MSIS 5633 – Predictive Analytics Technologies

(Section 67258)

MSIS 5633 – Term Project Team 1

Injury Severity Risk Factors in Automobile Crashes

Due Date

November 26, 2023

By

Sriharsha Vajjala and Marshall Proctor

Team Members

Marshall Proctor – Second year MS MIS student with a BSBA in MIS from Oklahoma State University

Sriharsha Vajjala - Second year MS MIS student, Oklahoma State University.

Executive Summary

Project Overview, Objectives, and Methodology

In this report, we'll be looking at the findings of a data analytics project aimed at understanding the key factors that contribute most significantly to higher injury severity in automobile crashes. Our dataset includes a wide variety of variables including time of day, date, weather conditions, manner of collision, presence of fire, distractions, intoxication, and vehicle condition. Using several number-based and machine learning models to evaluate the data, we want to determine patterns, correlations, and insights that can help improve vehicle safety measures and policies. The dataset used for the project was Crash Report Sampling System data from the National Highway Traffic Safety Administration, gathered from 2021 using information from police reports filed for each crash.

Key Findings:

- Our model indicates that the factors such as 'Ejection', 'Restraint System Usage', 'Airbag Misuse', and 'Fire Involvement in Crash' are crucial in contributing to high injury severity in vehicular incidents. These elements demonstrate a strong predictive relationship with the outcomes, suggesting that they are significant indicators of the potential for severe injuries.
- Additional factors including 'Pre-Crash Movements', 'Exceeding Speed Limit', 'Drinking', and 'Jackknife Events' also emerge as significant contributors to injury severity. Their presence in the model's variable importance ranking underscores their role in the circumstances leading to more serious injuries.

Recommendations:

- Based on our findings, it is recommended to enhance vehicle safety measures. This could include improving the design and enforcement of restraint systems, increasing public awareness campaigns on the proper use of airbags, and incorporating advanced materials and technology to reduce the likelihood of fires in crashes. These actions could significantly mitigate the risk of high-severity injuries in vehicular accidents.
- The significance of factors like 'Pre-Crash Movements', 'Exceeding Speed Limit', 'Drinking', and 'Jackknife Events' suggests a need for targeted preventive strategies. It is

recommended that policymakers implement stricter speed regulations, enforce anti-drunk driving laws more rigorously, and promote advanced driver-assistance systems to alert drivers of unsafe pre-crash movements. These measures address the behavioral and situational contributors to severe crash outcomes.

Through the remaining sections of the report, we will cover this information in greater detail to explain how these findings affect business applications, what the data means and how it was processed, which models were used and what insights were found using them, as well as how to apply the recommendations listed above.

Business Understanding

Business Objectives:

The overarching business objective is to enhance road safety and reduce the severity of injuries resulting from automobile crashes. This involves gaining a detailed understanding of the key determinants of injury severity, with the goal of informing targeted solutions and policies. By identifying influential factors, businesses and policymakers can develop more effective strategies to mitigate the impact of crashes on individuals and communities.

Stakeholders:

Stakeholders in this analysis include government agencies responsible for transportation and road safety, law enforcement agencies, automobile manufacturers, insurance companies, and public health organizations. Each of these stakeholders has a vested interest in understanding the factors influencing injury severity, as it directly impacts their roles and responsibilities within the context of road safety and public health.

Project Scope:

The scope of the project is an analysis of a diverse dataset including variables such as time of day, date, weather conditions, manner of collision, presence of fire, distractions, intoxication, and vehicle condition. The focus is on understanding the relationships and patterns that contribute to injury severity in automobile crashes. The findings will be used to generate actionable insights that inform solutions such as improved vehicle designs and new road safety policies.

Success Criteria:

The success of the project will be measured by the extent to which the identified factors contribute to insights and strategies for mitigating injury. Success criteria include the ability to:

- Provide a clear understanding of the factors influencing injury severity.
- Determine which factors have the most significant impact on severity.
- Develop clear and actionable recommendations for stakeholders to reduce the impact of these factors.

Risks and Limitations:

It is important to acknowledge potential risks and limitations associated with the project. These may include data quality issues, biases in the dataset, and external factors that could impact the findings. Addressing these risks will be an integral part of ensuring the reliability and applicability of the results to real-world scenarios.

Deployment Considerations:

The deployment of recommendations derived from this analysis will involve collaboration with relevant stakeholders. Recommendations will be tailored to the specific needs and responsibilities of each stakeholder group to ensure that the findings are translated into effective solutions for reducing injury severity.

By identifying key business objectives and stakeholders and establishing success criteria, the business understanding phase of the project provides a framework for the subsequent phases that begins in the next sections with data understanding.

Data Understanding

This study utilizes data from the Crash Report Sampling System (CRSS), focusing specifically on accidents that occurred in the year 2021. The CRSS dataset is comprehensive, containing various files, but our analysis is primarily concerned with files related to accidents, persons involved, vehicles, and distractions. The objective is to identify risk factors associated with automobile crashes, including jackknife incidents.

The accident file in the dataset includes key features such as weather conditions, the month and day of the crash, alcohol involvement, intersection type, manner of collision, among other variables. Complementarily, the vehicle file provides details about vehicle body type, model, speed, and the extent of damage. Additionally, the person file contains demographic and safety information about individuals involved in the crashes, including age, sex, and airbag usage.

A unique identifier called casenum in the accidents file facilitates an inner join with the vehicles file. This casenum, along with a vehicle number, is then used to perform a left join with the persons file and subsequently with the distractions file.

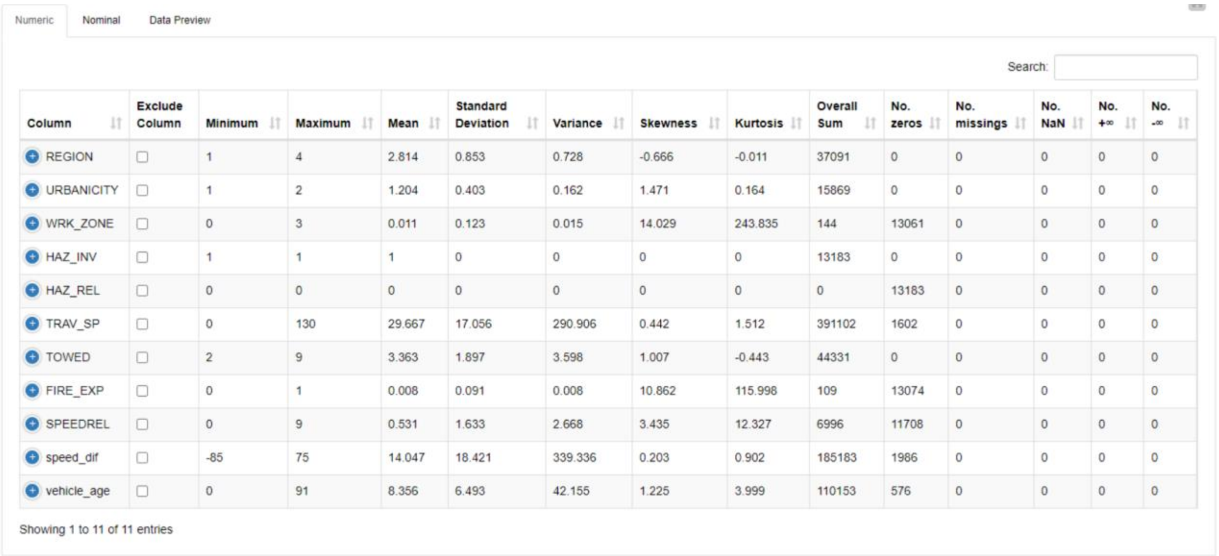
After conducting these joins, the consolidated dataset comprised 128,393 rows and 151 columns. However, to tailor the dataset more closely to our research objectives, we eliminated certain columns prior to performing the joins. This data curation process resulted in a dataset with a total of 54 columns. To gain a deeper understanding of this dataset, we conducted exploratory data analysis, focusing on key statistics such as missing values, means, medians, minimums, and maximums for all the columns in the raw dataset.

Data Preparation

In preparing the data for our analysis, we first removed columns that were not relevant to our study. This step helped us focus on the most important aspects of the data. We specifically chose to analyze accidents involving only automobiles, so we excluded data about other types of vehicles like motorcycles, farm vehicles, buses, and large trucks. We also focused our analysis on drivers. This meant we removed records of people who were not driving, like pedestrians and passengers. For our target variable, INJ_SEV, which shows how severe an injury is, we divided it into two groups: 'High' and 'Low'. We also removed records where no one was injured because we wanted to understand what factors lead to injuries.

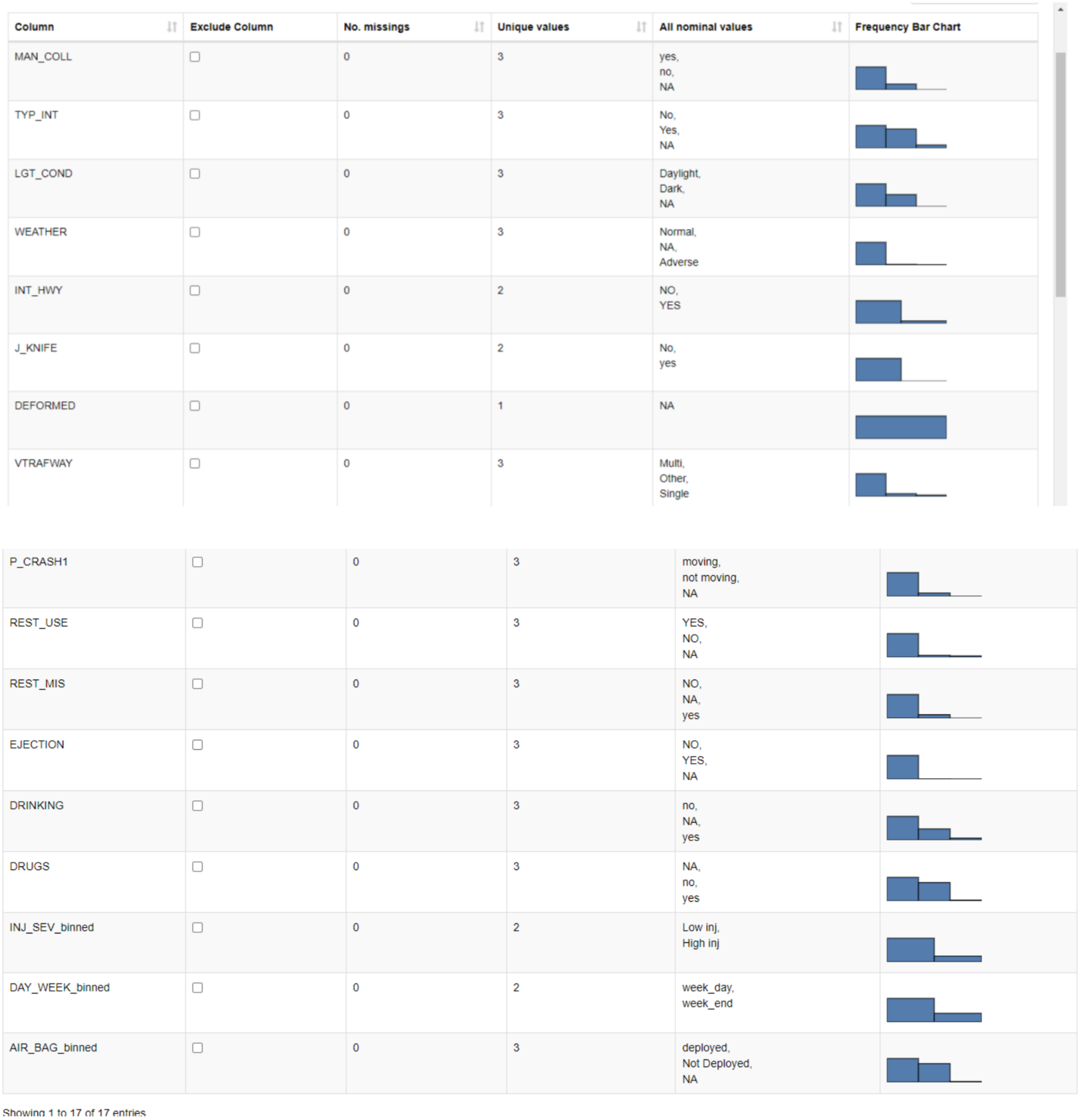
When it came to missing data in numerical variables like age and hour of the crash, we filled in these gaps using the median values of these variables. We also added new columns to our dataset. For example, we created a column to show the difference between a vehicle's speed and the speed limit and another one to show the age of the vehicle based on its model year and the year of the crash. Many of our categorical variables, like alcohol usage, drug usage, and airbag deployment, had a lot of missing or 'unreported' data. Instead of removing these records, we made new categories for them. This way, we could keep the data in our analysis without guessing what the missing values might be. We also grouped some of the data into simpler categories. For example, we classified the day of the accident as either a 'weekday' or 'weekend' and the hour of the crash as 'morning', 'afternoon', or 'night'. These steps in preparing the data were important to make sure our analysis would be focused on what we wanted to find out about the risk factors in automobile crashes. Also, we have removed the original columns after adding the derived columns to the dataset and removed the identity columns which we used to join different files.

Following the initial data cleanup and transformations, we applied a methodical approach to refine our dataset further and identify the most impactful features. We employed backward elimination(fig 2), a technique where we systematically removed one input feature at a time and trained the Random Forest model on the remaining features. This process was aimed at identifying features that, when removed, caused the least increase in the error rate. It allowed us to focus on variables that were most influential for our model's predictions. The backward elimination process was performed on the 54 columns we initially retained after data preprocessing. By iteratively removing the least significant features based on their impact on the model's error rate, we were able to reduce the number of variables in our dataset. The below image (fig 1) presents the final list of columns that remained in the dataset after completing the backward elimination process, along with the necessary data transformations and imputation. It also includes corresponding statistics for each column, providing a comprehensive overview of the features.



Column	Exclude Column	Minimum	Maximum	Mean	Standard Deviation	Variance	Skewness	Kurtosis	Overall Sum	No. zeros	No. missings	No. NaN	No. +∞	No. -∞
REGION	□	1	4	2.814	0.853	0.728	-0.666	-0.011	37091	0	0	0	0	0
URBANICITY	□	1	2	1.204	0.403	0.162	1.471	0.164	15869	0	0	0	0	0
WRK_ZONE	□	0	3	0.011	0.123	0.015	14.029	243.835	144	13061	0	0	0	0
HAZ_INV	□	1	1	1	0	0	0	0	13183	0	0	0	0	0
HAZ_REL	□	0	0	0	0	0	0	0	0	13183	0	0	0	0
TRAV_SP	□	0	130	29.667	17.056	290.906	0.442	1.512	391102	1602	0	0	0	0
TOWED	□	2	9	3.363	1.897	3.598	1.007	-0.443	44331	0	0	0	0	0
FIRE_EXP	□	0	1	0.008	0.091	0.008	10.862	115.998	109	13074	0	0	0	0
SPEEDREL	□	0	9	0.531	1.633	2.668	3.435	12.327	6996	11708	0	0	0	0
speed_dif	□	-85	75	14.047	18.421	339.336	0.203	0.902	185183	1986	0	0	0	0
vehicle_age	□	0	91	8.356	6.493	42.155	1.225	3.999	110153	576	0	0	0	0

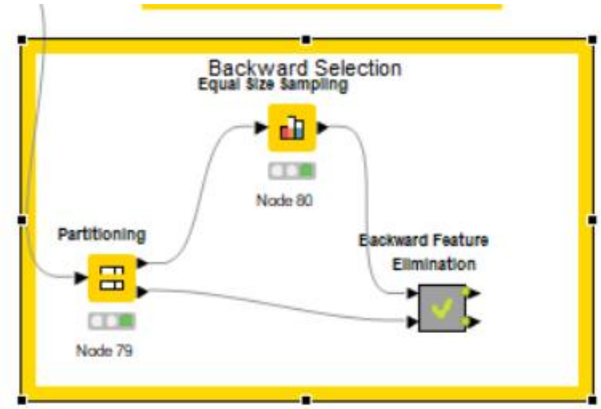
Showing 1 to 11 of 11 entries

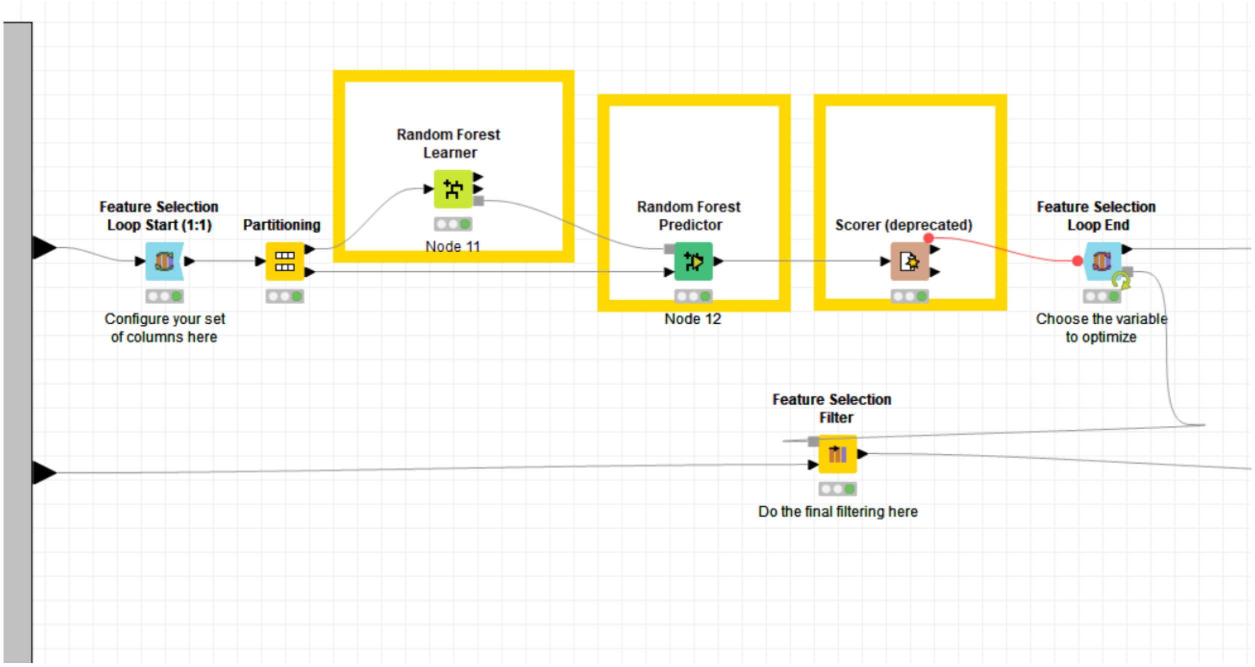


Column	Exclude Column	No. missings	Unique values	All nominal values	Frequency Bar Chart
MAN_COLL	□	0	3	yes, no, NA	
TYP_INT	□	0	3	No, Yes, NA	
LGT_COND	□	0	3	Daylight, Dark, NA	
WEATHER	□	0	3	Normal, NA, Adverse	
INT_HWY	□	0	2	NO, YES	
J_KNIFE	□	0	2	No, yes	
DEFORMED	□	0	1	NA	
VTRAFWAY	□	0	3	Multi, Other, Single	
P_CRASH1	□	0	3	moving, not moving, NA	
REST_USE	□	0	3	YES, NO, NA	
REST_MIS	□	0	3	NO, NA, yes	
EJECTION	□	0	3	NO, YES, NA	
DRINKING	□	0	3	no, NA, yes	
DRUGS	□	0	3	NA, no, yes	
INJ_SEV_binned	□	0	2	Low inj, High inj	
DAY_WEEK_binned	□	0	2	week_day, week_end	
AIR_BAG_binned	□	0	3	deployed, Not Deployed, NA	

Showing 1 to 17 of 17 entries

Fig 1: numeric and categorical variables





Our dataset exhibited a class imbalance, with a significantly higher number of records classified as 'low injury' compared to those labeled as 'high injury'. Specifically, 'low injury' cases constituted about 20% of the data. Such an imbalance can lead to challenges in model training, as it might cause the model to under-learn information about the minority class, in this case, the 'high injury' records. To address this issue and ensure a more balanced representation of both classes, we explored various sampling techniques.

After considering the options, we decided to implement under sampling. This approach involves reducing the number of instances from the majority class to match the number of instances in the minority class. By doing so, we created equal partitions of both 'low injury' and 'high injury' classes to be provided as training input to our models. This decision was instrumental in mitigating the risk of class imbalance, ensuring that our model would not be biased towards the majority class and would be able to learn effectively from both categories of injury severity. we chose under sampling as it provides a balanced train data set without artificially inflating the dataset size.

Prediction Models:

In our research, we are looking at the target variable INJ_SEV, which has two types: 'Low' and 'High' severity. This means we are dealing with a binary classification problem, where we have to categorize each case into one of these two types. To do this classification, we decided to use different models like Decision Trees, Logistic Regression, Neural Networks, and Random Forest. In the next parts of this report, we will talk about each of these models.

Decision Tree:

One of the fundamental approaches to solving classification problems is through the use of Decision Trees. In our project, we employed a Decision Tree to classify the severity of injuries in automobile accidents. A key advantage of Decision Trees, especially when compared to more complex models like Neural Networks, is their interpretability. The tree structure of a Decision Tree allows us to clearly see and understand the decision-making process within the model. This feature is particularly valuable for explaining the model's behavior to stakeholders who may not be versed in machine learning.

In our exploration of Decision Trees within the KNIME environment, we used various configurations to optimize the model's performance. This included experimenting with different quality measures such as Gini impurity and Gain Ratio. After comparing the results from these different configurations, we decided to use the Gini index as our quality measure and opted against pruning.

We have also used ROC (Receiver Operating Characteristic) curves to evaluate the performance of our Decision Tree model. The ROC curve, as depicted in Fig. , illustrates the diagnostic ability of the classifier by plotting the true positive rate against the false positive rate at various threshold settings. Additionally, we have compiled a confusion matrix, which provides a detailed breakdown of the model's predictions versus the actual outcomes. This, along with the accuracy statistics, gives us a comprehensive view of the model's performance.

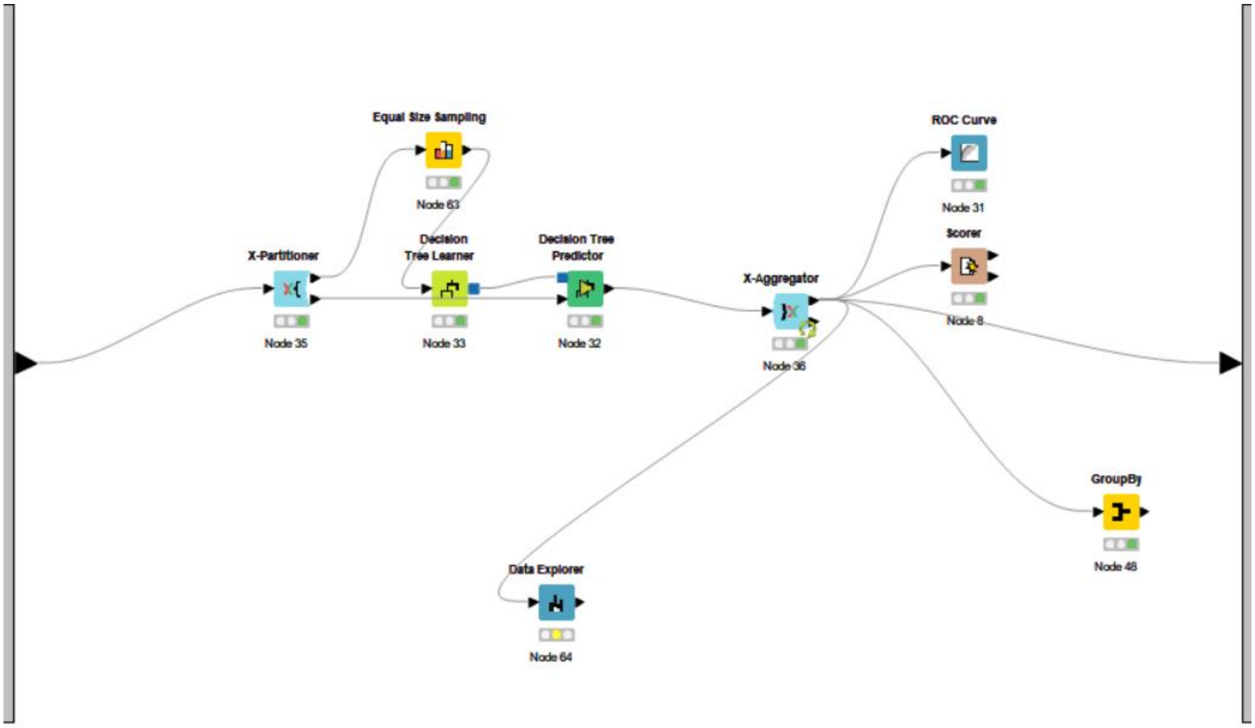
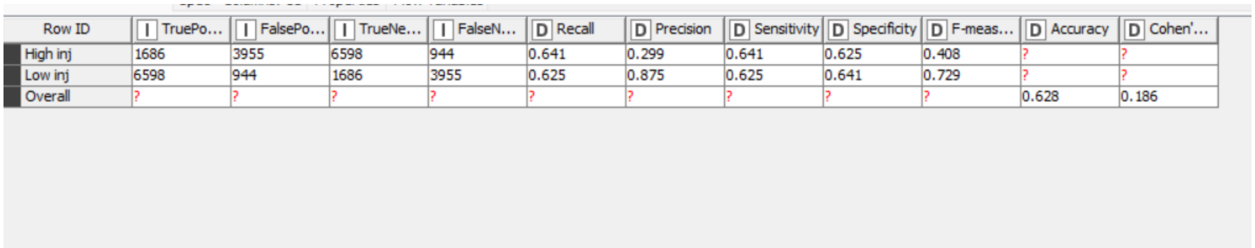


Fig 3 decision tree



Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specificity	D F-meas...	D Accuracy	D Cohen'...
High inj	1686	3955	6598	944	0.641	0.299	0.641	0.625	0.408	?	?
Low inj	6598	944	1686	3955	0.625	0.875	0.625	0.641	0.729	?	?
Overall	?	?	?	?	?	?	?	?	?	0.628	0.186

Fig 3a. Accuracy Statistics

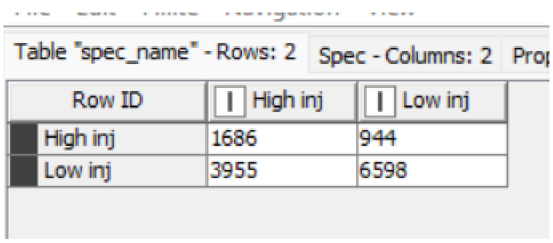
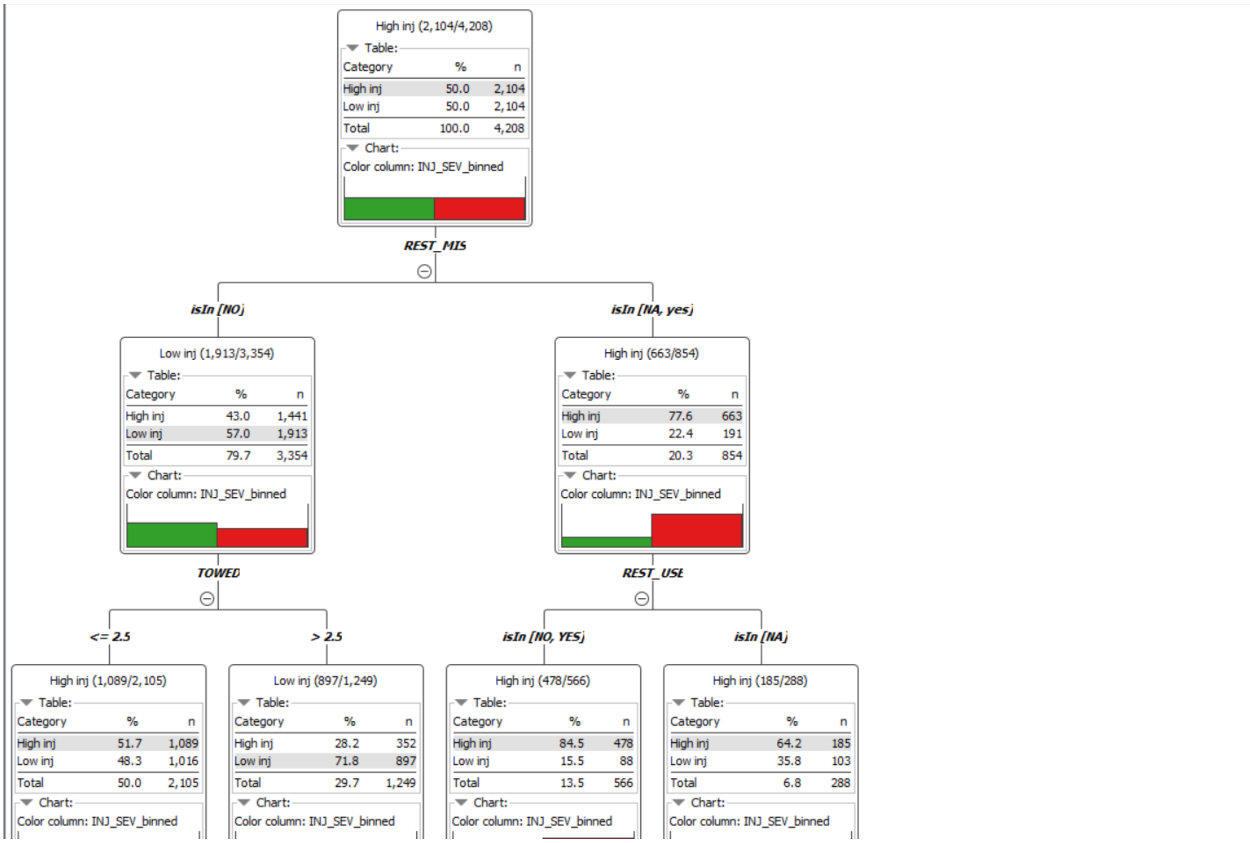


Table "spec_name" - Rows: 2 | Spec - Columns: 2 | Prop

Row ID	High inj	Low inj
High inj	1686	944
Low inj	3955	6598

Fig 3b. Confusion Matrix

The below image 3c shows the top 3 level splits in the decision tree, the variables at the top of the tree indicates that these variables are significant for the classification.



High inj (2,104/4,208)

Table:

Category	%	n
High inj	50.0	2,104
Low inj	50.0	2,104
Total	100.0	4,208

Chart:
Color column: INJ_SEV_binned

REST_MIS

isIn [NO]

Low inj (1,913/3,354)

Table:

Category	%	n
High inj	43.0	1,441
Low inj	57.0	1,913
Total	79.7	3,354

Chart:
Color column: INJ_SEV_binned

isIn [NA, yes]

High inj (663/854)

Table:

Category	%	n
High inj	77.6	663
Low inj	22.4	191
Total	20.3	854

Chart:
Color column: INJ_SEV_binned

TOWED

<= 2.5

High inj (1,089/2,105)

Table:

Category	%	n
High inj	51.7	1,089
Low inj	48.3	1,016
Total	50.0	2,105

Chart:
Color column: INJ_SEV_binned

> 2.5

Low inj (897/1,249)

Table:

Category	%	n
High inj	28.2	352
Low inj	71.8	897
Total	29.7	1,249

Chart:
Color column: INJ_SEV_binned

REST_USE

isIn [NO, YES]

High inj (478/566)

Table:

Category	%	n
High inj	84.5	478
Low inj	15.5	88
Total	13.5	566

Chart:
Color column: INJ_SEV_binned

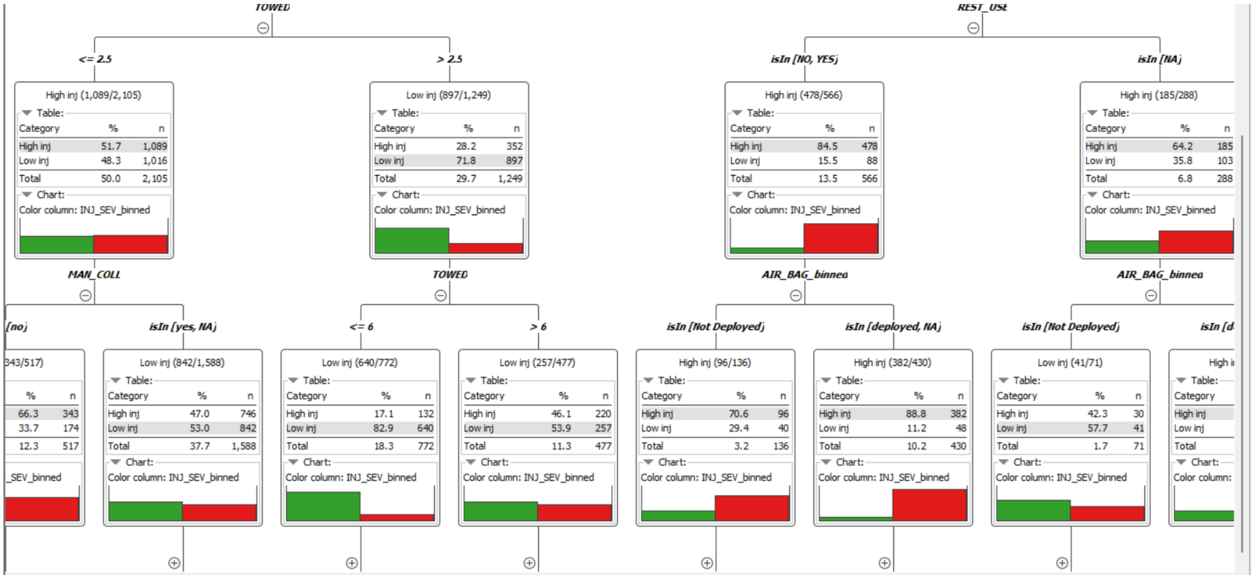
isIn [NA]

High inj (185/288)

Table:

Category	%	n
High inj	64.2	185
Low inj	35.8	103
Total	6.8	288

Chart:
Color column: INJ_SEV_binned



Random forest:

In addition to Decision Trees, we also used the Random Forest algorithm into our analysis to classify injury severity in automobile accidents. Random Forest, an ensemble method that combines multiple decision trees, was chosen for its robustness and ability to handle the complexities of our dataset. Unlike a single Decision Tree, Random Forest aggregates the predictions from numerous trees, which helps in reducing the risk of overfitting and increases the overall accuracy.

For our specific case, we carefully tuned the parameters of the Random Forest model in KNIME. This included adjusting the number of trees in the forest and the depth of each tree, as well as experimenting with different methods for sampling the data. The performance of the Random Forest model was evaluated using similar metrics as with the Decision Tree, including the ROC curve and confusion matrix. The results from the Random Forest were particularly insightful, as they provided a more nuanced understanding of the predictive power of our features when combined in an ensemble setting. This approach complemented our Decision Tree analysis, offering a broader view of the factors contributing to injury severity in car accidents.

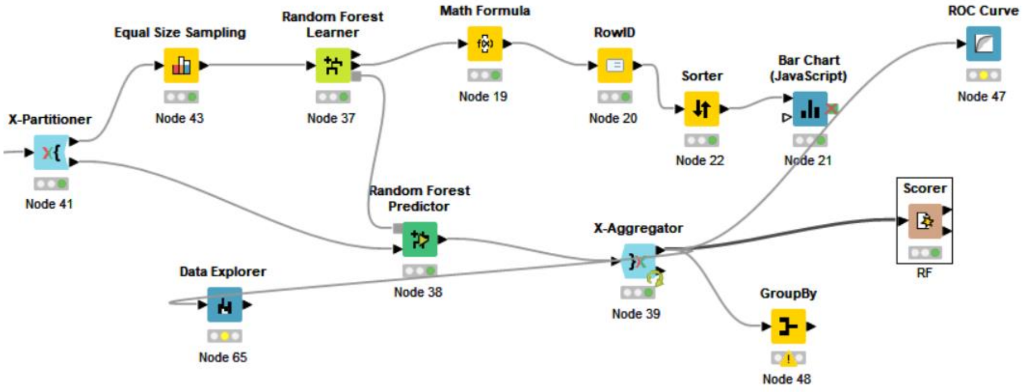


Fig 4. Random forest

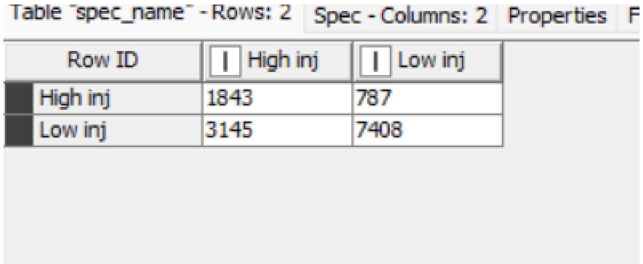


Table "spec_name" - Rows: 2 | Spec - Columns: 2 | Properties | F

Row ID	I High inj	I Low inj
High inj	1843	787
Low inj	3145	7408

Fig 4a. Confusion matrix

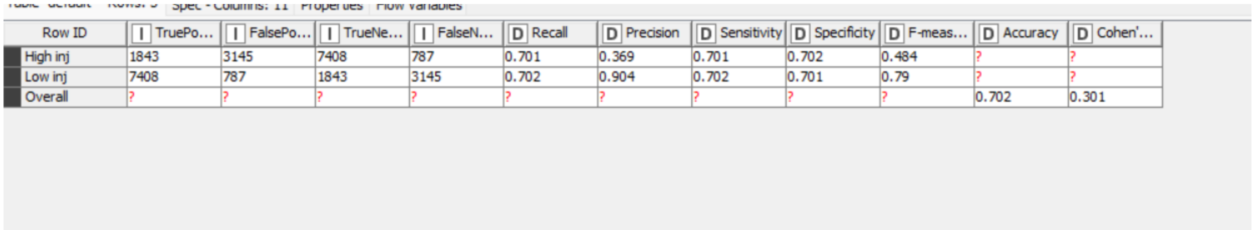


Table "default" - Rows: 3 | Spec - Columns: 11 | Properties | Flow Variables

Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specificity	D F-meas...	D Accuracy	D Cohen'...
High inj	1843	3145	7408	787	0.701	0.369	0.701	0.702	0.484	?	?
Low inj	7408	787	1843	3145	0.702	0.904	0.702	0.701	0.79	?	?
Overall	?	?	?	?	?	?	?	?	?	0.702	0.301

Fig 4b. Accuracy Statistics

Logistic Regression:

Another model we employed in our study was Logistic Regression. Although it's often associated with continuous outcomes, Logistic Regression is also effective for classification tasks. It predicts the probability of categorical outcomes, which made it suitable for our goal of classifying 'Low' and 'High' injury severity in car accidents.

For the configuration of Logistic Regression in our dataset, which contained many categorical variables, we implemented one-to-many encoding. This technique transformed categorical variables into several binary columns, allowing the Logistic Regression model to process them effectively. Furthermore, we utilized different normalization techniques, such as min-max scaling and z-score normalization, to address the issue of numerical variables with high values. These normalization steps were crucial to ensure that all variables contributed equally to the model, preventing any single variable with a larger scale from disproportionately influencing the results. Assessing the performance of the Logistic Regression model, we used the same metrics as for our other models, like ROC curves and confusion matrices.

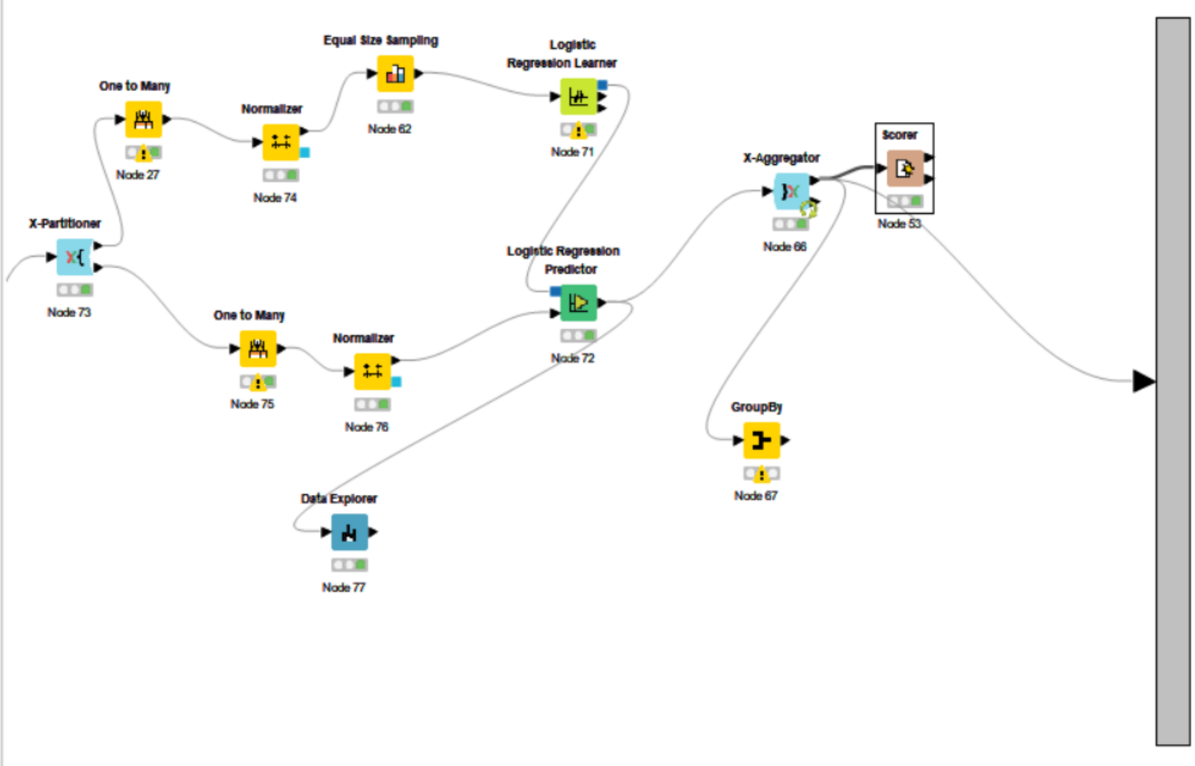
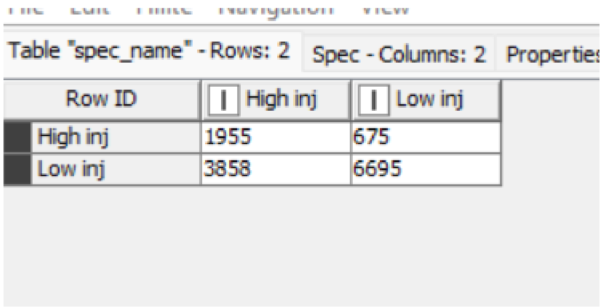


Fig. 5: Logistic Regression.



Row ID	High inj	Low inj
High inj	1955	675
Low inj	3858	6695

Fig 5a: confusion matrix

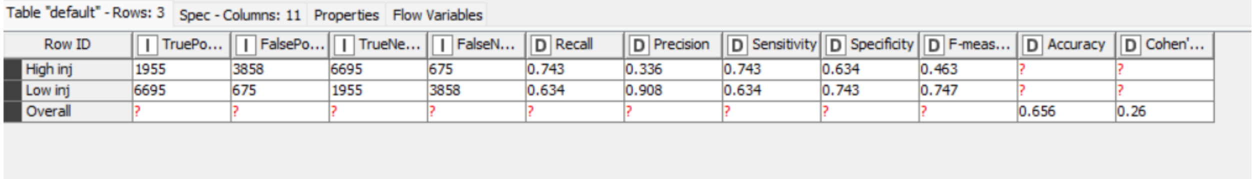


Table "default" - Rows: 3 | Spec - Columns: 11 | Properties | Flow Variables

Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specificity	D F-meas...	D Accuracy	D Cohen'...
High inj	1955	3858	6695	675	0.743	0.336	0.743	0.634	0.463	?	?
Low inj	6695	675	1955	3858	0.634	0.908	0.634	0.743	0.747	?	?
Overall	?	?	?	?	?	?	?	?	?	0.656	0.26

Fig 5b. Accuracy statistics

Neural Networks: In our project, we used a type of Neural Network called a Multi-Layer Perceptron (MLP) to classify injuries in car accidents as either 'Low' or 'High' severity. Neural Networks are good at finding patterns in complex data, and they work a bit like how the brain learns from experience. The MLP is a popular choice in many studies because it's good at solving both prediction and classification problems.

For our problem, the MLP seemed like a good fit. We set it up to have layers of nodes (or perceptrons), which take in our data and pass on information to the next layer. The MLP learns by adjusting itself based on the mistakes it makes in predicting. It gets better over time, much like learning from past experiences.

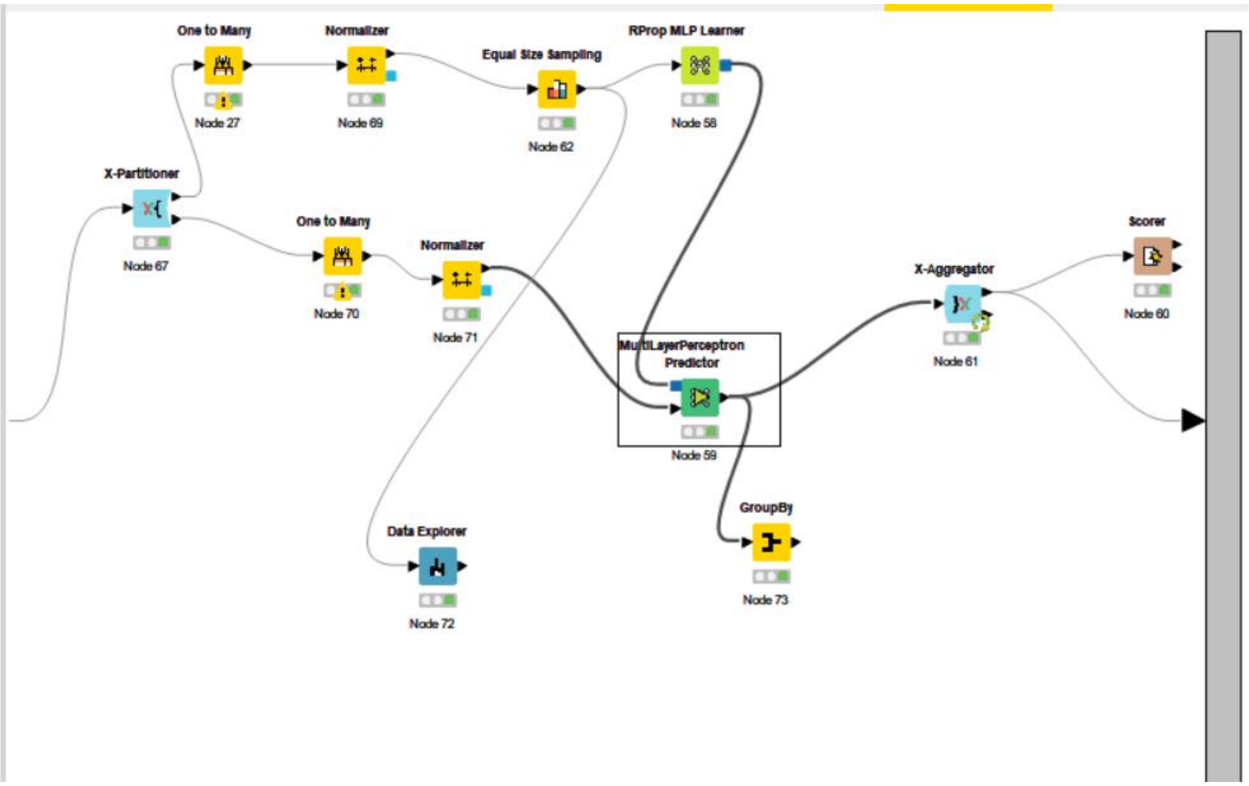
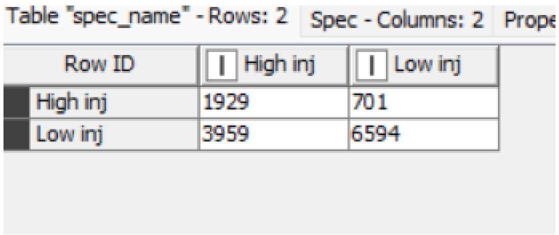
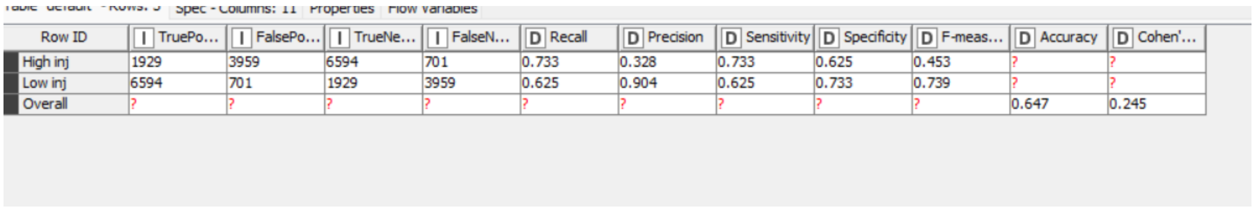


Fig 6 a: Neural networks



Row ID	High inj	Low inj
High inj	1929	701
Low inj	3959	6594

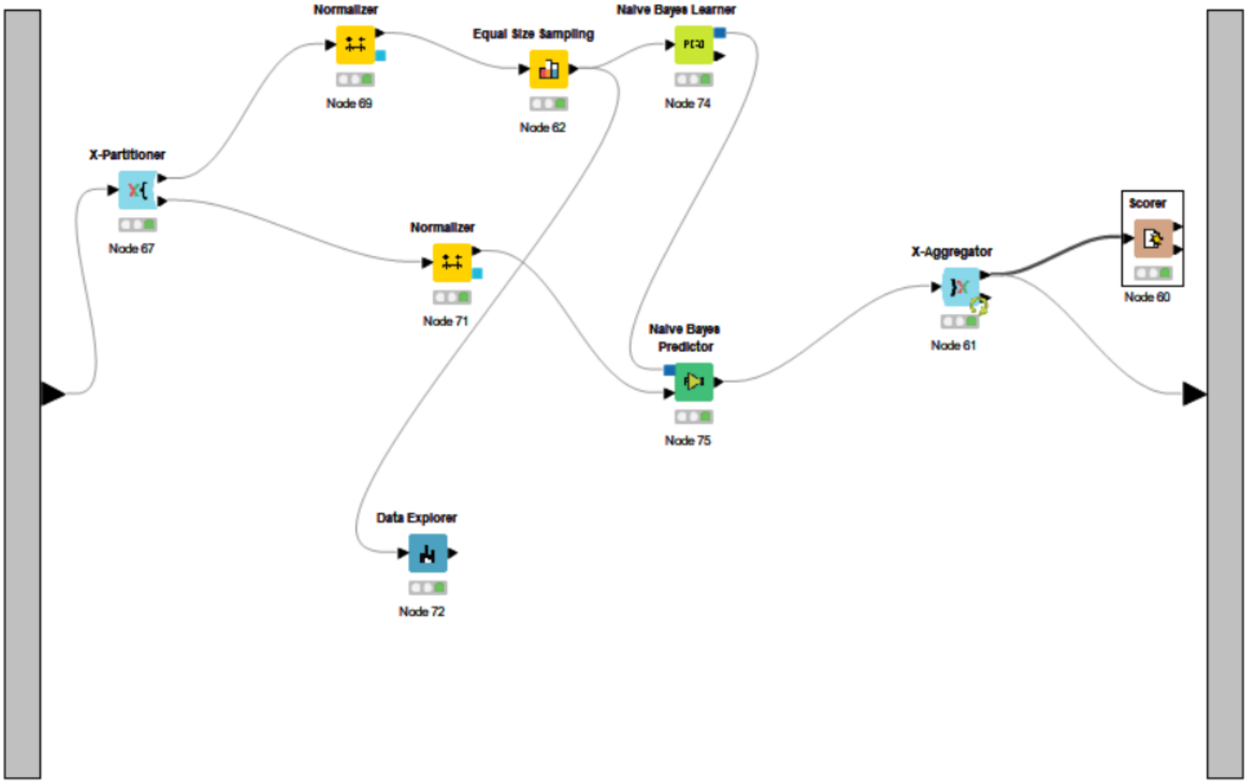
Fig:6 b confusion matrix

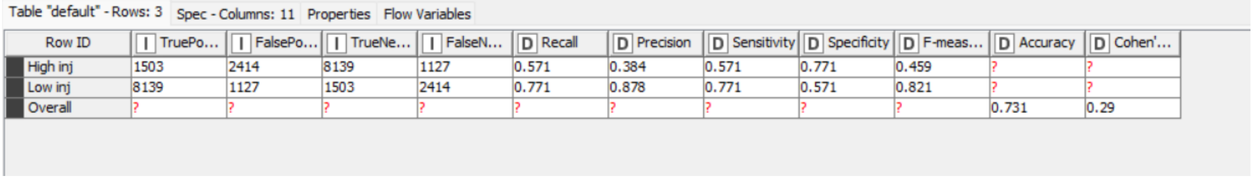


Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specificity	F-meas...	Accuracy	Cohen'...
High inj	1929	3959	6594	701	0.733	0.328	0.733	0.625	0.453	?	?
Low inj	6594	701	1929	3959	0.625	0.904	0.625	0.733	0.739	?	?
Overall	?	?	?	?	?	?	?	?	?	0.647	0.245

Fig 6c. Accuracy Statistics

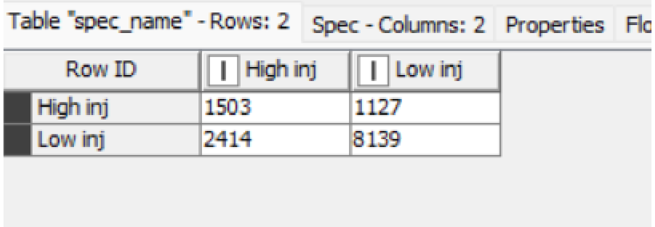
Naïve Bias:





Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specificity	D F-meas...	D Accuracy	D Cohen'...
High inj	1503	2414	8139	1127	0.571	0.384	0.571	0.771	0.459	?	?
Low inj	8139	1127	1503	2414	0.771	0.878	0.771	0.571	0.821	?	?
Overall	?	?	?	?	?	?	?	?	?	0.731	0.29

Fig 7a. Accuracy Statistics



Row ID	I High inj	I Low inj
High inj	1503	1127
Low inj	2414	8139

Fig 7b. Confusion matrix

K cross validation:

To ensure the robustness and reliability of our predictive models, including the Decision Trees, Logistic Regression, Neural Networks, and Random Forest, we implemented 10-fold cross-validation in our evaluation process. It is useful in scenarios where the aim is to predict the outcome of a target variable and understand how good the model performs on an independent dataset.

In 10-fold cross-validation, the data is randomly divided into 10 equal parts. During each iteration, one part is retained as the validation data for testing the model, and the remaining nine parts are used as training data. The process is repeated 10 times, with each of the 10 parts used exactly once as the validation data. This approach allows every observation from the original dataset to be used for both training and validation, and each observation is used for validation exactly once.

This method is beneficial because it provides a thorough insight into the model's performance, minimizing the potential bias that could come from a single train-test split. By using k-fold cross-validation, we can confidently assess the generalizability of our models over different subsets of data, ensuring that our models are not only accurate but also stable across various slices of the dataset.

Testing and Evaluation Metrics:

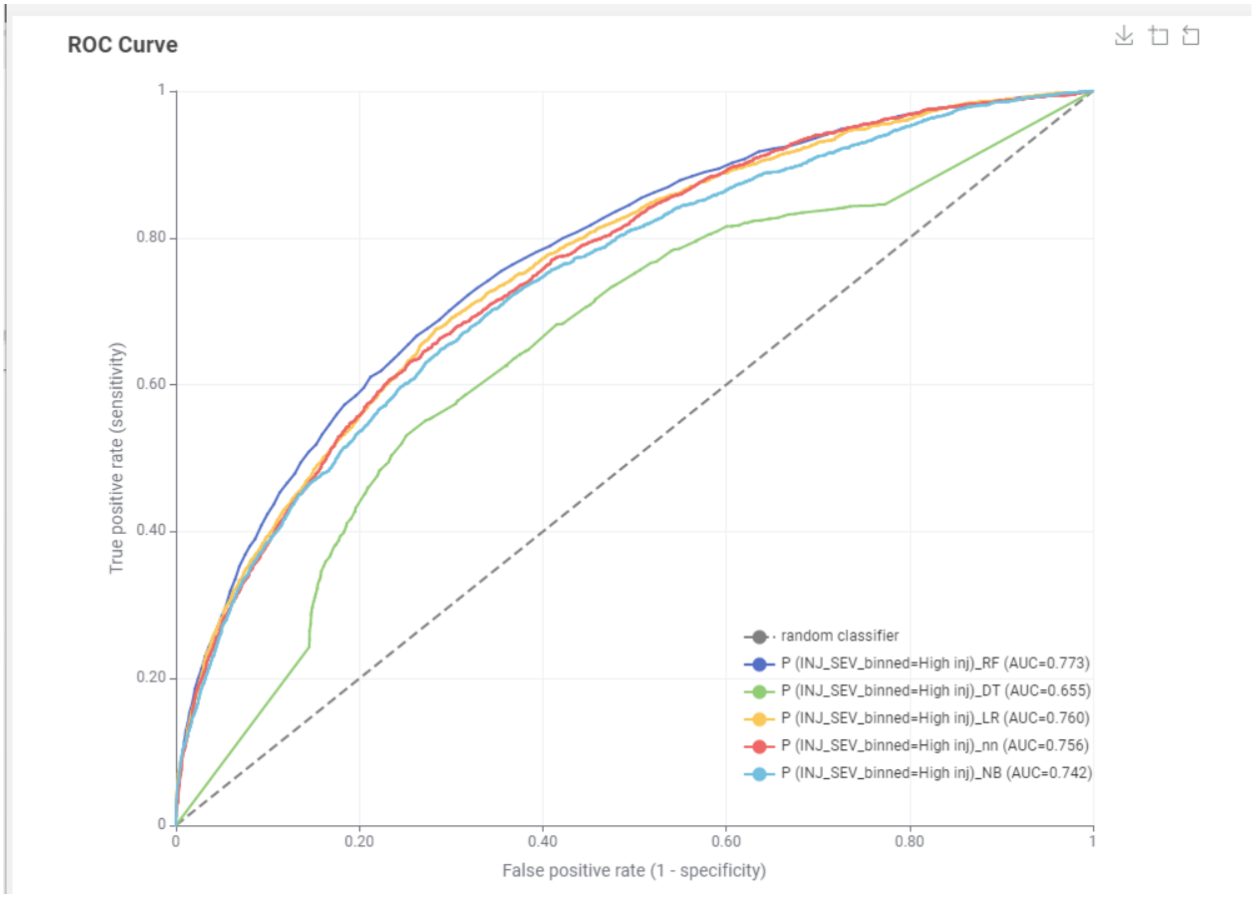
Accuracy = (TP +TN)/(TP +TN +FP+ FN)

Sensitivity= (TP)/(TP + FN)

Specificity= (TN)/(TN +FP)

AOC (Area under the curve)= [Sensitivity,1-Specificty]

Model	Accuracy	Sensitivity	Specificity	AUC value from ROC
Decission Tree	0.641	0.625	0.634	0.655
Random forest	0.702	0.701	0.702	0.773
Artificial Neural network	0.647	0.733	0.625	0.756
Logistic regression	0.656	0.743	0.634	0.760
Naïve Bias	0.731	0.571	0.771	0.742



Based on the measures like Accuracy, sensitivity, Specificity and AOC the Random Forest model performed well followed by logistic regression and then Neural network, Naïve bias whereas Decision tree performed the least in classifying the severity of injury. Although Naïve Bias model has the highest accuracy, the Area under the curve is smaller compared to Random forest, Logistic regression and MLP. Since the aim is to identifying factors for high injury severity random forest is considered as the best model

Variable Importance and Sensitivity analysis:

Upon developing four distinct models for predicting injury severity, our project's primary goal was to unearth the factors most indicative of high injury severity outcomes. The predictive power of Decision Trees, Logistic Regression, Neural Networks, and Random Forest was thoroughly assessed. Among these, the Random Forest model stood out, delivering the highest accuracy in our tests. Given its superior performance, we opted to delve deeper into the Random Forest model to understand which factors it deemed most crucial in forecasting high injury severity.

To visualize the significance of different variables as understood by the Random Forest model, we created a variable importance graph. This graph reflects a weighted average, considering both the frequency of a variable's selection across the forest and its presence at the top levels of the trees, where the most decisive splits occur. The resulting figure, a bar chart showcased within our report, highlights these dominant variables. It offers a clear illustration of which factors the Random Forest model repeatedly identified as pivotal when classifying the severity of injuries.

The variable importance chart, derived from our Random Forest model, provides a visual representation of the factors most associated with high injury severity in vehicle accidents. The analysis highlights a set of variables that stand out for their significant roles.

In our study, we sought to identify the key variables by conducting a comparative analysis(fig 8) of the model's variance. This entailed evaluating the variance with the full set of variables against the variance observed when each individual variable was excluded. This process was systematically applied to each column, and we computed the absolute difference in variances to ascertain the impact of each variable's removal. A significant change in variance pointed to a variable's importance within the model. Based on these variance differentials, we established a hierarchy of variable importance, enabling us to pinpoint the most influential factors in the model's predictive capability.

However, we found the results from this approach (fig 8) are different from the variable importance from random forest, the difference in feature significance order observed between the variance difference method and random forest's variable importance can be attributed to the distinct ways these methods measure feature importance. The variance difference method assesses how the absence of a feature affects the model's stability or variance, focusing on individual feature impact. In contrast, random forest's variable importance is based on how much a feature improves the model's predictions across multiple trees, capturing not just individual feature effects but also complex interactions between features. Therefore, these methods can provide different perspectives on which features are most influential in the dataset. After comparing both methods, we decided to use the variable importance from the random forest. This method takes into account how different features work together, which we think gives a fuller picture of what's important in our data. It's a strong way to see which features really matter for our predictions.

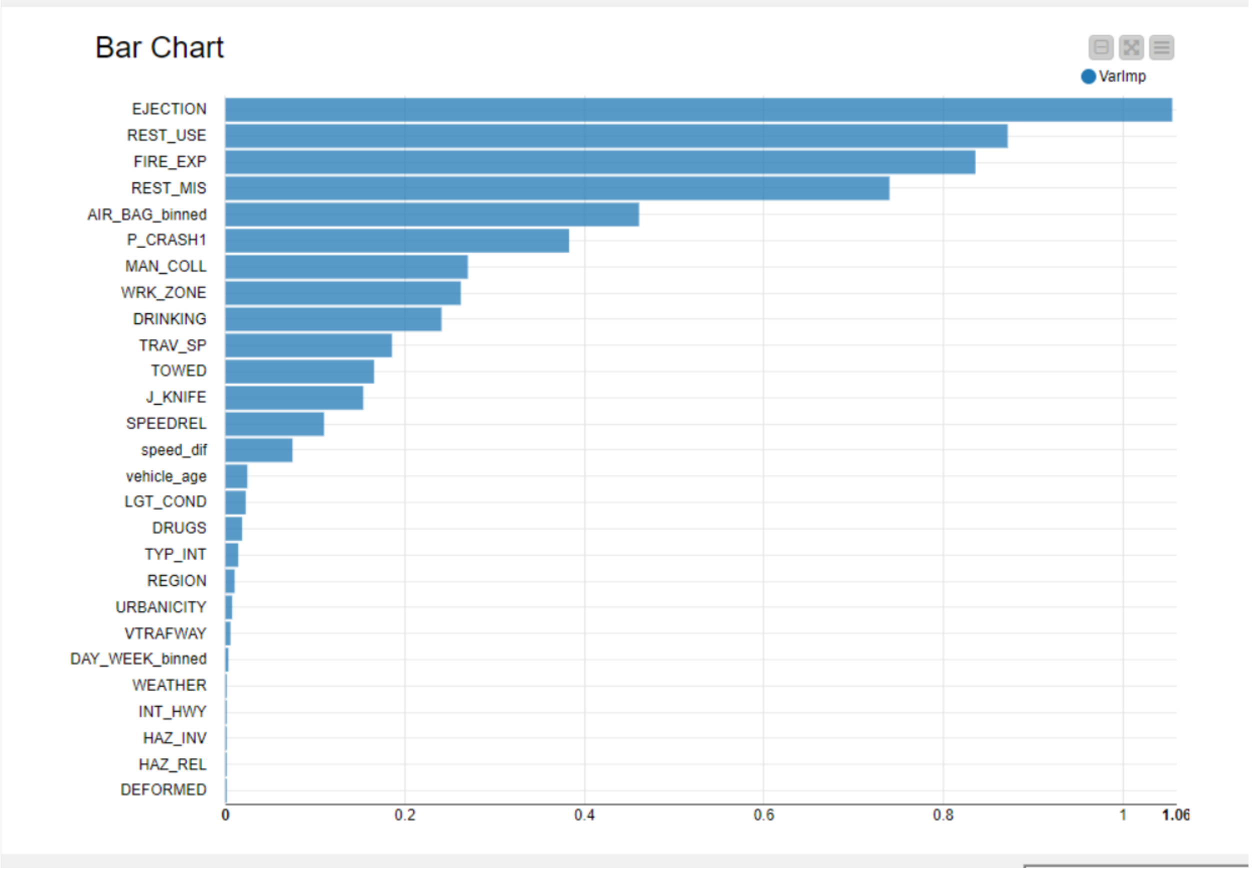
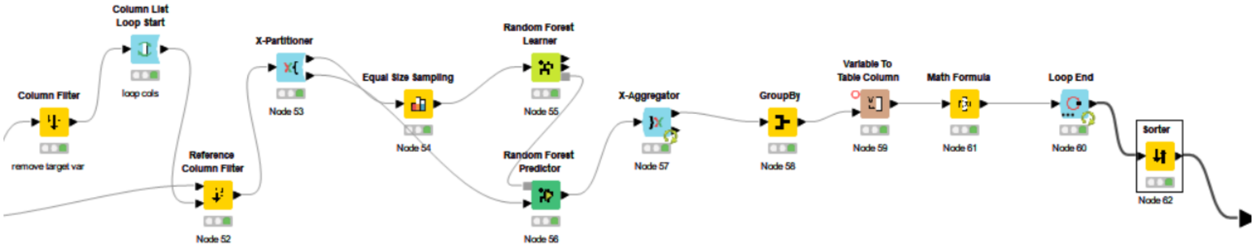
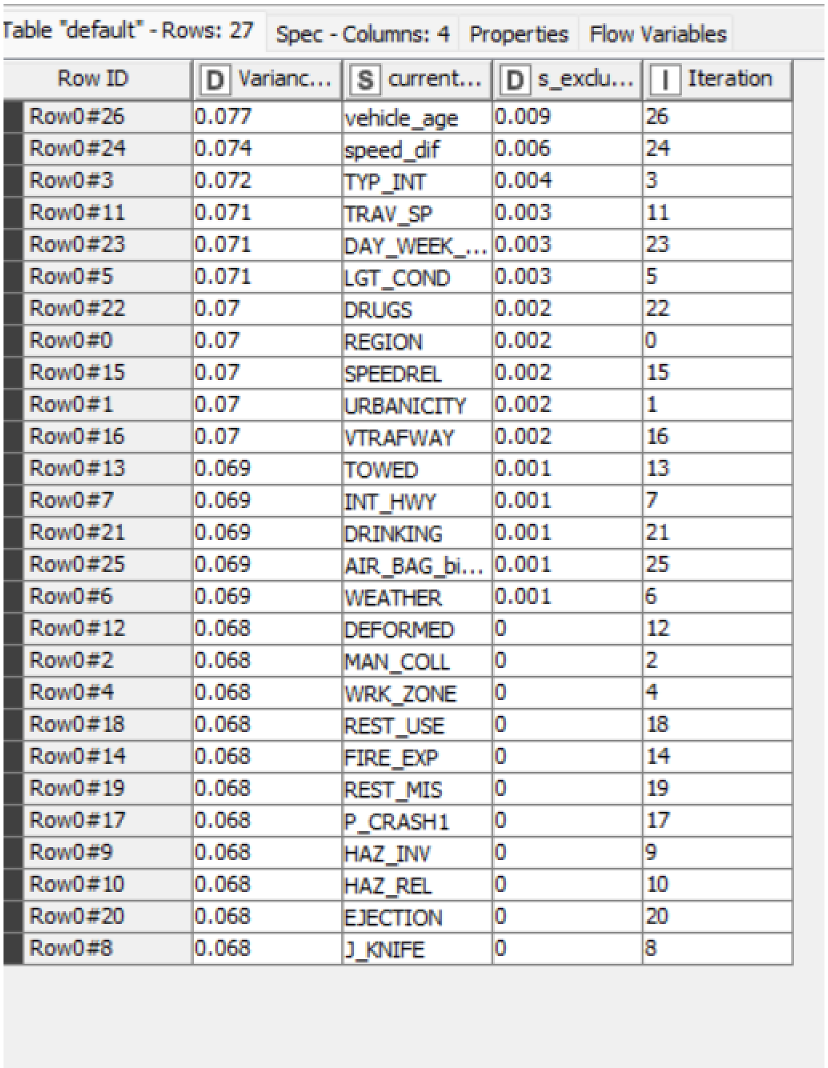


Fig 7. Variable importance





Row ID	D Varianc...	S current...	D s_exclu...	I Iteration
Row0#26	0.077	vehicle_age	0.009	26
Row0#24	0.074	speed_dif	0.006	24
Row0#3	0.072	TYP_INT	0.004	3
Row0#11	0.071	TRAV_SP	0.003	11
Row0#23	0.071	DAY_WEEK_...	0.003	23
Row0#5	0.071	LGT_COND	0.003	5
Row0#22	0.07	DRUGS	0.002	22
Row0#0	0.07	REGION	0.002	0
Row0#15	0.07	SPEEDREL	0.002	15
Row0#1	0.07	URBANICITY	0.002	1
Row0#16	0.07	VTRAFWAY	0.002	16
Row0#13	0.069	TOWED	0.001	13
Row0#7	0.069	INT_HWY	0.001	7
Row0#21	0.069	DRINKING	0.001	21
Row0#25	0.069	AIR_BAG_bi...	0.001	25
Row0#6	0.069	WEATHER	0.001	6
Row0#12	0.068	DEFORMED	0	12
Row0#2	0.068	MAN_COLL	0	2
Row0#4	0.068	WRK_ZONE	0	4
Row0#18	0.068	REST_USE	0	18
Row0#14	0.068	FIRE_EXP	0	14
Row0#19	0.068	REST_MIS	0	19
Row0#17	0.068	P_CRASH1	0	17
Row0#9	0.068	HAZ_INV	0	9
Row0#10	0.068	HAZ_REL	0	10
Row0#20	0.068	EJECTION	0	20
Row0#8	0.068	J_KNIFE	0	8

Fig 8.a. effect of single variable elimination.

At the forefront, 'Ejection' emerges as a critical factor. It represents the status of a person being ejected from the vehicle during a crash. This could range from not being ejected, partially ejected, to completely ejected, with each scenario contributing differently to the severity of injury sustained.

'Rest_Use', indicating the type of restraint equipment utilized by the occupants at the time of the crash, also shows substantial importance. This includes the use of safety features such as shoulder belts and lap belts. The presence and proper use of these restraints can play a decisive role in mitigating injury severity. 'Fire_Exp', which indicates whether there was a fire or explosion as a result of the crash, and 'Rest_Mis', describes any misuse of restraint equipment, have been identified.

The 'Air_bag_binned' variable captures whether airbags were deployed in the vehicles involved in the crash. Airbag deployment is a critical safety feature that can significantly alter the outcome of an accident, making it a key factor in assessing injury severity. 'P_CRASH1' provides insight into the driver's activity prior to the crash. This could range from safe driving practices to potentially dangerous behaviors, which can greatly affect the risk of severe injuries. 'MAN_COLL', or the manner of collision, describes how the vehicles impacted one another. This includes angles and points of impact, which are vital in understanding the dynamics and force involved in a crash. 'WRK_ZONE' indicates whether the crash occurred in or near a construction or work zone. Accidents in such areas can be particularly hazardous due to the presence of workers, equipment, and altered traffic patterns. Lastly, 'DRINKING' denotes whether alcohol consumption was a factor in the accident. Driving under the influence is a known risk factor that can exacerbate the chances of high injury severity due to impaired judgment and reaction times.

'TRAV_SP', indicating the traveling speed of the vehicle before the crash, is a critical factor. It's well-established that higher speeds can lead to more severe injuries, and this variable captures that aspect. 'TOWED' provides details on how the vehicle was removed from the scene, which can give indirect insights into the crash's severity. For instance, vehicles that are towed away are often substantially damaged, which can correlate with more serious injuries. The occurrence of a 'J_Knife' event, where a vehicle experiences a jackknife during an unstable situation, is another variable of note. This particular incident often results in more complex crash scenes and potentially higher injury severities due to the unpredictable nature of such accidents. 'SPEED_REL' takes into account the relative speed of the vehicle, categorizing it as traveling at high or low speeds, which affects both the risk and outcome of crashes. Lastly, 'SPEED_DIF' specifically quantifies how much the vehicle's speed exceeded the speed limit, providing a precise measure of speeding behavior. These variables, focused on the speed and control of the vehicle, contribute significantly to our model’s ability to predict injury severity.

In addition to the dominant factors previously discussed, our Random Forest model identified several other variables that contribute to injury severity, with less significance compared to the primary factors. 'Vehicle Age' accounts for the model year of the vehicle relative to the year of the accident, with older vehicles potentially lacking the latest safety features. 'Light Conditions' capture the visibility at the time of the crash, which can influence driver reaction time and the severity of an accident. The 'Region' variable might reflect geographic differences in road conditions or traffic laws that can impact crash outcomes.

'Drug Involvement' in the crash, while less significant than alcohol in our findings, still plays a role in assessing the risk factors. The 'Day of the Week' can indicate patterns in traffic volume or driver behavior that vary between weekdays and weekends. 'Weather' conditions at the time of the crash can drastically affect road safety, although in our model, its impact was less pronounced compared to other variables. Lastly, whether the vehicle was 'Carrying Hazardous Elements' could point to additional risks in the event of a crash, but again, its influence was relatively lower in our predictive model. While these variables showed less importance in our analysis, they still provide valuable context and should not be overlooked when considering the full scope of factors that can affect injury severity in automobile accidents.

Deployment:

Based on our findings, the Random Forest model has emerged as the most effective tool for predicting injury severity in vehicular accidents, due to its ability to account for complex feature interactions and provide a nuanced understanding of variable importance.

Should this model approach be considered for real-world application, the deployment phase would involve several critical steps:

The model would be integrated with present traffic management and vehicle safety systems, to ensure seamless communication between the model and the data sources. Deployment would also include mechanisms for the model to learn from new data over time, enhancing its accuracy and adaptability to changing conditions and emerging trends in road safety.

Summary and Conclusion:

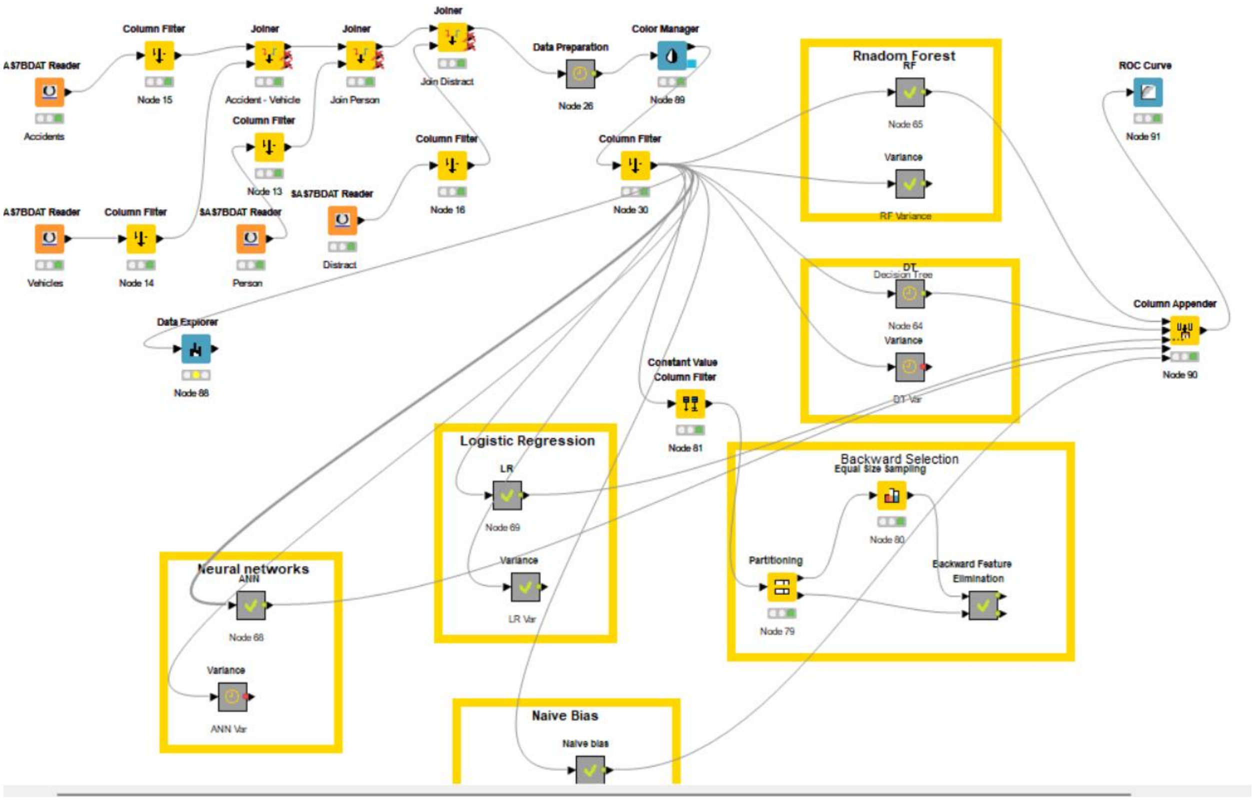
In our project, we set out to understand the factors that lead to severe injuries in car crashes involving drivers. We explored various predictive models, including Decision Trees, Logistic Regression, Neural Networks, and Random Forest, to determine which model worked best for this task. To ensure the reliability of our predictions, we employed a robust 10-fold cross-validation approach, which involves testing our models in ten different ways.

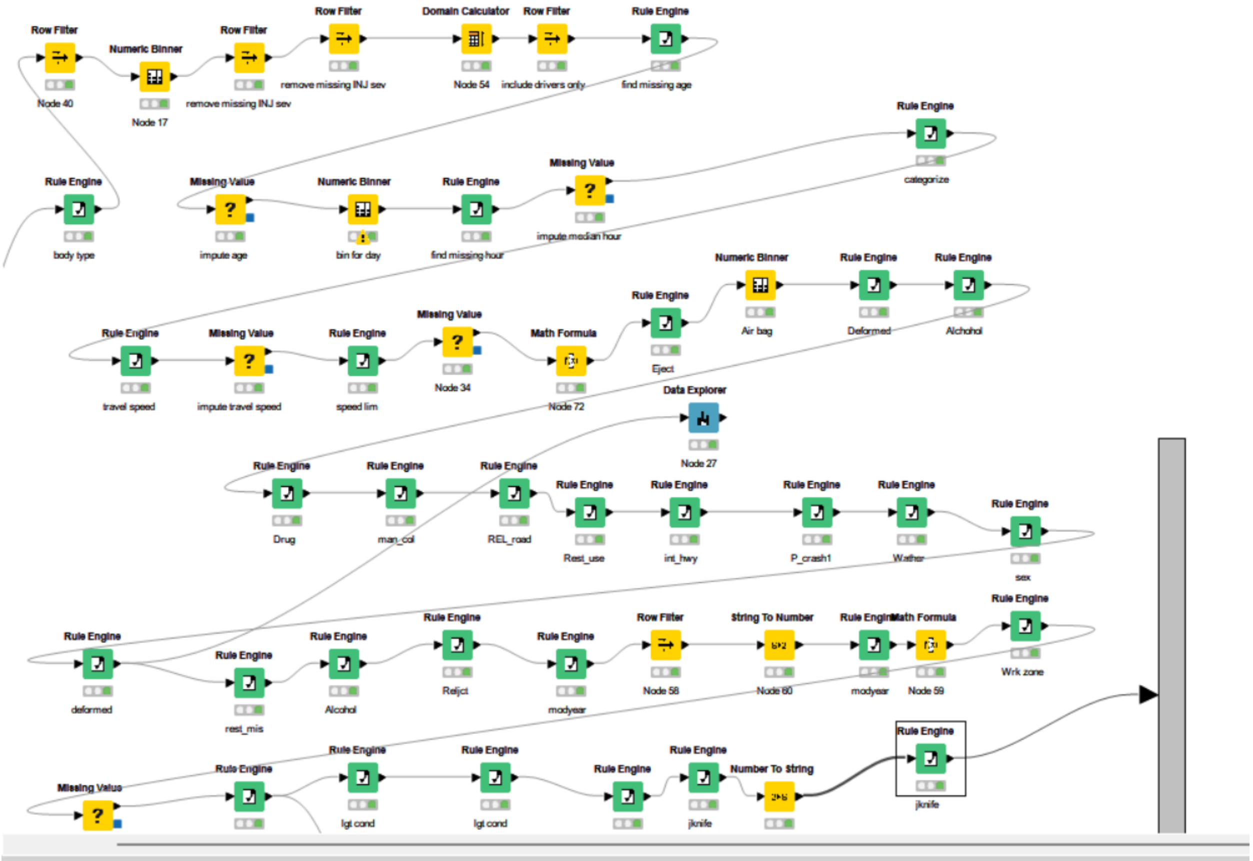
Among all the models we tested, the Random Forest model stood out as the most effective in predicting the severity of injuries sustained by drivers in car crashes. This model allowed us to identify the key factors that significantly influence injury severity.

For instance, we discovered that whether a person was ejected from the vehicle, whether they were using a seatbelt, and the functionality of the car's airbags were crucial factors in determining the severity of injuries. Additionally, factors such as the driver's activity before the crash, the manner of collision, drug involvement, and whether the crash occurred in a work zone also played important roles.

While these findings are certainly interesting, their practical implications are even more significant. They can inform the design of improved safety features in vehicles, promote awareness of safe driving practices, and help authorities plan for effective emergency responses to accidents. Our project underscores the power of data-driven insights in enhancing road safety and protecting drivers from severe injuries.

Complete Workflow:


A S7BDAT Reader
Accidents
Column Filter
Node 15
Joiner
Accident - Vehicle
Joiner
Join Person
Joiner
Join Distract
Data Preparation
Node 26
Color Manager
Node 89
Column Filter
Node 13
Column Filter
Node 16
Column Filter
Node 30
A S7BDAT Reader
Vehicles
Column Filter
Node 14
$A S7BDAT Reader
Person
$A S7BDAT Reader
Distract
Data Explorer
Node 88
Rnadom Forest
RF
Node 65
Variance
RF Variance
ROC Curve
Node 91
DT
Decision Tree
Node 64
Variance
DT Var
Column Appender
Node 90
Constant Value Column Filter
Node 81
Logistic Regression
LR
Node 69
Variance
LR Var
Backward Selection
Equal Size Sampling
Node 80
Partitioning
Node 79
Backward Feature Elimination
Neural networks
ANN
Node 68
Variance
ANN Var
Naive Bias
Naive bias


Row Filter
Node 40
Numeric Binner
Node 17
Row Filter
remove missing INJ sev
Row Filter
remove missing INJ sev
Domain Calculator
Node 54
Row Filter
include drivers only
Rule Engine
find missing age
Rule Engine
categorize
Rule Engine
body type
Missing Value
impute age
Numeric Binner
bin for day
Rule Engine
find missing hour
Missing Value
impute median hour
Numeric Binner
Air bag
Rule Engine
Deformed
Rule Engine
Alchohol
Rule Engine
travel speed
Missing Value
impute travel speed
Rule Engine
speed lim
Missing Value
Node 34
Math Formula
Node 72
Rule Engine
Eject
Data Explorer
Node 27
Rule Engine
Drug
Rule Engine
man_col
Rule Engine
REL_road
Rule Engine
Rest_use
Rule Engine
int_hwy
Rule Engine
P_crash1
Rule Engine
Wather
Rule Engine
sex
Rule Engine
deformed
Rule Engine
rest_mis
Rule Engine
Alcohol
Rule Engine
Reljct
Rule Engine
modyear
Row Filter
Node 58
String To Number
Node 60
Rule Engine
modyear
Math Formula
Node 59
Rule Engine
Wrk zone
Missing Value
Rule Engine
Rule Engine
lgt cond
Rule Engine
lgt cond
Rule Engine
Rule Engine
jknife
Number To String
Rule Engine
jknife