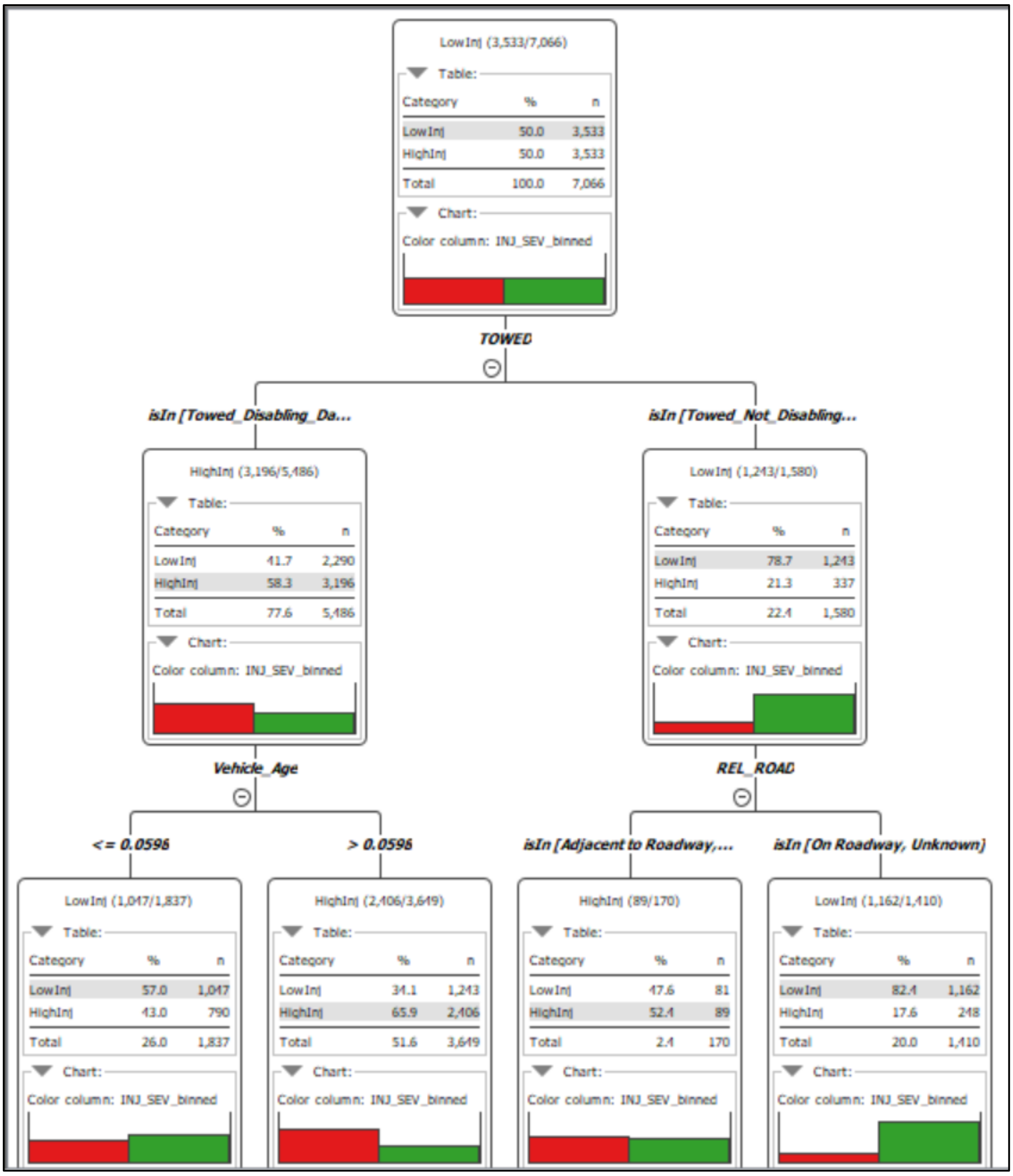
Accuracy: 62.562% Error: 37.438%

Cohen's kappa (κ): 0.168%

Figure 21: ROC Curve for Decision Tree

Figure 22: Confusion Matrix

From the decision tree view below, we see that vehicle age significantly influences injury severity, with newer vehicles associated with lower injuries. For older vehicles, the accident location becomes crucial, especially on major roads or adjacent areas, leading to higher injury rates. Additionally, the tow status of vehicles in accidents on non-major roads provides insights, with towed vehicles more likely to have lower injuries. These factors collectively highlight the nuanced relationship between vehicle age, accident location, and tow status in predicting injury severity outcomes.


TOWED
isIn [Towed_Disabling_Da...
isIn [Towed_Not_Disabling...
Vehicle_Age
<= 0.0596
> 0.0596
REL_ROAD
isIn [Adjacent to Roadway,...
isIn [On Roadway, Unknown]
Table:
Chart:
Color column: INJ_SEV_binned

LowInj (3,533/7,066)

Category	%	n
LowInj	50.0	3,533
HighInj	50.0	3,533
Total	100.0	7,066

HighInj (3,196/5,486)

Category	%	n
LowInj	41.7	2,290
HighInj	58.3	3,196
Total	77.6	5,486

LowInj (1,243/1,580)

Category	%	n
LowInj	78.7	1,243
HighInj	21.3	337
Total	22.4	1,580

LowInj (1,047/1,837)

Category	%	n
LowInj	57.0	1,047
HighInj	43.0	790
Total	26.0	1,837

HighInj (2,406/3,649)

Category	%	n
LowInj	34.1	1,243
HighInj	65.9	2,406
Total	51.6	3,649

HighInj (89/170)

Category	%	n
LowInj	47.6	81
HighInj	52.4	89
Total	2.4	170

LowInj (1,162/1,410)

Category	%	n
LowInj	82.4	1,162
HighInj	17.6	248
Total	20.0	1,410

Figure 23: Decision Tree View

6 Testing & Evaluation

6.1 Variable Importance Analysis

In our model, the "EJECTION" variable emerges as the most critical predictor, exerting significant influence on the model's outcomes. Following closely is the "TOWED" feature, which ranks as the second most significant, underscoring its substantial impact on predictions. Variables associated with airbags ("AIR_BAG_binned") and light conditions ("LGT_COND") exhibit moderate importance, contributing meaningfully to predictive accuracy. Conversely, "DAY_WEEK_str" and "INT_HWY" are identified as the least impactful variables in our model. Interestingly, the presence of missing data in certain variables ("Trav_SP_Miss", "Mod_Year_Miss") stands out as notably influential, suggesting that the absence of data may carry meaningful implications for prediction accuracy and model performance.

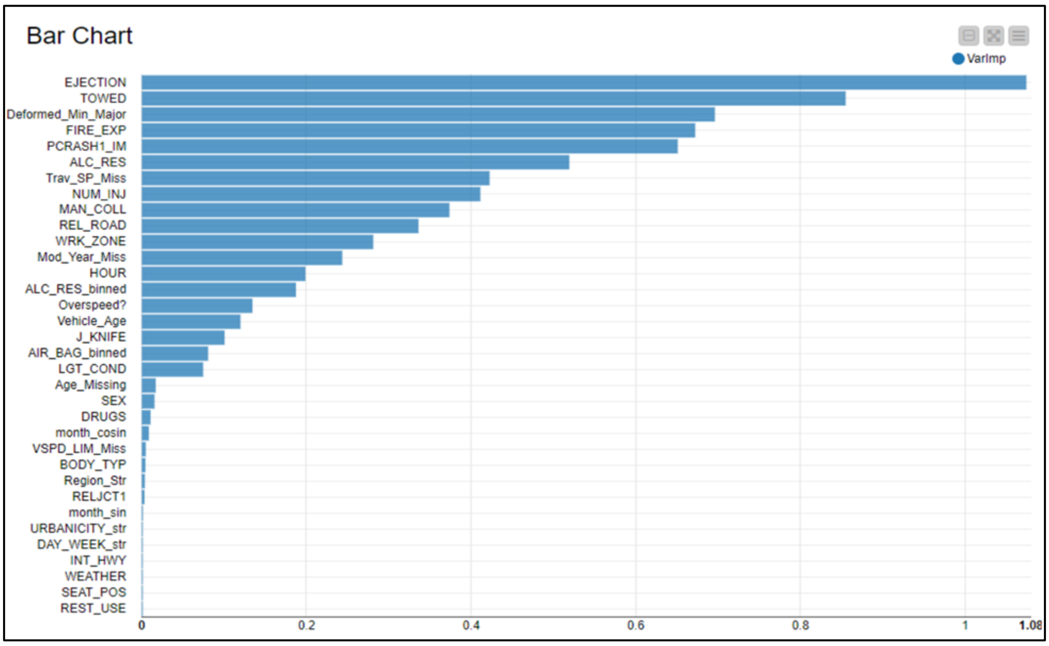


Figure 24: KNIME generated Variable Importance Chart

6.2 Model Evaluation – ROC Curve

Among the models evaluated, Gradient Boosting Tree (GBT) demonstrates superior performance with an AUC of 0.787, showcasing its effectiveness in accurately distinguishing between various levels of injury severities when compared to alternative models. Conversely, the Decision Tree model exhibits the lowest AUC of 0.634 among the models considered, indicating comparatively weaker discriminatory power in predicting injury severity outcomes. This highlights the notable advantage of employing Gradient Boosting Trees for robust and precise injury severity predictions.

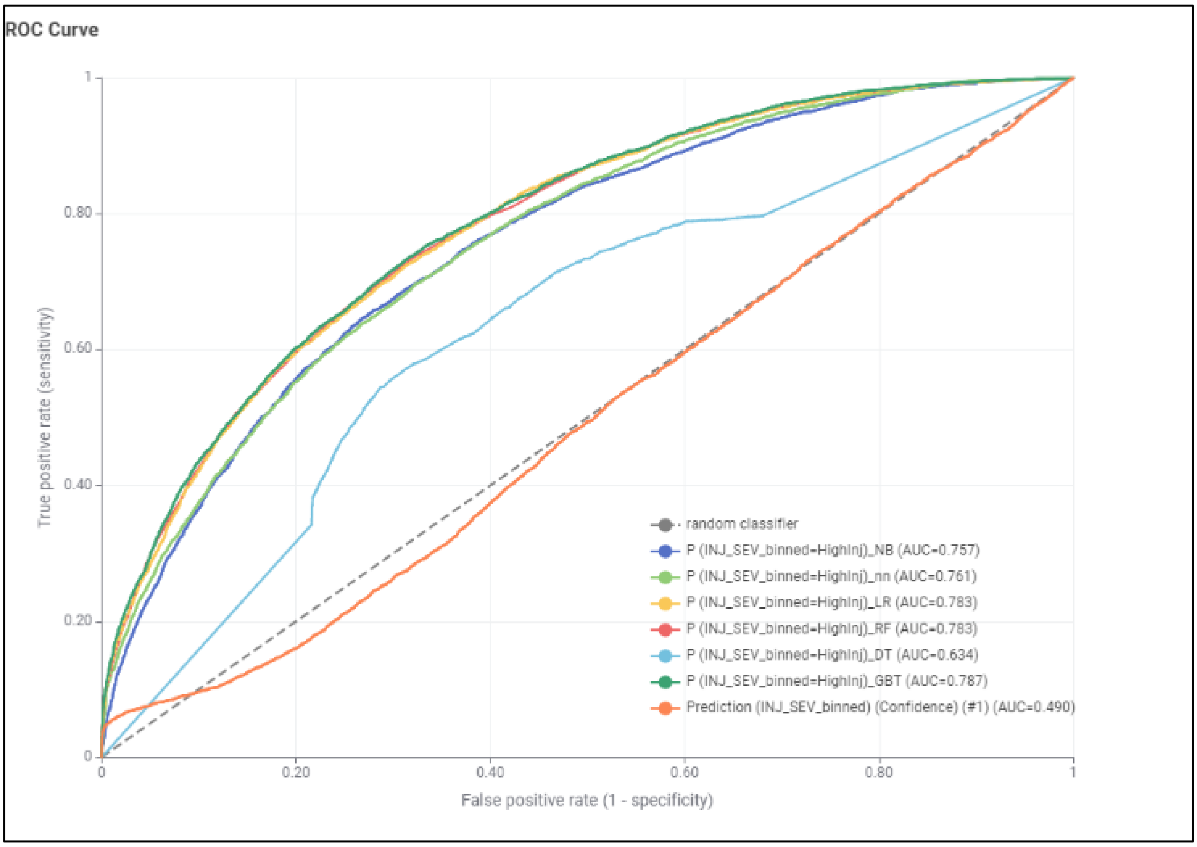


Figure 25: Combined ROC Curve

6.3 Model Summary

The table summarizes the performance metrics of various models for predicting injury severity. Among these models, the Gradient Boosting Tree (GBT) stands out with an accuracy of 70.142%, showcasing its effectiveness in overall predictions. GBT also exhibits the highest sensitivity of 91%, indicating its ability to accurately identify positive cases of high injury severity.

Model	**Accuracy**	**Sensitivity**	**Specificity**	**ROC Value**
Artificial Neural Network	65.7%	73.1%	64%	76.2%
Logistic Regression	66.8%	76%	64.6%	78.2%
Naïve Bayes	63.2%	77%	59.9%	75.7%
Random Forest	**75.3%**	61.4%	78.5%	78.4%
Decision Tree	62.6%	61.8%	62.8%	63.4%
Gradient Boosting Tree	70.142%	**91%**	**82.3%**	**78.7%**

Table. 1: Model Summary

Additionally, GBT maintains a respectable specificity of 82.3%, highlighting its capability to correctly identify negative cases of low injury severity. Despite its lower accuracy, the Random Forest model also performs well with a balanced sensitivity and specificity, resulting in a high ROC value of 78.4%.

Logistic Regression and Artificial Neural Network models also demonstrate competitive performance across the metrics, with Logistic Regression slightly edging out in accuracy and sensitivity. However, both models fall slightly behind in terms of ROC value compared to GBT and Random Forest.

Naïve Bayes exhibits lower accuracy and specificity compared to other models but maintains a high sensitivity, indicating its strength in identifying positive cases accurately. On the other hand, the Decision Tree model shows the lowest performance across all metrics, suggesting limitations in its predictive power for injury severity outcomes in this context.

Overall, GBT emerges as the top-performing model in this comparison, followed closely by Random Forest, Logistic Regression, and Artificial Neural Network models

7 Deployment

During the deployment phase, insights from predictive models will be instrumental in reinforcing vehicular safety standards, with a focus on preventing ejections and regulating towing practices. These factors have been identified as critical determinants of injury severity and will receive priority in our safety recommendations. By leveraging model findings, we aim to advocate for enhanced driving safety laws and practices, collaborating closely with roadway safety organizations and policymakers.

Engagement with key stakeholders such as the NHTSA and insurance agencies are essential to ensure that traffic regulations and emergency protocols are in line with our predictive model outcomes. This alignment is particularly crucial for aspects such as airbag effectiveness and lighting conditions, which play significant roles in preventing severe injuries. Securing funding and allocating resources effectively will support safety campaigns and technological advancements, addressing challenges posed by weather-related risks and driver distractions.

Additionally, establishing a continuous improvement loop through training based on model findings and sensitivity analyses will maintain the accuracy and relevance of the predictive model amidst evolving road safety dynamics.

8 Conclusion & Future Scope

8.1 Conclusion

Throughout this project, we have successfully employed several advanced predictive modeling techniques, including decision trees, random forests, and neural networks. Notably, the Random Forest model excelled, achieving the highest Receiver Operating Characteristic (ROC) score, indicating its robustness in classifying the injury severity based on the variables studied. The analytical strength of this model suggests that it could serve as a reliable tool in understanding and predicting injury outcomes in vehicle accidents.

The models generated through this project have provided pivotal insights, particularly highlighting the importance of public education programs and vehicular safety enhancements in reducing the severity of injuries from accidents. These insights reinforce the notion that well-implemented safety features and informed driving behaviors significantly mitigate the risks associated with road accidents. The application of these models is not merely academic; they have practical implications that can extend to real-world applications, providing continuous and valuable guidance to manufacturers, policy makers, and public health officials.

Furthermore, the consistent application of these predictive models promises to yield ongoing benefits, offering data-driven insights that can guide the development of targeted interventions. These interventions can be specifically tailored to address the most significant factors contributing to high injury severity, thus enhancing road safety more effectively. As these models are integrated and updated with new data over time, their predictive accuracy and reliability are expected to improve, making them indispensable tools in the field of road safety analysis.

8.2 Future Scope

The potential for future enhancements and applications of this project is vast. By incorporating a broader array of datasets, including those from newer or unexplored demographics and geographic regions, the models can be significantly enriched. This expansion would likely uncover new insights and improve the generalizability of the model predictions across different contexts and conditions. Moreover, experimenting with cutting-edge machine learning algorithms could unlock higher levels of model accuracy and provide deeper insights into complex interdependencies within the data.

An exciting prospect for this research lies in the integration of these models into real-time monitoring systems. Such systems could revolutionize how traffic safety measures are implemented, allowing for immediate responsiveness to emerging risks and adapting interventions in real-time based on live data analysis. This proactive approach in risk management could greatly enhance the efficacy of safety measures, ultimately leading to a substantial reduction in severe traffic-related injuries. As we move forward, the focus will also be on refining these models through continuous feedback loops and sensitivity analyses, ensuring they remain aligned with the latest trends and safety standards in traffic management and vehicle safety technology.

9 References

[1] KNIME AG. KNIME Analytics Platform (2021). URL Available at: https://www.knime.com/knime-software.

[2] Delen, Dursun & Tomak, Leman & Topuz, Kazim & Eryarsoy, Enes. (2017). Investigating injury severity risk factors in automobile crashes with predictive analytics and sensitivity analysis methods. Journal of Transport & Health. 4. 10.1016/j.jth.2017.01.009.

[3] Newly released estimates show traffic fatalities reached a 16-Year high in 2021. (2022, May 17). NHTSA. https://www.nhtsa.gov/press-releases/early-estimate-2021-traffic-fatalities