

# **MSIS 5633-Predictive Analytics Technologies**

## **Driver injury severity analysis in automobile crashes**

**Due Date:**

**Nov 26, 2023**

**By**

**Team 09**

**Naveen Varma Patsamatla**

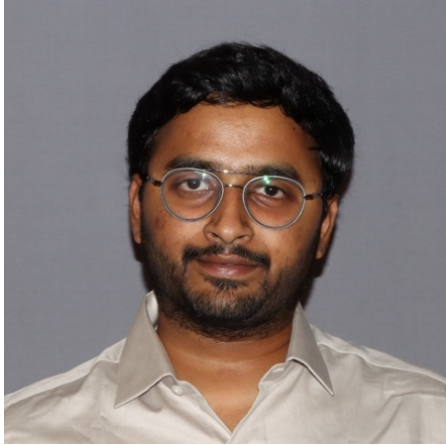
**Deva Darshita Pillai**

# Table of Contents

<b>Executive Summary</b> .....	5
<b>CRISP-DM Methodology</b> .....	6-7
<b>Business Understanding</b> .....	7-9
A. Business Objective.....	8-9
B. Data Mining Goals.....	9
C. Project Planning.....	9
<b>Data Understanding</b> .....	9-14
A. Data Collection.....	10
B. Describing Data.....	11
C. Exploring Data.....	12-13
D. Verifying Data Quality.....	14
<b>Data Preparation</b> .....	14-24
A. Selecting Data.....	14-16
1. Row Filtering	
2. Feature Selection	
B. Cleaning Data.....	16-22
1. Data Replacement	
2. Data Type Conversion	
3. Imputations	
4. Binning/ Grouping	
5. Handling Outliers	

C. Constructing Data.....	22
D. Formatting Data.....	23-24
<b>Modeling.....</b>	<b>24-29</b>
A. Selecting Model Techniques.....	25
B. Generate Test Design.....	25
C. Building Model.....	25-29
1. Decision Tree	
2. Random Forest	
3. Gradient Boosted Tree	
4. Naïve Bayes	
5. Artificial Neural Networks – MLP	
6. Logistic Regression	
<b>Evaluation.....</b>	<b>30-39</b>
A. Evaluation Results.....	30-31
B. Combined ROC.....	32
C. Variable Importance using Sensitivity Analysis.....	32-38
D. Decision Tree Branches.....	39
<b>Deployment.....</b>	<b>39-40</b>
<b>Conclusion.....</b>	<b>40-41</b>

## **Team Members**



**Naveen Varma Patsamatla (A20377440)**



**Deva Darshita Pillai (A20392692)**

## **Executive Summary**

Following a directive to examine risk factors for injury severity in car crashes, our team conducted a thorough analysis. Leveraging an authentic crash dataset from the U.S. Department of Transportation, sourced specifically from the National Highway Traffic Safety Administration, our focus was on approximately one percent of all reported domestic (U.S.) automobile crashes. To ensure methodological consistency, we homogenized the data into driver oriented and restricted ourselves to automobiles. Later we have use data balancing techniques to ensure our models are not biased. We employed six prevalent predictive analytics algorithms—decision tree, random forest, gradient boosted tree multi-layered perceptron, logistic regression, and naive bayes classifier—within the KNIME application, and we delved into the complex relationships between injury severity levels and associated risk factors.

Our analysis peaked with a variable importance using sensitivity analysis study, showing us which crash-related factors matter most for different injury severity levels. This helped us grasp how each risk factor contributes to the predicted outcomes in a clear and detailed way.

The information we discovered using predictive analytics is valuable. It helps us see new possibilities in the data. As technology improves and safety measures grow, we think our findings can be crucial for avoiding crashes and reducing injuries when accidents happen.

With more vehicles on the road and rising driver distractions, our goal is to make a real difference by sharing these findings. We want to help turn the tide of growing severe injuries from accidents, creating a safer road. Our thorough analysis makes these insights valuable for anyone dedicated to improving road safety.

## CRISP – DM Methodology

CRISP-DM, which stands for Cross-Industry Process for Data Mining, presents a meticulously structured approach for orchestrating data mining projects. This method, shaped through collaboration with over 200 organizations, extends itself as an open standard, accessible for universal use. Originally designed with a focus on data mining, its versatility empowers its application across diverse analytical styles and methodologies. The added advantage of this methodology is we can move from one phase to any other phase with ease. For our Analysis we are using this methodology

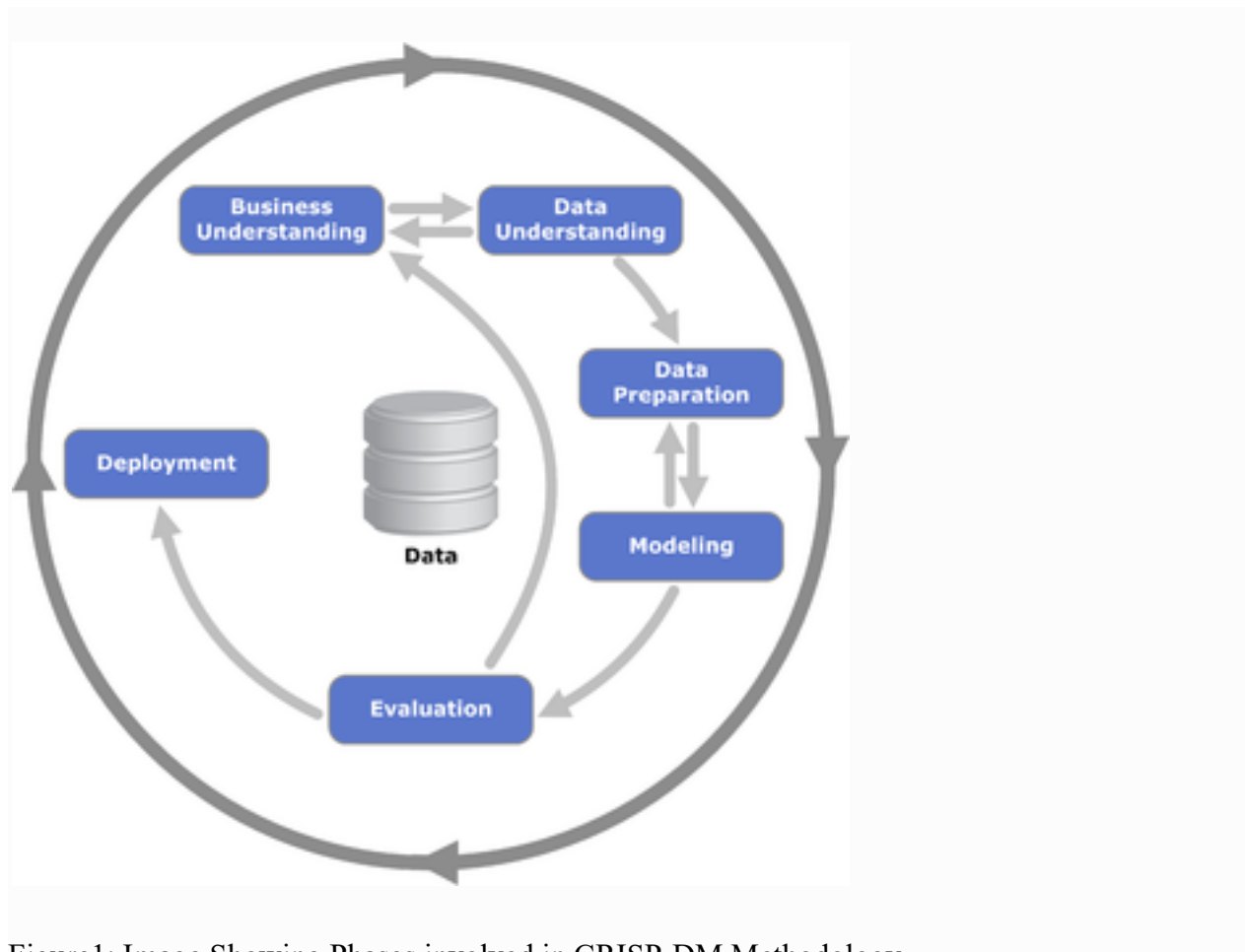


Figure1: Image Showing Phases involved in CRISP-DM Methodology

The CRISP-DM process model unfolds through six key phases:

- **Business Understanding:** This initial phase involves gaining a profound understanding of the business problem at hand, addressing its goals and objectives comprehensively.
- **Data Understanding:** In this stage, a thorough examination of the data is conducted to enhance understanding and evaluate its characteristics.
- **Data Preparation:** The data is prepared and modified to a state suitable for analysis, ensuring it is well-suited for subsequent stages.
- **Modeling:** Analytical techniques are developed in this phase to construct a robust model, one that can effectively address the identified business problem.
- **Evaluation:** Models generated in the previous phase are compared, and the best-performing one is selected for implementation.
- **Deployment:** The final model is presented to the business in a manner that facilitates seamless integration into everyday operations, ensuring practical applicability.

## Phase 1: Business Understanding

Transportation, often hailed as the lifeblood of the economy, has seen road transport play a pivotal role in our evolutionary journey. Despite incredible technological strides, the challenge of road safety persists, costing the U.S. a staggering \$800 billion annually in road accidents. With over 5 million road accidents annually for the past decade, the National Highway Traffic Safety Administration (NHTSA) projects a concerning 10.5% increase in motor vehicle crash fatalities, reaching 42,915 in 2021 a record not seen since 2005, marking the highest annual percentage surge in the history of the Fatality Analysis Reporting System [1] .

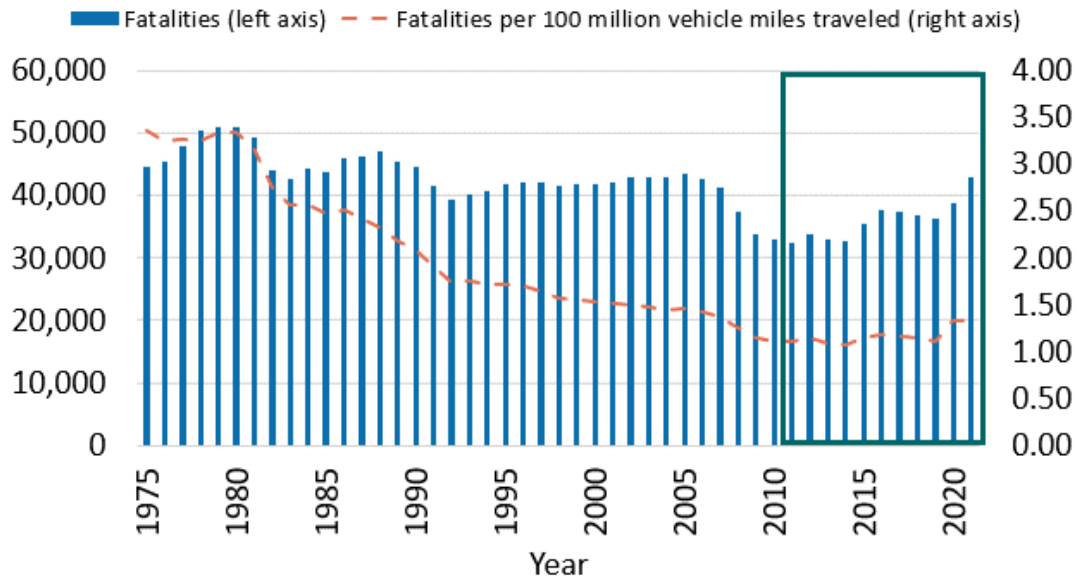


Figure2: Image Showing fatality rate progression over the last 30 years

While we've witnessed an overall decline in car crash-related deaths for last 30 years due to improved road infrastructure and automobile safety features, but progress have been stalled over the last decade. Regulations have played a pivotal role in driving down fatality rates. In our pursuit of this mission, our analysis taps into the wealth of insights within the General Estimates System (GES) database from NHTSA. This data goldmine, drawn from thousands of accidents nationwide, is a crucial asset for uncovering the causes of severe injuries in car crashes. These insights become a catalyst for industry leaders striving for safety innovations and governments making policy decisions rooted in data. Our analysis involves the exciting task of classifying injury severity and unraveling the factors tied to specific injury classes, including drivers' control abilities and behavior, safety restraint usage, external conditions, crash types, driver demographics, vehicle related characteristics and other critical factors. Let's dive into the data to drive progress and make our roads safer!

### A) Business Objective:

- Uncover and understand the various factors that significantly impact the severity of injuries sustained by drivers.



- Provide valuable suggestions and recommendations derived from the insights gleaned through our analysis, contributing to informed decision-making and potential improvements in safety measures or practices.

## B) Data Mining Goals:

- Utilizing analytical methods to identify and highlight the variables that play a pivotal role in determining the severity of injuries experienced by drivers. Here we used sensitivity analysis to explain the variable importance.
- Creating a better analytics model with the existing dataset to forecast the probability of injury severity, providing valuable insights for informed decision-making and safety enhancement strategies.
- In pursuit of our goals, we've embraced the **Knime Data Analytics platform**. Its impressive flexibility empowers us to stay focused on problem-solving rather than getting entangled in coding complexities.

## C) Project Planning:

Phase	Duration	Challenges
Business understanding	1 week	Lack of understanding
Data understanding	1 week	Data related problems, Domain Knowledge
Data preparation	2 weeks	Data Quality issues and cleaning
Modeling	1 week	finding good and balanced model
Evaluation\Presentation	1 week	Variables impacting the results

Table1 : Table Showing the project plan

## Phase 2: Data Understanding

Essentially, Data Understanding represents a critical initial stage in the data analysis process, involving the collection, description, exploration, and validation of data to prepare for more comprehensive analysis and modeling.

## A) Data Collection

Initiating our analysis, data is consolidated from the National Highway Traffic Safety Administration (NHTSA) in the form of four distinct SAS files (Accident, Vehicle, Person, Distract). These files contain information on road conditions, environmental impact, vehicles, persons, and distractions related to accidents. The dataset comprises Accident (54,200 records, 46 attributes), Vehicle (95,785 records, 88 attributes), Person (133,734 records, 59 attributes), and Distract (98,845 records, 11 attributes). Given the common occurrence of accidents involving vehicles and passengers, the KNIME tool utilizes Joiner nodes to merge these files into a unified dataset. Left joins are executed based on case numbers and vehicle numbers, as illustrated in Figure 1. Following data consolidation, the resultant dataset consists of person-level records, with one record per individual involved in a reported automobile crash. At this stage, prior to data cleaning, preprocessing, and slicing/dicing, the comprehensive dataset encompasses 128,589 unique person and incorporates 204 attributes in the final dataset.

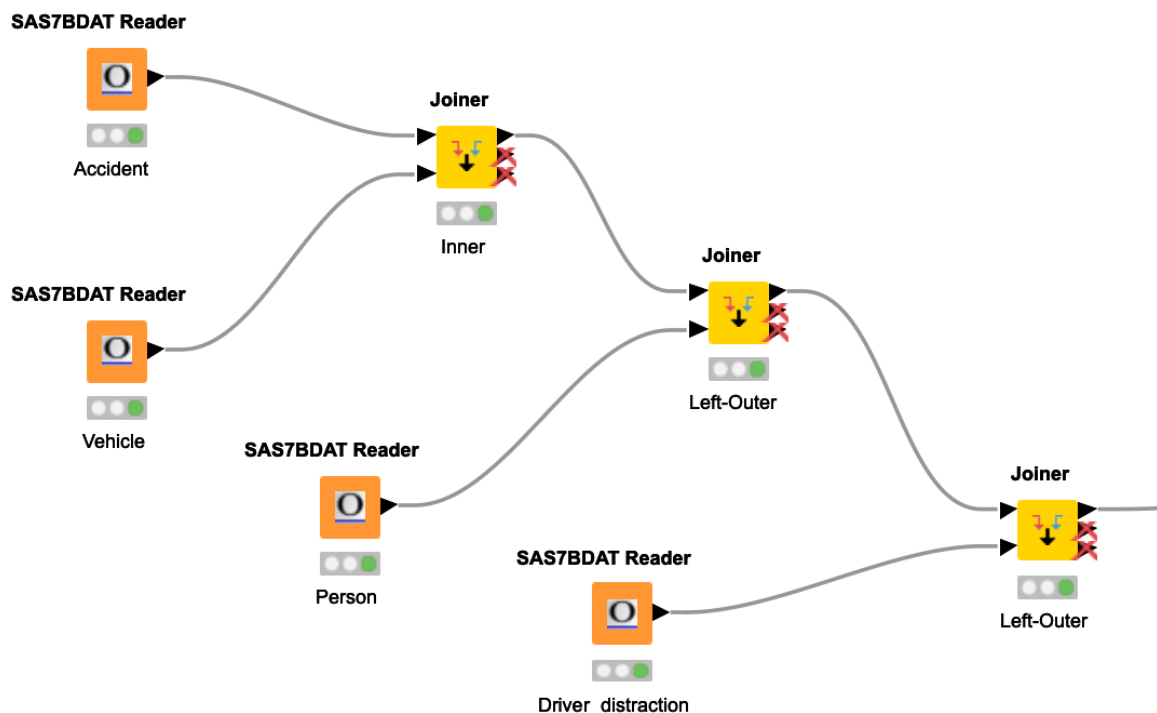


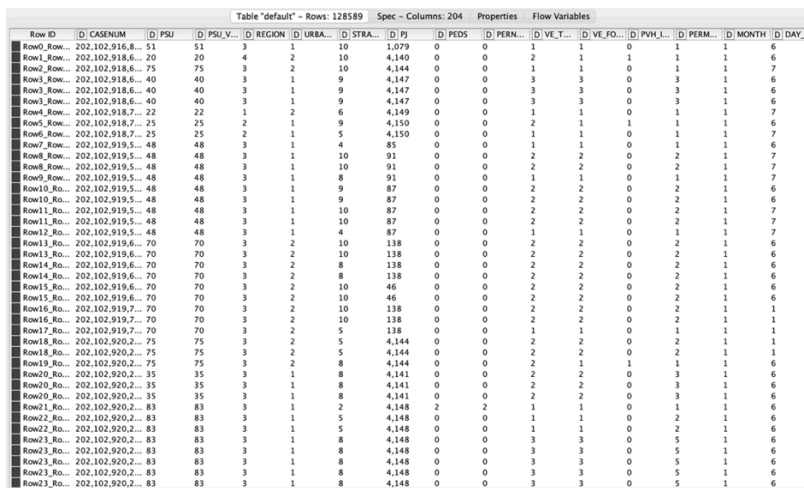
Figure3: Image Showing Joining different sas files using Joiner nodes to create a single final flat file

## B) Describing the Data:

In the second phase of our analysis, we present an overview of the dataset, encompassing 128,589 records and comprising 204 features or attributes. Notably, numerous duplicate columns resulting from the joins are identified, necessitating their removal in subsequent stages of the analysis. These attributes are categorized into two main groups: 198 are numeric features, while 6 are nominal, binomial features.

A crucial aspect of this task involves data transformation, as Knime defaults to treating number columns as numeric and string columns as nominal datatypes, a convention that does not align with our specific scenario. Consequently, we undertake the transformation of data into their appropriate datatypes, as outlined in the CRSS Analytical User's Manual (2016-2021).

The dataset itself contains information pertaining to accidents, including impact details, road surface conditions, weather conditions, light conditions, time of day, and month of the year. Additionally, it encompasses technical characteristics of the vehicles involved, such as the age and body type of the vehicle. Information related to individuals involved in the accidents is also present, covering demographics, injury-related data, and situational context. Furthermore, the dataset captures details about any distractions affecting the drivers.



Row ID	CASENUM	PSU	REGION	URBAN	STRA	PJ	PEDS	PERM	VE	T	VE	FO	PVH	L	PERM	MONTH	DAY
Row0_Row...	202.102.916.8...	51	51	3	1	10	1,079	0	0	1	1	0	1	1	1	6	
Row1_Row...	202.102.918.6...	20	20	4	2	10	4,140	0	0	2	1	1	1	1	1	6	
Row2_Row...	202.102.918.6...	75	75	3	2	10	4,144	0	0	1	1	0	1	1	1	7	
Row3_Row...	202.102.918.6...	40	40	3	1	9	4,147	0	0	3	3	0	3	1	6		
Row4_Row...	202.102.918.6...	40	40	3	1	9	4,147	0	0	3	3	0	3	1	6		
Row5_Row...	202.102.918.7...	22	22	1	2	6	4,149	0	0	1	1	0	1	1	7		
Row6_Row...	202.102.918.7...	25	25	2	1	5	4,150	0	0	2	1	1	1	1	6		
Row7_Row...	202.102.919.5...	48	48	3	1	4	85	0	0	1	1	0	1	1	6		
Row8_Row...	202.102.919.5...	48	48	3	1	10	91	0	0	2	2	0	2	1	7		
Row9_Row...	202.102.919.5...	48	48	3	1	8	91	0	0	1	1	0	1	1	7		
Row10_Row...	202.102.919.5...	48	48	3	1	9	87	0	0	2	2	0	2	1	6		
Row11_Row...	202.102.919.5...	48	48	3	1	10	87	0	0	2	2	0	2	1	7		
Row12_Row...	202.102.919.5...	48	48	3	1	4	87	0	0	1	1	0	1	1	7		
Row13_Row...	202.102.919.6...	70	70	3	2	10	138	0	0	2	2	0	2	1	6		
Row14_Row...	202.102.919.6...	70	70	3	2	8	138	0	0	2	2	0	2	1	6		
Row15_Row...	202.102.919.6...	70	70	3	2	10	46	0	0	2	2	0	2	1	6		
Row16_Row...	202.102.919.7...	70	70	3	2	10	138	0	0	2	2	0	2	1	1		
Row17_Row...	202.102.919.7...	70	70	3	2	5	138	0	0	1	1	0	1	1	1		
Row18_Row...	202.102.920.2...	75	75	3	2	5	4,144	0	0	2	2	0	2	1	1		
Row19_Row...	202.102.920.2...	75	75	3	2	8	4,144	0	0	2	1	1	1	1	6		
Row20_Row...	202.102.920.2...	35	35	3	1	8	4,141	0	0	2	2	0	2	1	6		
Row21_Row...	202.102.920.2...	35	35	3	1	8	4,141	0	0	2	2	0	3	1	6		
Row22_Row...	202.102.920.2...	83	83	3	1	5	4,148	0	0	1	1	0	2	1	6		
Row23_Row...	202.102.920.2...	83	83	3	1	8	4,148	0	0	3	3	0	5	1	6		
Row24_Row...	202.102.920.2...	83	83	3	1	8	4,148	0	0	3	3	0	5	1	6		
Row25_Row...	202.102.920.2...	83	83	3	1	8	4,148	0	0	3	3	0	5	1	6		
Row26_Row...	202.102.920.2...	83	83	3	1	8	4,148	0	0	3	3	0	5	1	6		
Row27_Row...	202.102.920.2...	83	83	3	1	8	4,148	0	0	3	3	0	5	1	6		

Figure 4: Image Showing snapshot of the file data before data preprocessing

### C) Exploring the Data:

Here, we delve into an in-depth exploration of the dataset through the application of exploratory data analysis (EDA). This process involves the application of statistical concepts such as mean, median, and mode, alongside the utilization of graphical and visual analysis tools. To enhance our understanding of the data distribution for each feature, we employ measures such as standard deviation, variance, and skewness.

To execute this analysis, we leverage the capabilities of the data explorer node, a potent and valuable tool. Additionally, Boxplots and Numeric Outlier nodes within the Knime analytics tool play a crucial role in this exploration. Through these tools, we address various aspects, including the handling of missing values, identification of impactful outliers and anomalies, homogenization of the data, feature selection for model training, addressing data imbalances, and discerning relationships between attributes and their impact on the dependent variable.

This comprehensive exploratory approach equips us with valuable insights, aiding in the refinement of the dataset and informing subsequent steps in the analysis process.

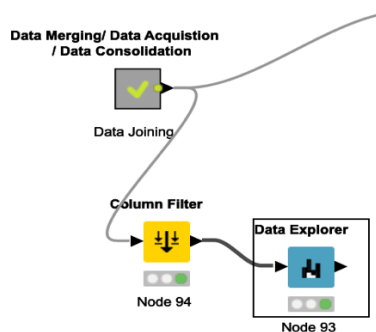


Figure 5: Image Showing Data Explorer node

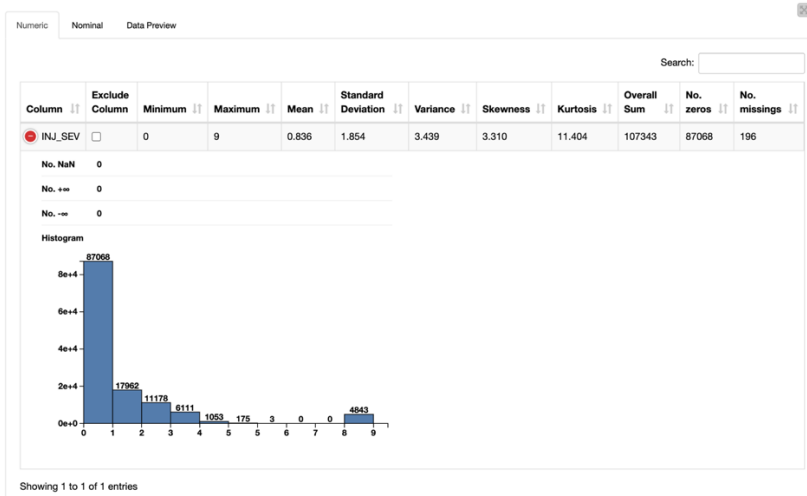


Figure 6: Image Showing Distribution of INJ\_SEV variable before Data Preprocessing in Data Explorer node

The variable INJ\_SEV, identified as the dependent variable within the dataset, holds significant importance. To gain deeper insights into this attribute, we refer to the output of the data explorer node. This source of information provides a clear understanding of the attribute's characteristics. Notably, the data explorer node categorizes our dependent variable under Numeric datatype, while it is, in fact, nominal. This conversion issue is addressed in later sections of the report.

Examining the histogram, we observe that the distribution of the data exhibits right skewness and is platy kurtotic. Most values are concentrated in a specific region with low variance. This information proves valuable in understanding the central tendencies and dispersion of the data for the dependent variable.

Additionally, the data explorer node is instrumental in identifying the count of missing values associated with INJ\_SEV, providing crucial insights into data completeness. This comprehensive overview aids in refining our understanding of the dependent variable, guiding subsequent steps in the analysis, and informing necessary data preprocessing actions.

## D) Verifying Data Quality:

In the fourth task, our evaluation of the data quality reveals several challenges that need to be addressed. Notably, there are numerous missing and null values present within the dataset, posing a significant obstacle to the integrity of our analysis. Additionally, the identification of outliers raises concerns about their potential impact on the robustness of our findings.

Furthermore, a notable observation is the imbalance in the distribution of the target variable, suggesting a potential bias in the model towards the majority class. To mitigate this, it becomes imperative to implement data balancing techniques, ensuring a more equitable representation of different classes in the dataset.

The prevalence of missing values underscores the need for extensive transformations, smoothing, imputations, and cleaning procedures. The data, in its current state, is not optimal, requiring diligent efforts to enhance its quality and reliability for subsequent stages of analysis and modeling. This recognition prompts the initiation of comprehensive data preprocessing measures to rectify these issues and foster a more robust analytical foundation.

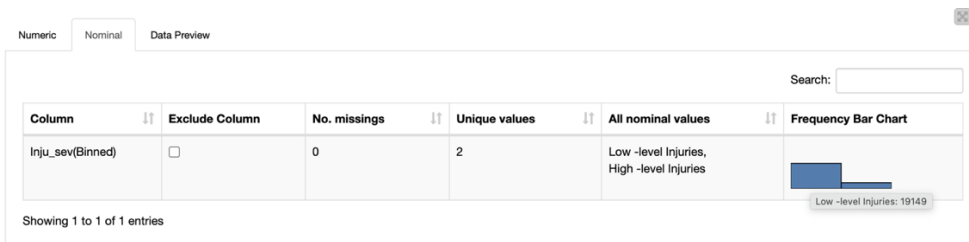


Figure 7: Image Showing data Imbalance of INJ\_SEV after Data Preprocessing in Data Explorer node

## Phase 3: Data Preparation

In this phase, we focus on getting the data ready for training our model. We go through several essential steps, such as cleaning the data, transforming it, and reducing it. This phase is crucial because the quality and quantity of the data have a significant impact on the accuracy of our

predictions. The better the data we provide to our model, the better the predictions it can make. The figure shows complete Data Preparation we implemented

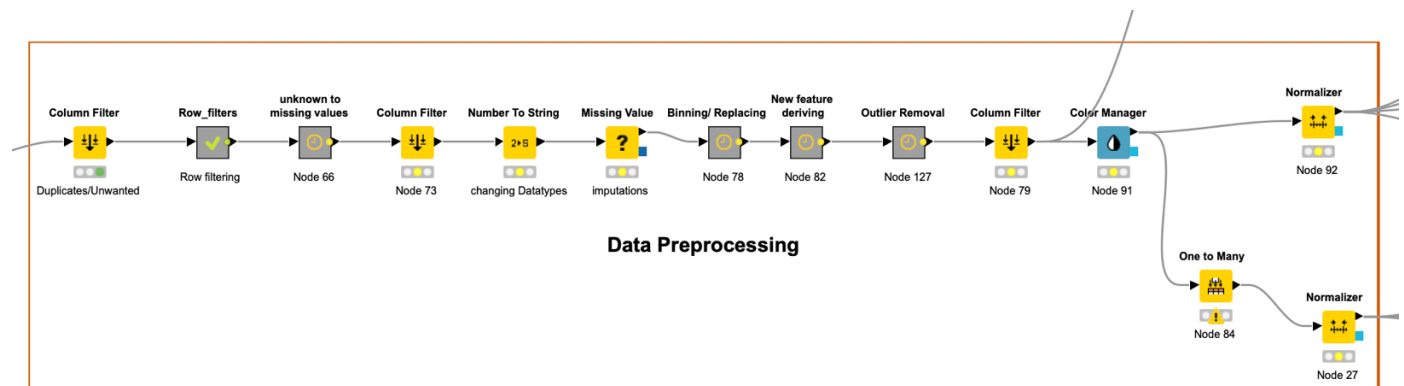


Figure 8: Image Showing Data Preprocessing workflow

## A) Selecting Data:

Selecting the right set of features or variables is a crucial aspect of model training, as incorporating unnecessary or irrelevant data may result in overfitting. Overfitting makes the model too sensitive to noise in the training data, leading to inaccurate results when applied to new data. In our case, out of the initial 204 variables, we carefully choose 30 potential variables.

### 1) Row Filtering:

In the initial steps of data preprocessing, we employed the **Row Filter** node functionality in Knime to enhance the homogeneity of the dataset. Our focus narrowed specifically to drivers in automobiles who sustained some level of injury. To achieve this, we selectively chose person records with seat position (seat\_pos) equal to 11 and person type (per\_typ) equal to 1, indicating the driver category.

Additionally, to maintain our focus on automobiles, we excluded person records associated with non-motorists, such as pedestrians, tricycles, motorcycles, farm-related vehicles, and bicycles.

Ensuring that there was at least some level of damage or injury to the vehicle, we excluded true matches for the variable deformed equal to 0. This step helped filter out instances where no damage occurred to the vehicle.

Finally, to streamline the dataset for our analysis, we removed records where there were no injuries, as well as records of individuals who had died prior to the recorded incident. This process left us with a dataset (23636 rows) specifically tailored to drivers in automobiles who experienced low and high-level injuries, aligning with the focus on injury severity for this subset of the data.

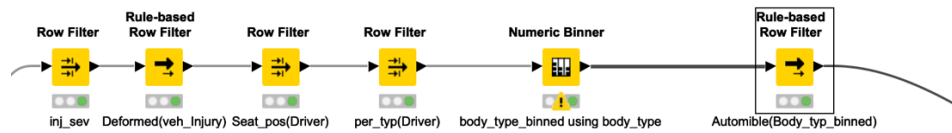


Figure 8: Image Showing removing rows using row filter

## 2) Feature Selection:

The selection process involves employing various statistical measures, forward and backward selection methods, and leveraging both common sense and domain knowledge by eliminating unwanted and not useful features through stages with help of **COLUMN FILTER** in knime. Additionally, insights from the CRSS Manual contribute to the decision-making process. This meticulous approach aims to eliminate unwanted features, ensuring that the chosen variables contribute meaningfully to the model's predictive capabilities. By doing so, we enhance the model's generalization to new data and avoid the pitfalls associated with overfitting. Finally we choose 29 features to train the models.

## B) **Cleaning Data:**

This step involves handling missing data, dealing with outliers (either by imputing or removing them), addressing anomalies, and encoding the data in a way suitable for the model. Additionally, we've performed data normalization, specifically using min-max normalization for numerical predictive models. We also used equal-sampling technique to ensure that there is equal number of low-level injuries and high-level injuries to train our model so that our model is not biased.

### 1) Data Replacement:

Here, we standardized the treatment of records containing unknown values, as specified in the CRSS manual. Instances where unknowns were represented by codes such as 999, 888, 99, 98, etc., were replaced as null values ( "?"). Standardizing the representation of missing values is a



preparatory step for further data imputation. we simplify the subsequent imputation process using the missing value node in Knime. There are a whole lot of replacements made as shown in the figure using **Math Formula** node.

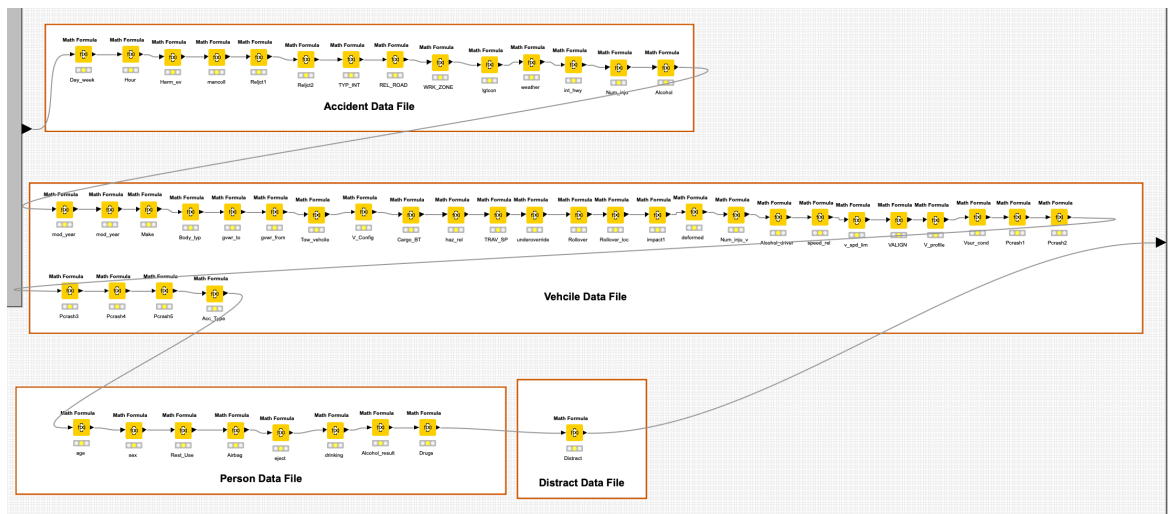


Figure 9: Image Showing series of math formula nodes used to bring out the null values

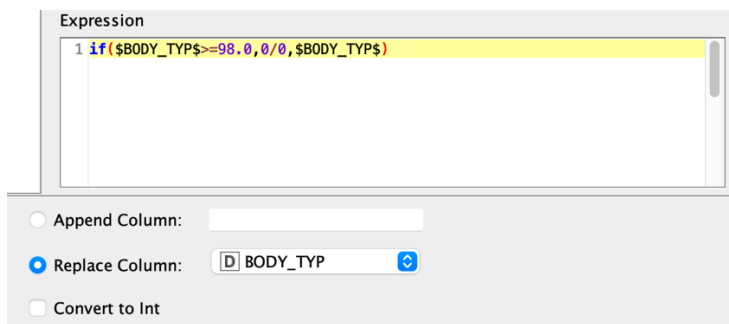


Figure 10: Image Showing logic used to bring the null values out

## 2) Datatype Conversion:

As noted previously, Knime's default treatment of nominal (categorical) values as numeric (double) prompted the need for corrective action. To rectify this, we employed the **Number to String Node** in Knime. This node facilitated the conversion of columns that were meant to represent nominal values or labels back to their appropriate data type. At this point by removing few columns and records we are left with 9 numeric and 51 nominal variables after datatype conversion.

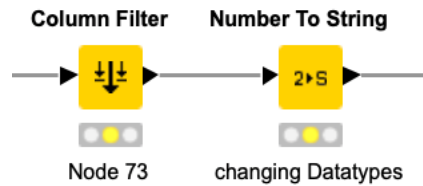


Figure 11: Image Showing data type conversion using number to string node

### 3) Imputations:

The utilization of the **Missing Value Node** in Knime facilitated the imputation process, simplifying the handling of missing data in the dataset. Imputation strategies were tailored to the nature of the variables:

**Nominal Variables:** For nominal variables, imputation was performed using the most frequent value. Additionally, a fixed value of "unknown" was employed for many other variables based on specific requirements.

**Numeric Variables:** Imputation for numeric variables varied based on the distribution of the data. For variables exhibiting a skewed distribution, the median was used as an imputation strategy. On the other hand, for variables with a distribution close to normal, imputation was done using the rounded mean. These strategies aimed to maintain the central tendency of the data while accounting for its distributional characteristics.

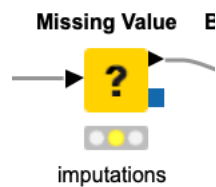


Figure 12: Image Showing using missing value node to do imputations

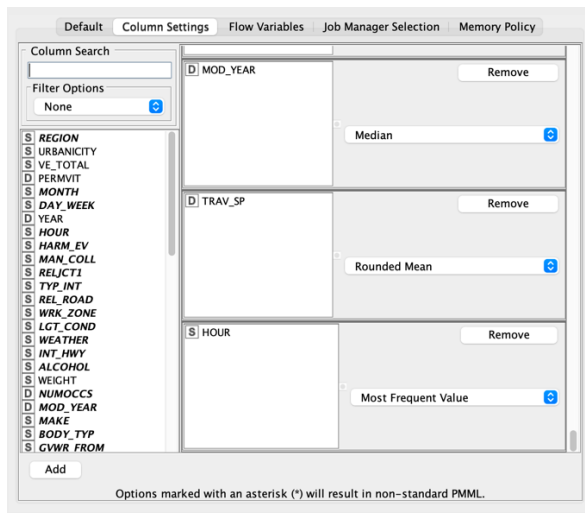


Figure 13: Image Showing different types of imputations used for different variables based on their datatype and data distribution

#### 4)Binning/Grouping:

Binning or grouping data is a crucial step in analysis, as it can simplify models and potentially improve results. However, it's important to note that binning can also lead to a loss of granularity in the data, potentially impacting the performance of models. The key lies in selecting the appropriate variables for binning based on the specific characteristics of the dataset. Rule Engine Node in Knime was utilized for binning variables. We Binned lot variables but later used only few based on the performances.

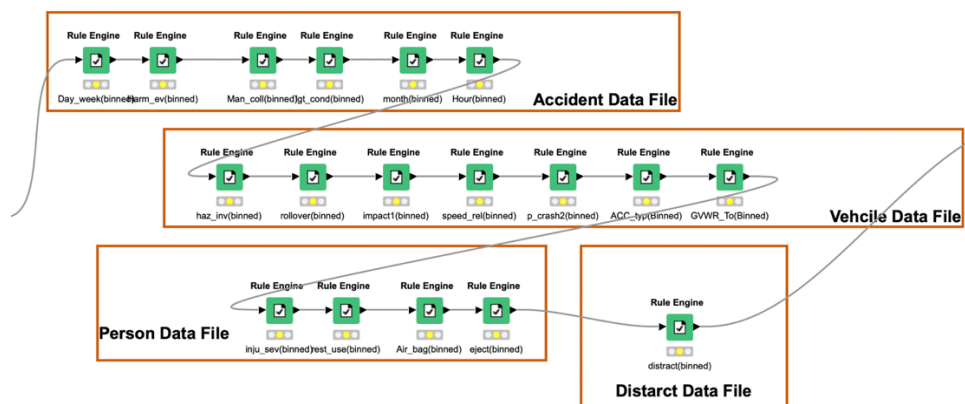


Figure 14: Image Showing series of Rule Engine nodes used to bin the variables

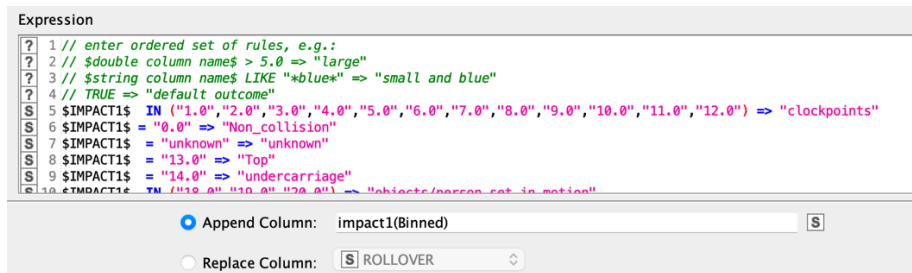


Figure 15: Image Showing the logic used to bin the variables

Variable	Bins	Number of Bins
Day_week (Binned)	Weekday, weekend	2
Harm_ev(Binned)	Non-Collison Harmful event, collision with MVIT, collision with object not fixed, collision with object fixed	4
Man_coll(Binned)	Front, angle, other, not with MVIT, sideswipe, rear, unknown	7
Lgt_cond(Binned)	Dark, Daylight, Dawn, Dusk, other, unknown	6
month(Binned)	Q1,Q2,Q3,Q4	4
hour(Binned)	Night, Morning ,Afternoon, Evening	4
Haz_inv(Binned)	No, Yes, unknown	3
Rollover(Binned)	No, Yes, unknown	3
Impact1(Binned)	Clock points, Non-Collison, Top, undercarriage, objects set in motion, left side, right side, unknown	8
Speed_rel(Binned)	No,Yes,unknown	3
P_crash2(Binned)	V_loc, VT, OMVIL, OMVEL, PPNM, O/A, Unknown	7
Acc_typ(Binned)	No Impact,1A,1B,1C,2D,2E,2F,3G,3I,4J,4K,5L,6M,other crash type ,unknown	15
GVWR_To(Binned)	Less-weight, medium-weight, heavy-weight, unknown	4
Inju_sev(Binned)	Low-level injuries, high-level injuries	2
Rest_use(Binned)	No, Yes, unknown	3
Air_bag(Binned)	Deployed, Not-Deployed, unknown	3

eject(Binned)	Not ejected, ejected, unknown	3
distract(Binned)	Not distracted, distracted, unknown	3
Body_typ(Binned)	1.0,2.0,3.0,4.0,5.0,6.0,7.0,8.0,not-required,unknown	10

Table2: Table showing the list of binned variables and respective bins

### **5) Handling Outliers:**

Certainly, addressing outliers is a critical aspect of preparing a dataset for modeling. Outliers can exert undue influence on statistical analyses and machine learning models, potentially leading to skewed results. Here in our case, we identified we few outliers/Anomalies in few columns and removed those data points as there are only few records. This process has also been done through stages wherever it is necessary.

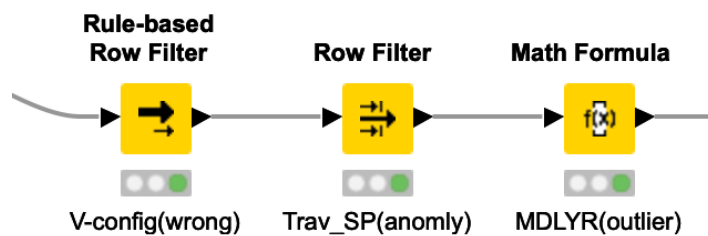


Figure 16: Image Showing series of row filter eliminating the outliers

Here is list of Outlier columns:

Variable	Outliers observed	Handling outlier
V_CONFIG	V_CONFIG =20 or V_CONFIG =21 These rows do not seem to be matching with the Body_typ as described in the manual	Removed outlier rows
TRAV_SP	One value (152.0)	Removed outlier row
Not injured	Values Less than 0	Removed outlier rows

MOD_YEAR	2022	Changed them to 2021 though they are considered to be 2022 as this put Age vehicle negative values which is not true
----------	------	--

Table 3: Table showing list of outliers observed and how they are handled

### C) Constructing Data:

Introducing new features through feature engineering is a strategic approach to unlock additional insights and potentially enhance model performance. The aim of this process is to provide the models with a richer understanding of the dataset, potentially leading to improved predictive capabilities. But only a select few derived features had a positive impact is insightful.

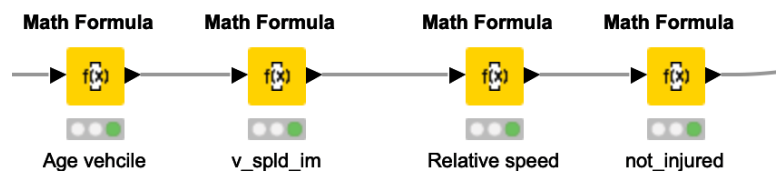


Figure 17: Image Showing series of math formula nodes used to derive new variables

Variable	Datatype	Derived From	Description
Age_Vehcile	Numeric	YEAR, MOD_YEAR	Tells the Age of the Vehicle
Relative Speed	Numeric	TRAV_SP,VSPD_LIM	Speed of the vehicle relative to speed limit
Not Injured	Numeric	NUMOCCS,NUM_INJV	Number of people in the vehicle that are not injured

Table 4: Table showing list of variables derived and their description

## D) Formatting Data:

We've performed data encoding using **one to many** node, which involves converting necessary columns from nominal to numeric format to make them compatible with number-based models like Artificial Neural Networks (ANN), Logistic Regression, and among others. Also, we used **Normalizer** node (min-max normalization) to massage and smoothen the data

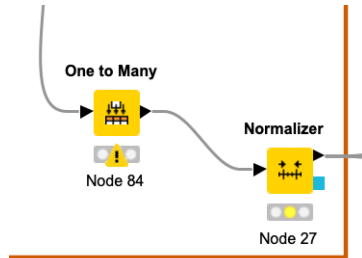


Figure 18: Image showing using one to many node to for data encoding

As we wrap up the Data Preprocessing phase, our dataset now stands at 23,608 rows and 29 columns, having undergone various transformations and optimizations (Table 6). With this refined dataset, the next step involves training our model. However, before diving into the modeling phase, we acknowledge and address the issue of data imbalance.

Variables	Data type	Statistics
URBANICITY	Binary	Mode: 1.0
MONTH	Nominal	Mode: 10.0
HOURL	Nominal	Mode: 17.0
RELJCT1	Nominal	Mode: 0.0
MAKE	Nominal	Mode: 20.0
J_KNIFE	Nominal	Mode: 0.0
V_CONFIG	Nominal	Mode: 0.0
CARGO_BT	Nominal	Mode: 0.0
TRAV_SP	Numeric	Mean:28.179 ,SD:16.93
UNDEROVERRIDE	Nominal	Mode: 0.0
IMPACT1	Nominal	Mode: 12.0

DEFORMED	Nominal	Mode: 6.0
VPROFILE	Nominal	Mode: 1.0
P_CRASH1	Nominal	Mode: 1.0
AGE	Numeric	Mean:40.947,SD:17.488
SEX	Nominal	Mode: 1.0
DRINKING	Nominal	Mode: 0.0
DRUGS	Nominal	Mode: unknown
BODY_TYP_binned	Nominal	Mode: 1.0
P_CRASH2(Binned)	Nominal	Mode: VT
ACC_TYP(Binned)	Nominal	Mode: 2D
GVWR_TO(Binned)	Nominal	Mode: Less weight
Inju_sev(Binned) - Dependent	Binary	Mode: Low level injuries
Rest_use(Binned)	Binary	Mode: yes
Air_Bag(Binned)	Binary	Mode: Deployed
Eject(Binned)	Binary	Mode: Not Ejected
Distract(Binned)	Nominal	Mode: unknown
Age_vehicle	Numeric	Mean:8.214 , SD:6.961
Relative speed	Numeric	Mean:-16.327 ,SD:17.877

Table 5: Table showing list of variables used for model training and their respective statistics

## Phase 4: Modeling

Modeling is important aspect as choosing appropriate model to the dataset is often hard and yields in better results. In the stage of constructing our models, we employed a range of modeling methods to address the core objective of forecasting the seriousness of injuries in car accidents. This came after we had readied the dataset. We transformed our dependent variable (injury severity) as a binary one, having only two potential results: low or high injury severity. This channelized us into a binomial classification problem.



## A) Selecting Model techniques:

There is no single machine learning method that universally suits this specific data mining task. Consequently, we experimented with different models, exploring both set-based and number-based approaches.

Here is the list of the models used:

**Set-based models:** Decision tree (DT), Random Forest(RF), Gradient Boosted Trees(GBT), Naïve Bayes(NB)

**Number-based models:** Artificial neural networks(ANN), Logistic Regression(LR)

## B) Generate Test Design:

We incorporated K-fold cross validation(K=10) methodology so in that way we use all our for both testing and training. In this method dataset is divided to 10 subsets and trained for 10 times each time using 9 folds for training and remaining fold for testing and finally all the testing results were averaged. Even though this tedious process we are much focused on Accuracy and balanced model We did this using **X-partitioner** and **X-Aggregator** node in knime.

## C) Building Model:

The workflow is nearly identical for both set-based models and number-based models, except for incorporating the "one to many" node for data encoding in number-based models. This encoding step was previously covered during the data formatting stage (3D) of data preparation.

### 1)Decision Tree (DT):

The Decision Tree is a set-based non-parametric model utilized for classification and regression tasks, is designed to improve our understanding of the data. Structured like a tree, it consists of internal nodes, branches, and leaves. Leaves represent predictions or outcomes, while internal nodes use various metrics to create additional branching nodes. In our implementation of the decision tree, as illustrated in Figure (5), we introduced an x-partitioner node at the start of a cross-validation loop. Additionally, we incorporated an **equal-size sampling** node(Removes rows from the input data set such that the values in a categorical column are equally distributed) to balance the input dataset by removing rows, ensuring even distribution in categorical columns. Subsequently, we trained the decision tree learner and predictor, necessitating the division of data

into training and testing sets. Finally, we used the x-aggregator node to gather iterations and results from the predictor node, enabling the comparison of predicted and actual class outcomes. This workflow design remains consistent across all the models we employed.

Model able to generate better results for this Parameter settings :

Quality Measure – Gain ratio , Pruning method – MDL , Binary nominal splits – 10, number of threads -8 .

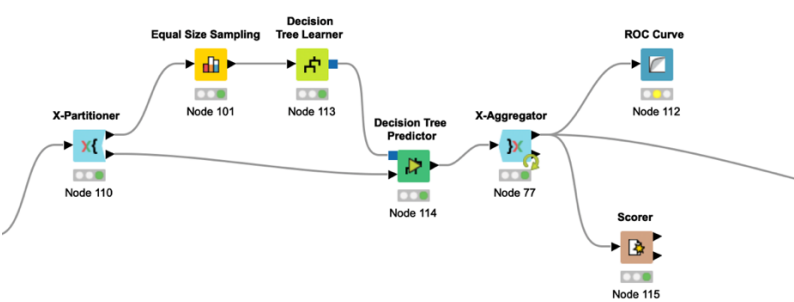


Figure 19: Image showing Decision Tree workflow

**Results with balancing:**

Inju_sev(Bi...	Low -level I...	High -level...
Low -level...	13823	5309
High -leve...	1472	3004

**Results without balancing:**

Inju_sev(Bi...	Low -level I...	High -level...
Low -level...	18455	677
High -leve...	3367	1109

The impact of equal-size sampling is evident in this scenario. Given that the predominant class is low-level injuries, and the minority class is high-level injuries, the results show significantly improved accuracy for the low-level class. Conversely, predictions for the high-level class suffer without the balancing effect introduced by equal-size sampling.

**2)Random Forest (RF)**

Random Forest is a set-based supervised ensemble model that combines multiple decision trees independent trees to predict the target variable. It makes decisions based on the majority vote of the trees. It's great for solving tricky classification problems by understanding how different factors work together to influence the outcome. The workflow remains same except for learner and predictor.

Model able to generate better results for this Parameter settings :

Split criterion – Gini Index , Number of models – 200 (trying to be cost effective)

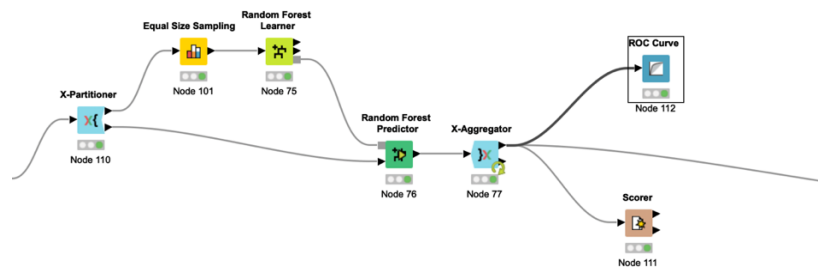


Figure 20: Image showing using Random Forest workflow

### **3)Gradient Boosted Trees**

A Gradient Boosting Tree is a set-based and specialized form of gradient boosting, emphasizing the utilization of decision trees as weaker learners in a group. This robust machine learning method constructs a predictive model by combining numerous decision trees, each aimed at rectifying the mistakes of the ones that came before it. The workflow remains same except for learner and predictor.

Model able to generate better results for this Parameter settings :

Tree depth – 3 , Number of models – 200 , use binary split for nominal columns, learning rate -0.1

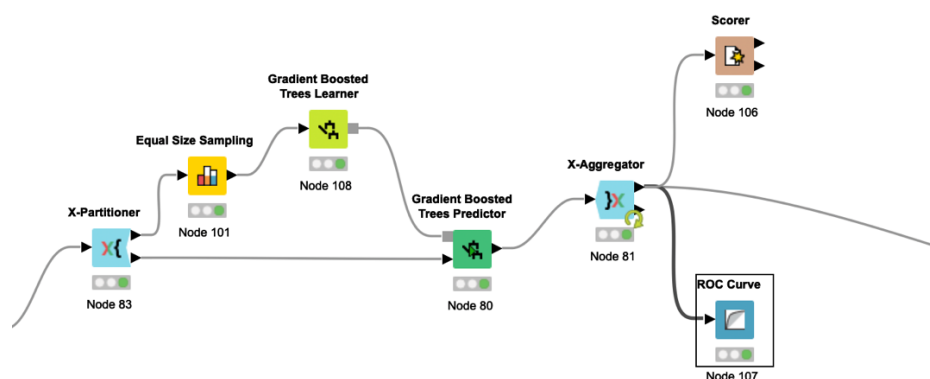


Figure 21: Image showing using Gradient Boosted Tree workflow

### **4)Naïve Bayes (NB)**

Naive Bayes, a set-based algorithm in supervised machine learning, is widely applied for classification tasks using Bayes' theorem. A notable benefit of this classifier is its efficiency in

requiring a small training dataset to estimate parameters, making it well-suited for situations with limited data. The general process is similar, differing only in the learning and prediction steps. Model able to generate better results for standard and default Parameter settings

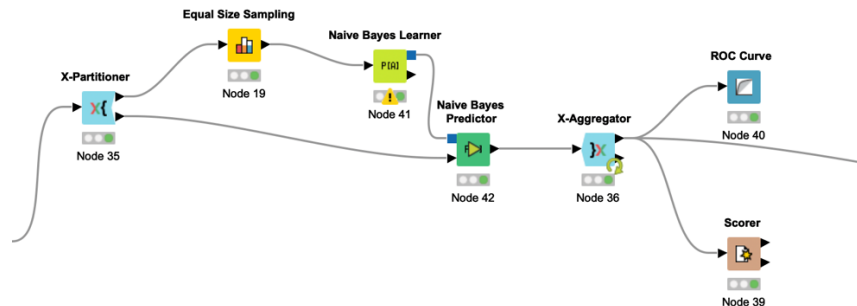


Figure 22: Image showing using Naïve Bayes workflow

## **5)Artificial Neural Network (ANN-MLP)**

ANN- MLP (Multilayer Perceptron) is a number-based designed to mimic the biological neural networks in the human brain. Neurons in ANN calculate their outputs based on weights from various inputs or the previous layer, employing an activation function. During training, the model adjusts the weights between connections to improve performance. The workflow remains same except for learner and predictor.

Model able to generate better results for this Parameter settings :

Number of hidden layers – 3 , Number of hidden neurons per layer – 10

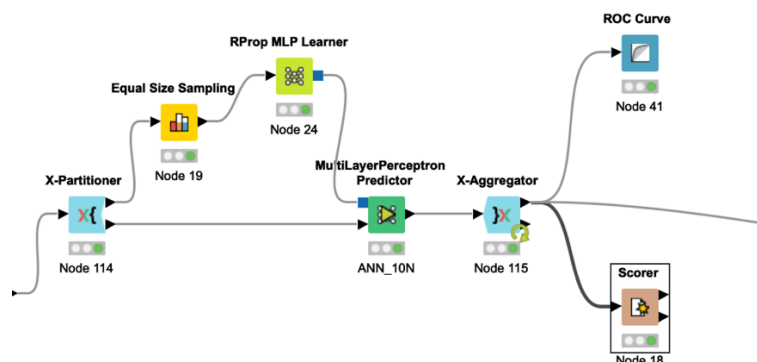


Figure 23: Image showing using ANN-MLP workflow

## 6) Logistic Regression (LR)

Logistic Regression is a commonly used number-based model for classification tasks. It employs logistic or sigmoid functions to map values and make predictions. The workflow is same except for learner and predictor.

Model able to generate better results for this Parameter settings :

Solver – stochastic average gradient

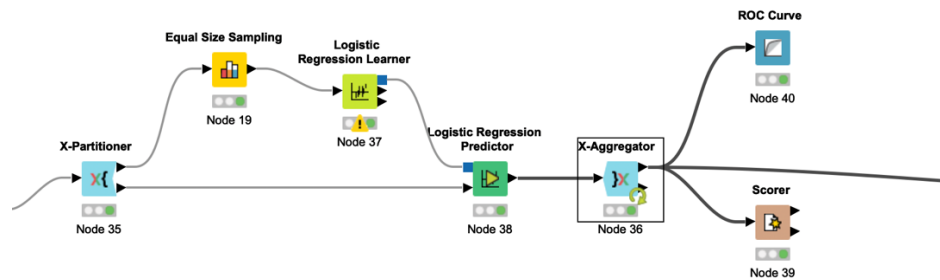


Figure 24: Image showing using Logistic Regression workflow

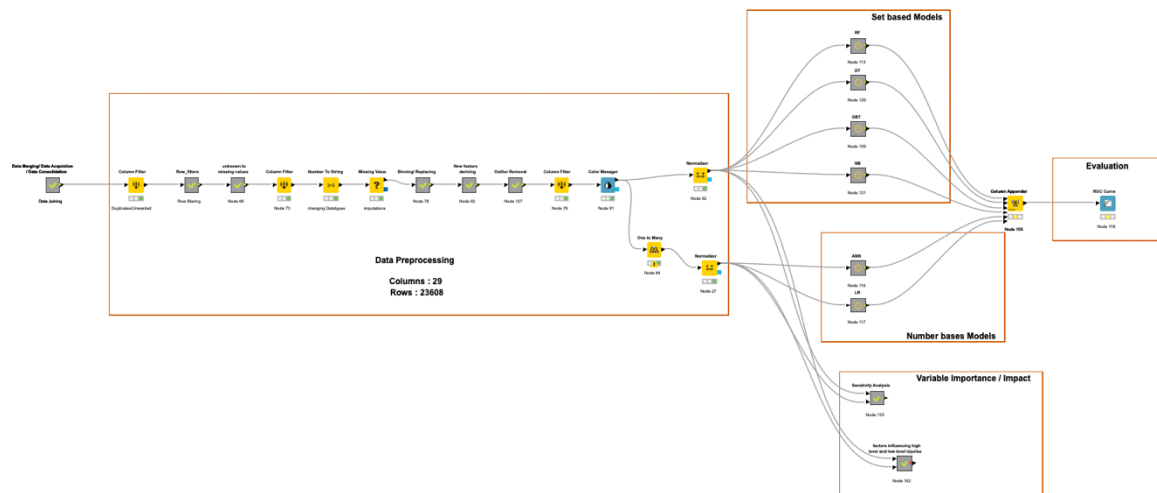


Figure 41: Image showing complete workflow of the entire Project (Canvas)

## Phase 5: Evaluation

Evaluation is very important it is a measure of verifying the model whether the predictions are close to the actual values. To maintain pseudo randomness during training the model we used fixed random seed =7666 for all the models. Tabulated evaluations of each model have been listed below based on 10-fold cross validation.

### A) Evaluation Results:

Model Type	AUC	ACCURACY	SENSITIVITY	SPECIFICITY
Random Forest(RF)	0.793	81.01%	0.88	0.56
Decision Tree(DT)	0.742	71.3%	0.723	0.671
Gradient Boosted Trees(GBT)	0.799	73.1%	0.738	0.703
Naïve Bayes (NB)	0.773	73%	0.747	0.66
Artificial Neural Networks(ANN-MLP)	0.730	68.7%	0.692	0.669
Logistic Regression (LR)	0.796	73.28%	0.738	0.707

Table 6: Table showing using Evaluation metrics for different models

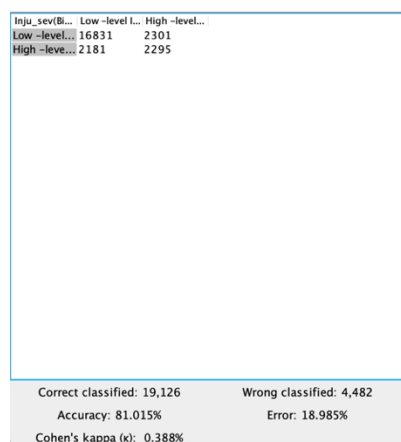


Figure 26: RF-Confusion Matrix

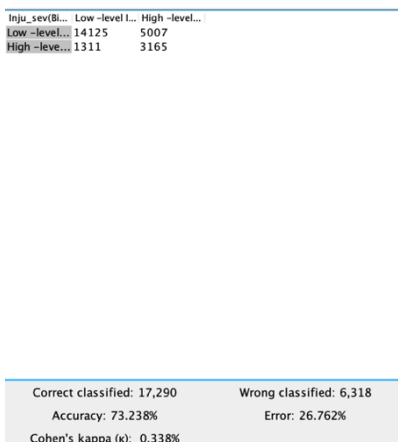


Figure 27: LR-Confusion Matrix

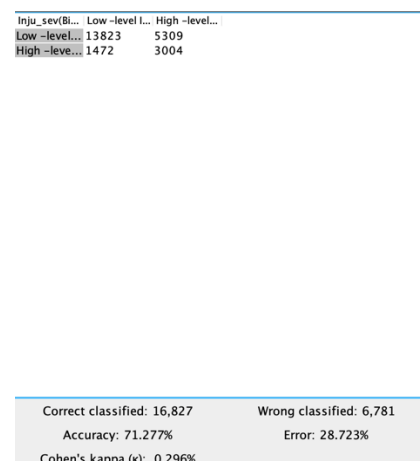


Figure 28: DT-Confusion Matrix

Inju_sev(Bi...	Low-level I...	High-level...
Low-level...	14111	5021
High-level...	1331	3145

Correct classified: 17,256	Wrong classified: 6,352
Accuracy: 73.094%	Error: 26.906%
Cohen's kappa (k): 0.335%	

Figure 29: GBT-Confusion Matrix

Inju_sev(Bi...	Low-level I...	High-level...
Low-level...	14285	4847
High-level...	1524	2952

Correct classified: 17,237	Wrong classified: 6,371
Accuracy: 73.013%	Error: 26.987%
Cohen's kappa (k): 0.316%	

Figure30: NB-Confusion Matrix

Inju_sev(Bi...	Low-level I...	High-level...
Low-level...	13231	5901
High-level...	1480	2996

Correct classified: 16,227	Wrong classified: 7,381
Accuracy: 68.735%	Error: 31.265%
Cohen's kappa (k): 0.262%	

Figure31: ANN-Confusion Matrix

Table 6 shows predictive accuracies for all the models implemented From the results we can infer that the Random Forest model showcased exceptional performance with an impressive 81.01% accuracy and AUC of 0.793 and following to it Logistic Regression has next better results with accuracy of 73.2% and AUC of 0.796. On the lower spectrum, the ANN-MLP achieved a modest 68.7% accuracy with an AUC of 0.730. From the confusion matrices (from Figure 26-31)notably, all the models face challenges in producing accurate predictions for high-level injuries, attributed to the scarcity of data and a limited number of variables explaining these high-level injuries.

## B) Combined ROC Curve:

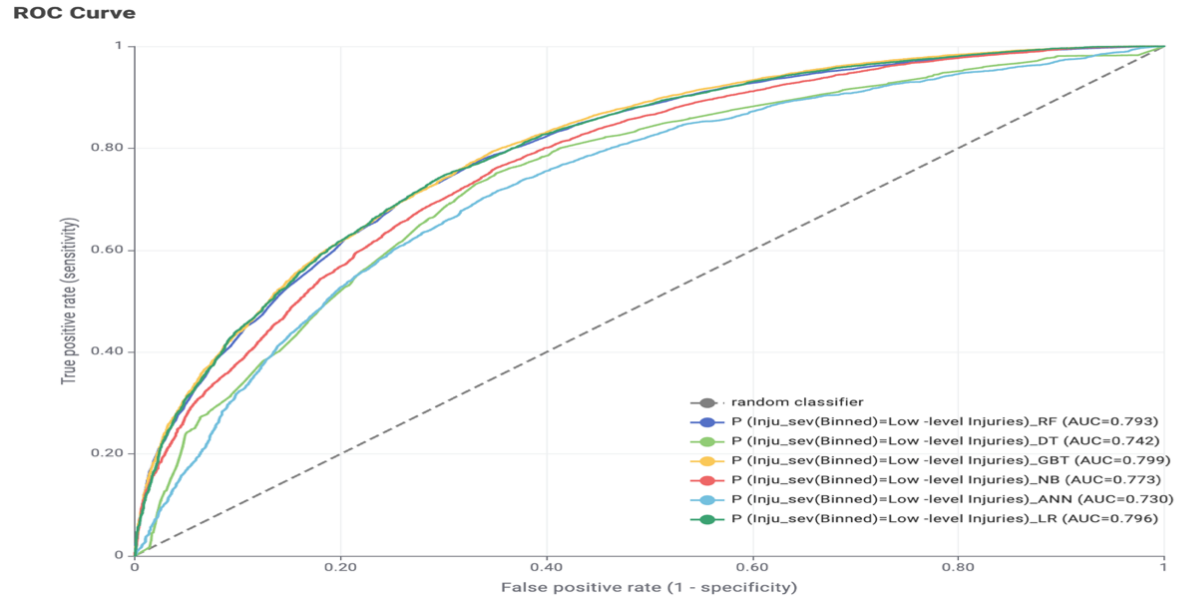


Figure 32: Image Comparing ROC curves for different models

From the Combined ROC chart, we can infer that all the models have close AUC values and among all Gradient Boosted Tree and Logistic regression have little better results(power of prediction).

## C) Variable Importance using Sensitivity Analysis:

While the accuracy measures from all four model types are significant to validate our methodology, the primary aim of this study was to pinpoint and prioritize the key risk factors that influence the severity of injuries sustained by drivers in automobile crashes. One way of doing this is using sensitivity analysis for finding variable importance.

This method involves systematically eliminating input variables one at a time from the model and assessing the impact of their absence on the predictive performance of the machine learning model. For each input variable, the model is trained and tested to gauge its individual contribution and importance. This analytical approach is commonly applied to the best four models as this process is really time consuming. So, we ensembled the results of RF,LR,GBT,DT.



To consolidate the sensitivity analysis results across all four model types, we employed ( Equation 1) to ensure a proper ensemble. This process aimed to weigh the contributions of each model type based on their cross-validation accuracy. The resulting fused variable importance values were normalized, and subsequently presented graphically (Figure 33).

Total Accuracy = RF Accuracy + DT Accuracy + NB Accuracy + LR Accuracy

RF\_ VI = Vector consisting of RF Variable importance values for included variables

DT\_ VI = Vector consisting of DT Variable importance values for included variables

NB\_ VI = Vector consisting of NB Variable importance values for included variables

LR\_ VI = Vector consisting of LR Variable importance values for included variables

RF\_ W = weight of RF = (RF Accuracy / Total Accuracy)

DT\_ W = weight of DT = (DT Accuracy / Total Accuracy)

LR\_ W = weight of LR = (LR Accuracy / Total Accuracy)

NB\_ W = weight of NB = (NB Accuracy / Total Accuracy)

Ensembled Variable Importance = (RF\_ VI \* RF\_ W) + (DT\_ VI \* DT\_ W) + (NB\_ VI \* NB\_ W) + (LR\_ VI \* LR\_ W) ----- equation 1

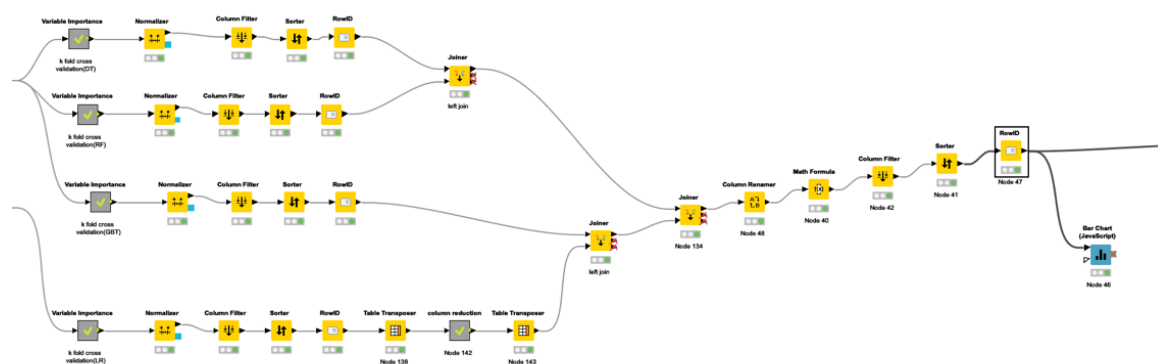


Figure 33: Image showing workflow for sensitivity Analysis

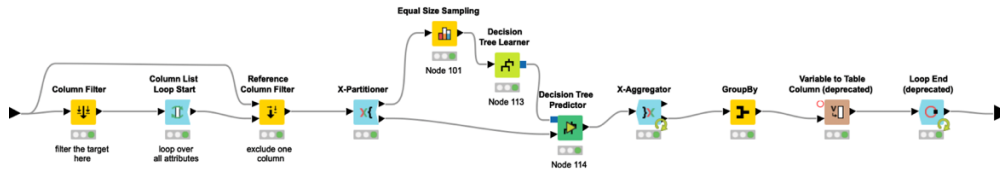


Figure 34: Image showing workflow inside Variable Importance meta node of sensitivity Analysis

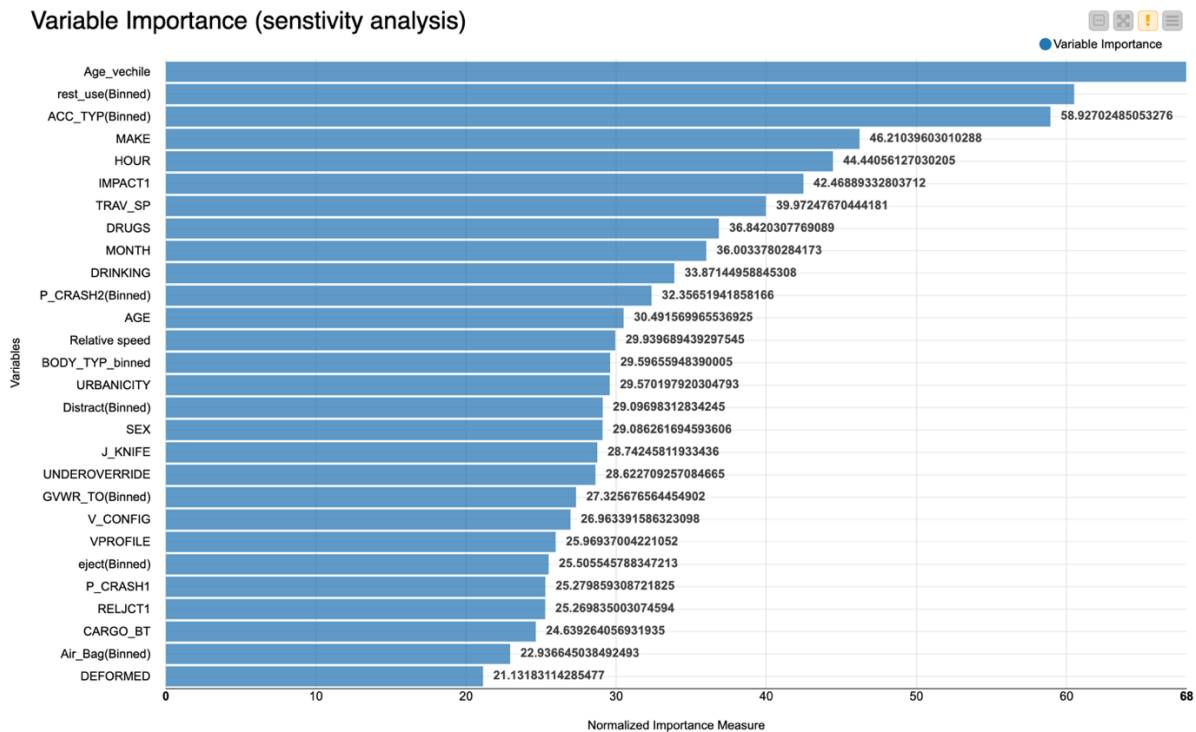


Figure 35: Image showing Variable Importance using sensitivity Analysis

Figure 35 displays the variables of significance concerning a crash resulted in some level of injury. A thorough examination of the sensitivity analysis results unveils four distinct risk groups, Both groups and attributes inside the group are arranged from most to least importance

### Group1(Age\_vehicle,Rest\_used(Binned),ACC\_TYPE,MAKE) – TOP

- Notably, Age\_vehicle takes the lead, emphasizing that the age of the vehicle significantly influences the injuries. This aligns with the intuitive understanding that older cars may have weaker structural integrity and fewer safety features.

- Following closely in importance in this group is Rest\_used (Binned), highlighting the crucial role of seat belt or restraint usage in determining injuries.
- ACC\_TYPE follows, shedding light on the nature of the accident, whether it occurred in the same trafficway, same direction, involved changing trafficway, or featured a turning vehicle.
- Finally, the Make of the car, indicating the manufacturer, stands out as a crucial factor. This result is unsurprising, considering that certain manufacturers prioritize safety systems, albeit potentially at a higher cost, while others may not provide the same level of safety features.

#### **Group 2 (HOUR, Impact1, Trav\_sp, Drugs)**

- HOUR of the Crash: Influences injury severity, with a notable observation that accidents are more prevalent during late hours, particularly around 5 pm.
- Impact1: Describes the area on the vehicle that produced the initial instance of injury, indicating clock points (Top, Left, Right).
- Trav\_sp: Provides information about the speed of the vehicle during the crash.
- Drugs: Indicates whether the driver was under the influence of any drugs during the accident.

#### **Group 3 (MONTH, Drinking, P\_CRASH2 (Binned), Age)**

- MONTH: Specifies the month in which the crash occurred, with a significant concentration observed in October.
- Drinking: Indicates whether the driver consumed alcohol during the crash.
- P\_CRASH2 (Binned): Describes the critical event leading to the crash, encompassing factors like pedestrians, animals, objects, or loss of vehicle control.
- Age: Represents the age of the driver.

#### **Group 4 (RELJCT1, CARGO\_BT, Airbag (Binned), Deformed) – last few variables**

- RELJCT1: Determines whether the accident is related to a junction.
- CARGO\_BT: Describes the primary cargo-carrying capability of the vehicle.
- Airbag (Binned): Indicates whether airbags deployed during the crash.
- Deformed: Specifies the amount of damage sustained by the vehicle, whether it's minor or involves functional damage.

Notably, Group 4 variables have been found to be relatively marginal in significance. With its comprehensive representation of other variables, did not find them to be highly significant.

Since our aim is to identify risk factors that impact severity of driver injuries, we further classified the risk factors for high-level and low-level .To make this happen we've also developed a specialized workflow (Figure )named as **Factors influencing Low-level and High-level Injuries**. This process categorizes features based on their impact on low-level and high-level injuries in sensitivity analysis. We induced row filter in sensitivity analysis to filter low-level injuries separately and high – level injuries separately and train the model. This gives variable importance on high-level injuries separately and low-level injuries separately. And finally, we used a scatter plot to show compare the results shown figure(39).

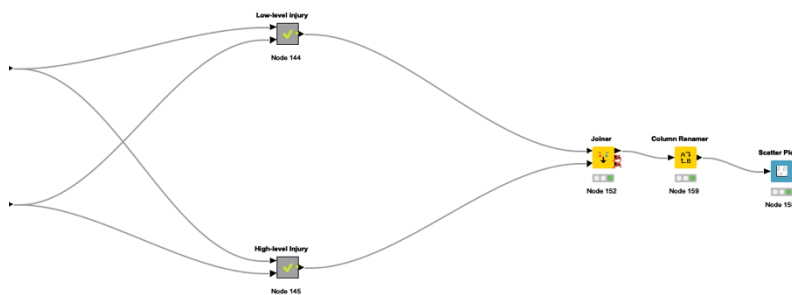


Figure 36: Image showing workflow of Factors influencing Low-level and High-level Injuries.

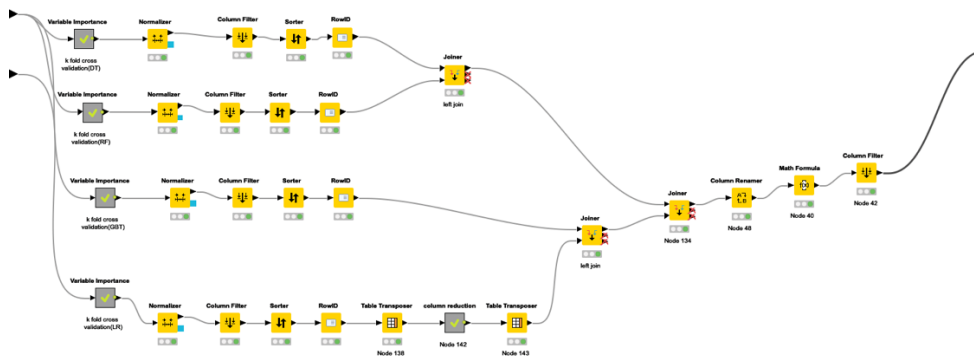


Figure 37: Image showing workflow inside Low-level injury meta node

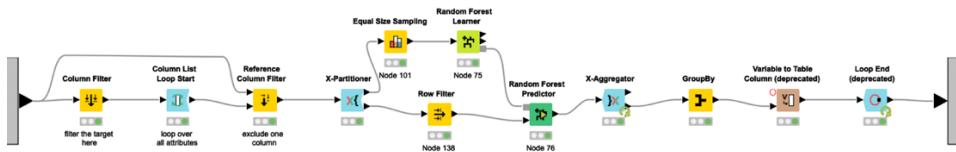


Figure 38: Image showing workflow inside Variable Importance meta node

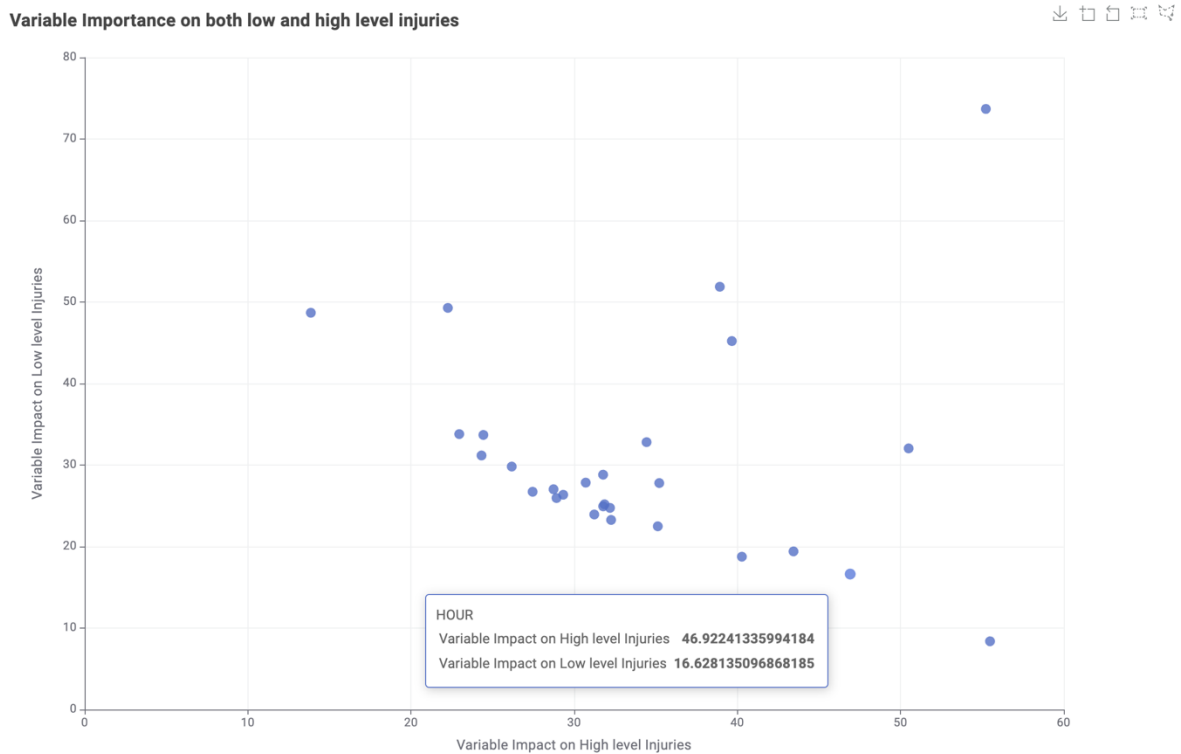


Figure 39: Image showing Variable Importance using sensitivity Analysis

From the results of the analysis, we found the following findings

Factors (TOP) Influencing Minor-Level Injuries:

- Deformed
- Age
- Airbag (Binned)
- Eject (Binned)

Factors(TOP) Influencing Major-Level Injuries:

- MONTH
- IMPACT1

- HOUR
- Rest\_used
- Make

It is evident from Figure 39 that only a few variables reveal the pattern behind low-level injuries, including Deformed, Age, Airbag (Binned), and Eject (Binned). Conversely, for high-level injuries, there are fewer variables explaining the pattern, which is expected due to the limited amount of data for high-level injuries. These variables include MONTH, IMPACT1, HOUR, Rest\_used, and Make.

Notably, there is one variable, Age\_vehicle, that effectively explains extremely better both high-level and low-level injuries. This variable falls in the top right corner of the graph, along an imaginary diagonal line passing through the center point of both axes. In the quest for a robust model, having more variables along this diagonal line is desirable, as they can effectively explain both low-level and high-level injuries. Unfortunately, only two other variables, TRAV\_SP and ACC\_TYP, fall on the higher part of this line.

## D) Decision Tree Branches:

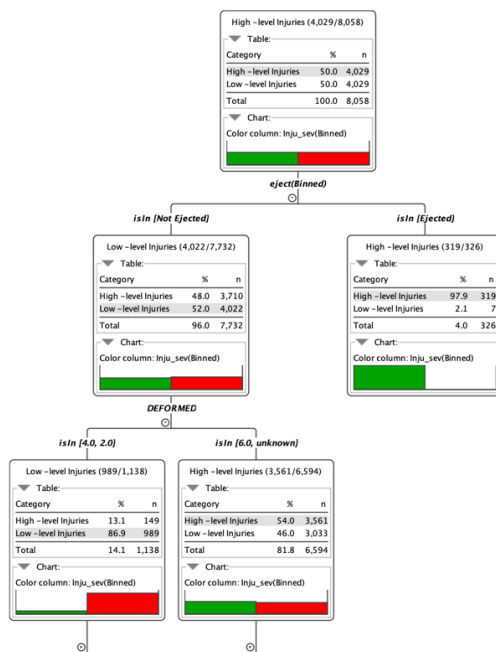


Figure 40: Image Showing Decision Tree Branches

The Decision Tree depicted in Figure 40 reveals diverse decision pathways that trace individuals and their corresponding association with the binary decision of "Yes or No" regarding the occurrence of severe injuries in a crash. In constructing this tree, we implemented the MDL pruning method, resulting in the pruning of certain branches as we delve into deeper levels.

Examining the tree, if the input variable "eject (binned)" indicates ejection, there is a substantial 97.9% likelihood of high-level injuries, in contrast to a 48% chance if the input suggests no ejection. This highlights the significant influence of the "eject (binned)" variable.

Furthermore, the positioning of variables used for splits is crucial; those closer to the root have a more pronounced impact. This underscores the importance of variable influence in the decision-making process of the model.

## **Phase 6: Deployment**

As we wrap up the final phase of the CRISP-DM data mining process, our efforts have been focused on systematically organizing and presenting the knowledge and insights gained throughout the project. This presentation is meticulously designed to ensure that our findings are not only accessible but also comprehensible to key stakeholders and end-users, providing them with the necessary information to make informed decisions.

The crux of our exploration has revolved around the pressing issue of automobile crashes and the associated severity of injuries. While undoubtedly a significant challenge, we firmly believe that our key stakeholders, comprising both government and corporate entities, possess the potential to be catalysts for positive change. Through the utilization of the innovations and regulatory insights derived from our findings, these stakeholders can play a pivotal role in shaping safer roads for the future.

Among the various models employed, the Random Forest Model has emerged as our standout performer. This powerful tool not only aids in predicting outcomes but also serves as a proactive measure for preventing undesirable incidents. We envision that key decision-makers in both

government and corporate sectors can leverage this model effectively, utilizing it as a valuable resource to enhance road safety strategies and mitigate potential risks.

## **Conclusion**

Upon receiving real-life data, it became evident that dedicating sufficient time to comprehend and preprocess data before constructing predictive models was crucial. The quality of our interaction with the data significantly influenced overall accuracy, underscoring its importance in the project.

In the development of six distinct models, encompassing both number and set-based approaches, addressing data imbalance through equal-size sampling, and employing 10-fold cross-validation proved pivotal. Among the varied outcomes, the random forest model emerged as the most effective, boasting an 81.01% accuracy, 88% sensitivity, 56% specificity, and a noteworthy 79.3% AUC.

Transitioning from model specifications, identifying impactful variables through sensitivity analysis was imperative. In the realm of road safety and mitigating human injury, the study pinpointed crucial factors for both major and minor injuries. Variables such as month, impact points, hour of the crash, seat belt usage, vehicle age, and manufacturer played significant roles in major injuries, while factors like deformation type, driver's age, airbag deployment, and ejection status were vital for minor injuries.

Drawing insights from our analysis, we offer recommendations for drivers, car manufacturers, and authorities. Drivers are advised to use safety restraints properly, choose cars with functional airbags, abstain from substances impairing alertness, be mindful of vehicle age, older drivers should be more careful. Car manufacturers should invest in robust designs, incorporate safety features, and develop intelligent driving systems. Authorities are urged to enforce strict measures against drunk and dangerous driving.

Ultimately, to leverage our findings effectively, presenting this information to key stakeholders is crucial. We anticipate that our results can inform decision-making in various domains, leading to refined regulations and innovations for a safer future on the roads. The overarching goal is to



enhance safety and reduce injuries, fostering a collective commitment to creating a safer road environment.

## Citations

1. *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>
2. Delen, D., Tomak, L., Topuz, K., & Eryarsoy, E. (2017). Investigating injury severity risk factors in automobile crashes with predictive analytics and sensitivity analysis methods. *Journal of Transport & Health*, 4, 118–131. <https://doi.org/10.1016/j.jth.2017.01.0>