MSIS – 5633 PREDICTIVE ANALYTICS TECHNOLOGIES

KNIME TERM PROJECT

Examining the factors influencing the severity of injuries in Automobile accidents using predictive analytics.

Due Date

November 26, 2023

By

Rachael Schaefer (A20381063)

Rishitha Ganagoni (A20398497)

Team Page

Rishitha Ganagoni

Rachael Schaefer

Table of Contents

Executive Summary

In this analysis, our team was assigned the task of examining factors influencing injury severity in automobile crashes. We utilized various predictive model techniques and sensitivity analysis methods to gain insights that could contribute to a better understanding of these incidents, aiming to mitigate future hazards and improve the well-being of those involved. Our study revealed numerous impactful factors contributing to the severity of injuries sustained in car accidents. Specifically, restraint (seatbelt) use and ejection were identified as important variables related to injury severity

For this analytics project, we employed a comprehensive crash dataset obtained from the U.S. Department of Transportation, specifically the National Highway Traffic Safety Administration. This dataset represented approximately one percent of all reported domestic automobile crashes in the United States, as documented by law enforcement agencies. To ensure consistent sampling, Bootstrap sampling was utilized for one of the models. We conducted the analysis using five widely used predictive analytics algorithms (decision tree, random forest, multi-layered perceptron, logistic regression, and naive Bayes classifier) within the KNIME application. These models were instrumental in capturing the intricate relationships between injury severity levels and various crash-related risk factors.

Additionally, we applied sensitivity analysis to the trained prediction models, identifying the hierarchy of importance among crash-related factors concerning different injury severity levels. By employing diverse prediction methods instead of relying on a singular approach, we generated a coherent list of risk factors appropriately ranked based on their contributions to the predicted outcomes.

The results yielded invaluable insights through our predictive analytics approach, revealing undeveloped value within the dataset and highlighting the relative importance of crash-related risk factors across different levels of injury severity. As technology advances and improved safety measures are implemented, we anticipate that our findings can contribute to preventing crashes or minimizing the impact of injuries sustained in such incidents. Given the increasing number of vehicles on the road and growing driver distractions, we aim to make a positive impact by sharing our findings and potentially mitigating the trend of more severe injuries associated with accidents.

Business Understanding:

As automobiles have progressed, they have gained speed and improved safety through innovations like airbags and seat belts. Recent advancements include driver-assist technology in newer vehicles which prevents lane drift and alerts drivers to blind spots and potential forward collisions. Despite these technological improvements, there has been a long-term increase in motor vehicle accident deaths. This trend was notably evident in the years 2020-2022, a period marked by reduced traffic due to the COVID-19 pandemic. However, crash severity also rose, attributed to risky behaviors such as drunk driving, improper lane changes, turning without signaling, and driving without seatbelts. Law enforcement reports indicate a common occurrence of speeding, supported by New York City traffic data revealing a 108% increase in traffic speed on Midtown Avenue. Even traditionally cautious areas like school zones have experienced a surge in speeding violations. Although it was initially hypothesized that less crowded roads during the pandemic led to increased speeding, the return to normal road congestion did not result in a decrease in speeding incidents. [i]

According to the National Highway Traffic Safety Administration, there was a 17% rise in vehicle fatalities from 2019 to 2021. This represents a significant and unprecedented surge, marking the highest increase since the 1940s. [ii] Numerous stakeholders are acknowledging this trend and endeavoring to change its trajectory. Entities such as insurance companies, local governments, and car manufacturers have implemented various interventions to enhance road safety. In an effort to reduce risk, insurance companies have explored telemetric devices that utilize in-car sensors to gather information about drivers' habits. State transportation administrations have adopted a humorous approach to capture drivers' attention, employing slogans like "drinking and driving go together like peas and guac" or "Hocus pocus drive with focus," drawing on popular culture to promote safety awareness. However, the effectiveness of these campaigns is still uncertain.[iii] As mentioned earlier, vehicles now boast a growing array of automated safety features; nevertheless, manufacturers are in a competitive race to introduce fully autonomous cars to the market. The revolution has begun gradually, with San Francisco permitting the operation of driverless taxis. However, incidents of car crashes involving these autonomous taxis have prompted regulators to limit their numbers on the road. It is inevitable that driverless technology will encounter failures at some point. The crucial question to be addressed by the data is whether autonomous vehicles are safer than those operated by humans. The information provided by the National Highway Traffic Safety Administration will be pivotal in evaluating the risk associated with autonomous vehicles.

The underlying factors contributing to traffic accidents and the severity of resulting injuries have become a focal point of investigation. The goal is not only to prevent potential casualties but also to diminish injury severity and associated costs. To further explore this topic, it is essential to examine historical data and pinpoint the key factors that have a significant impact on the severity of injuries. These factors include a range of elements, such as behavioral or demographic characteristics (e.g., drug/alcohol use, seatbelt usage, gender, age), situational crash characteristics (e.g., road type, direction of impact), environmental factors (e.g., road surface condition, time of day, visibility), and vehicle characteristics (e.g., body type, weight, age, maintenance level). Employing widely used predictive modeling techniques through KNIME, our objective is to pinpoint the person, vehicle, and accident-related risk factors that wield the most significant impact on the severity of injuries sustained in car crashes.

Data Understanding

The data used in our project, related to crashes, was sourced from the National Highway Traffic Safety Administration database. It was provided in four distinct text files, each associated with different aspects of the accident, including road conditions, environmental impact, etc. These files covered information on the accident, individuals involved (demographics, injuries, and situational context), the vehicles (specific features), and any distractions.

Moving on to the further step, after defining the problem and gaining a comprehensive understanding of our scope, it is crucial to establish a foundational understanding of the datasets. This involves dedicating significant time to massage, cleaning, and identifying key variables that form the basis of our project. Specifically, members of the data science community pay close attention to columns with missing values, as this provides additional context for the project.

Gathering information:

The National Highway Traffic Safety Administration (NATSA) aims to reduce the human and property toll of motor vehicle accidents, which cause thousands of deaths, injuries and billions in damages on an annual basis. The agency publishes the Crash Report Sampling

System (CRSS) to distribute data gathered from police-reported crashes. This report provides a comprehensive overview to identify safety issues, measure trends, and support safety initiatives. This report is a priority of the NHTSA as it enhances transparency and makes data accessible. The data is gathered from crashes reported to police departments. There is always the risk that a crash is not reported, however, most states require that accidents are reported to police if there is an injury that requires more than basic first aid.

There are more than 6 million vehicle crashes reported to the police each year. Of those reported to the police, the data provided was chosen from 60 sites across America and taken directly from the police report, crash diagrams, or the written summary. The data was provided in four statistical analysis software files. The *accident* data file includes information such as the weather, date, light condition, and manner of collusion. The *vehicle* data file includes the car model year, number of injured vehicle occupants, and initial contact point. The *person* data file includes police reported alcohol involvement, ejection, injury severity, seating position, age, and sex. The *distract* file contains information about the driver distractions. All data included years 2016 – 2021. Merging all the data sets with the help of left join into one single file.

File Name	**Number of Rows**	**Number of Columns**
Accident	54,200	46
Distract	95,845	11
Person	133,734	59
Vehicle	95,785	88

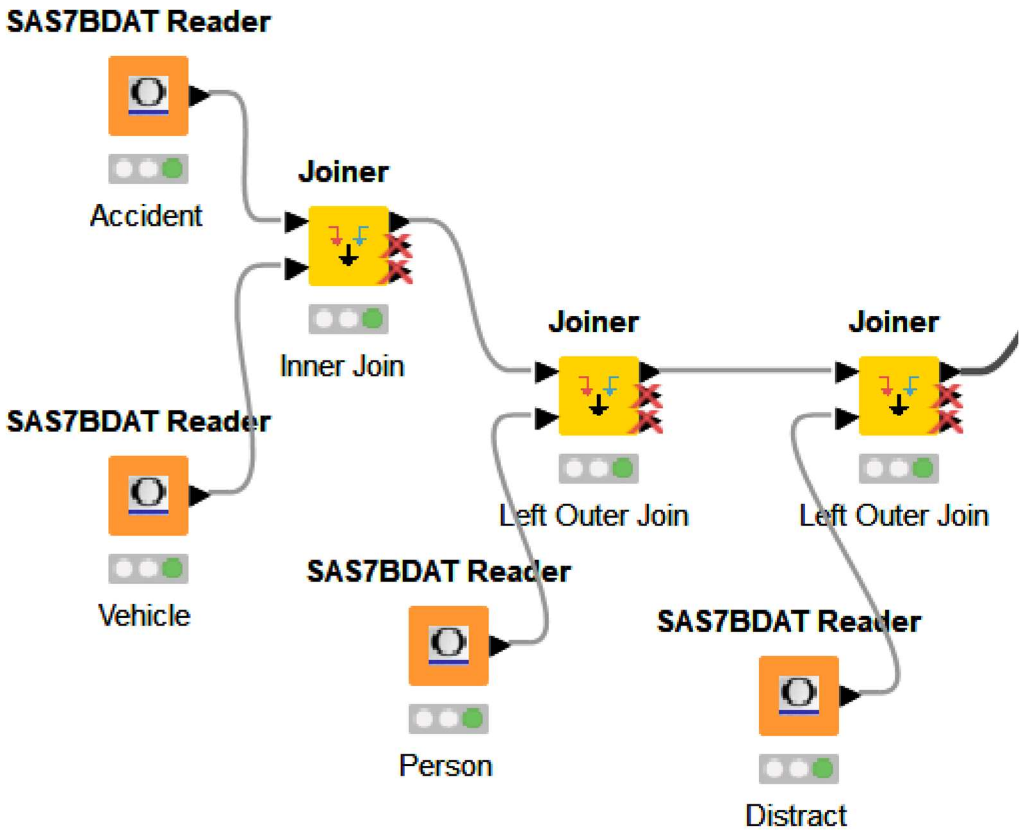


Figure 1: Showing the picture of merging all the data files with left join.

Describing Data:

Volume:

The ultimate dataset comprises 128,589 records and includes 204 attributes. This extensive number of attributes can impede the efficiency of designing and running certain models. Therefore, it is crucial to prioritize the cleaning of the data as a vital step in the analysis process.

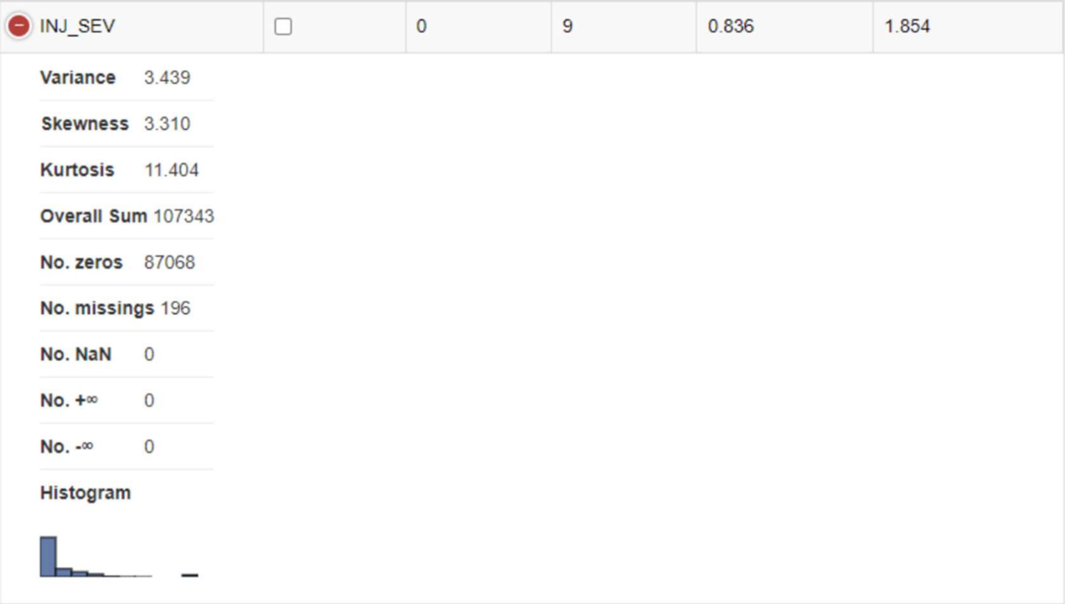
Categories:

The dataset comprises both numeric and nominal values, consisting of a total of 198 numeric attributes and six nominal attributes. The predominant portion of the dataset is characterized by numeric attributes.

Exploring the Data:

This integrates various statistical analysis tools, enabling users to perform statistical tests, correlation analysis, and other advanced statistical operations on the data.

Within the Knime tool, the Statistics and Data Explorer nodes play a crucial role in examining the consolidated dataset. The primary focus is on the INJ_SEV variable, which serves as a dependent attribute in the dataset. To gain deeper insights into this particular attribute, one can consult the results generated by the statistics node. The characteristic demonstrates positive skewness, indicating that most of the data is clustered on the left side of the distribution, with a few outlier values present on the right side.



⊖ INJ_SEV	□	0	9	0.836	1.854

Variance	3.439
Skewness	3.310
Kurtosis	11.404
Overall Sum	107343
No. zeros	87068
No. missings	196
No. NaN	0
No. +∞	0
No. -∞	0

Histogram

Figure 2: Dependent Variable named INJ_SEV

Variable List:

Variable	**Data Type**	**Mean**	**Standard Deviation**
YEAR	Numeric	2021	0
NUMOCCS	Numeric	1.762	1.242
HIT_RUN	Numeric	0.058	0.235
MOD_YEAR	Numeric	2300.782	1489.275
MAKE	Numeric	33.687	20.709
BODY_TYP	Numeric	18.368	20.400
DR_PRES	Numeric	0.998	0.042
SPEEDREL	Numeric	0.257	1.018
PER_TYP	Numeric	1.257	0.454
SEAT_POS	Numeric	13.699	10.539
REST_USE	Numeric	7.699	18.232
AIR_BAG	Numeric	21.075	18.272
DRDISTRACT	Numeric	56.203	46.854

Table 1: All Numeric Variables with Mean and Standard Deviations

Variable	**Data Type**	**Nominal Values**
DAY_WEEK	Nominal	Weekday, Weekend
MAN_COLL	Nominal	Front-to-rear, Angle, Not Collision, Sideswipe - Same Direction, Front-to-front,[...],Sideswipe - Opposite Direction, Rear-to-Side, Other, Unknown, Rear-to-Rear
WRK_ZONE	Nominal	Construction, Maintenance, Work Zone Others, Utility
LGT_COND	Nominal	Daylight, Dark-lighted, Dark-Not lighted, Dusk, Dawn, Dark-Unknown lighting, Light Others
WEATHER	Nominal	Clear, Cloudy, Rain, weather Others, Snow,[...], Severe Crosswinds, Sleet or Hail,

		Blowing Snow, Freezing Rain or Drizzle, Blowing Sand, Soil, dirt
INT_HWY	Nominal	No, Yes
UNDEROVERRIDE	Nominal	Underride, Underride Others, Override
ROLLOVER	Nominal	No Rollover, Rollover, Tripped by object, Rollover Others, Rollover, Untripped
DEFORMED	Nominal	No Damages, Major Damages, Deformed Others, Minor Damages
VEH_ALCH	Nominal	Veh Alcohol Others, No Alcohol
VSPD_LIM	Nominal	Speed Limit, vspd_lim Others, No Statutory limit
VSURCOND	Nominal	Wet, Snow, Vsurcond Others, Dry, Ice/Frost, Sand, Water, Slush, Oil
AGE	Nominal	Adults, Grown-up, Old, Teenager, Kids, Less than a year
SEX	Nominal	Male, Female, Binary
INJ_SEV	Nominal	No Injuries, Low Injury, High Injury, Unknown
EJECTION	Nominal	Not Ejected, Ejection Others, Totally Ejected, Partially Ejected
DRUGS	Nominal	Other Drugs, Drugs involved
BODY_TYP_binned	Nominal	Vehicles, Tractor, MotorCycles, Body Type Others, Farm Equipment

Table 2: All Nominal Variables

Preparation of Data:

Overview:

This section encapsulates the pivotal and time-demanding phase of this project. Data preparation encompasses fundamental tasks inherent in any analytical project, including the meticulous selection of a data subset, partitioning it into training and test datasets, and addressing missing values through either removal or replacement. The KNIME tool boasts numerous built-in nodes designed to execute these tasks seamlessly, and we will thoroughly explore and elaborate on these nodes in the following sections.

Selecting the Data:

Out of the initial pool of 204 variables, we've narrowed down our focus to a subset of 30 variables. Numerous factors influence the severity of a driver, including light conditions, weather, road facilities, and more. Our analysis reveals that these specifically chosen variables have a more pronounced impact on driver severity. The compiled list of selected variables comprises:

DAY_WEEK , MAN_COLL, WRK_ZONE , LGT_COND, WEATHER , INT_HWY ,UNDEROVERRIDE, ROLLOVER , DEFORMED, VEH_ALCH , VSPD_LIM , VSURCOND , AGE , SEX , INJ_SEV , EJECTION , DRUGS , BODY_TYP_binned, SPEEDREL, PER_TYP, SEAT_POS, REST_USE, AIR_BAG , DRDISTRACT, YEAR, NUMOCCS, HIT_RUN, MOD_YEAR, MAKE, BODY_TYP,DR_PRES.

These variables are deemed particularly influential in understanding and assessing driver severity.

Cleaning of Data:

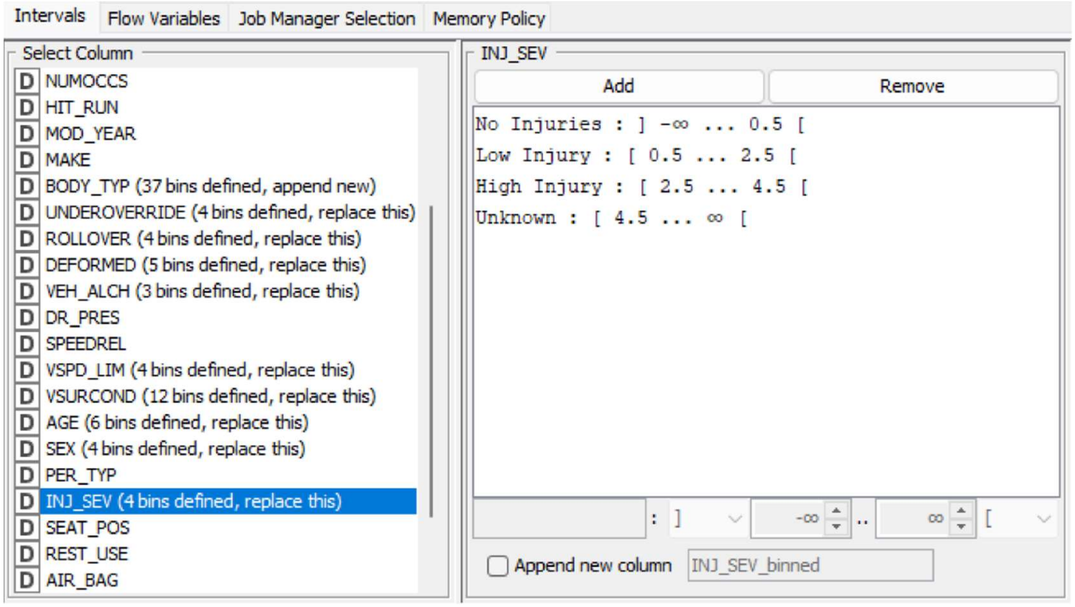
In this project we have used the some nodes to clean the data and get the appropriate data to find the severity of the data. The nodes are Column Filter, Numeric Binner, Row Filter, Rule Engine, Rule Based Row Filter, Math Formula, Outlier Removal.

Column Filter:

This node serves the purpose of discerning essential attributes. It identifies 30 attributes from the initial pool of 204 as required to the study.

Numeric Binner:

The Numeric Binner node in KNIME is employed to establish attribute bins, which involves grouping similar values into a single bin. The resulting bins can then be used to overwrite the original column or generate a new column. It allows the user to convert a continuous numeric variable into discrete categories, making it suitable for algorithms that handle categorical data more effectively. It allows for the customization of bin widths, enabling users to tailor the grouping of values based on domain knowledge or specific requirements of the analysis. The list of variables which are binned are:
DAY_WEEK, MAN_COLL, WRK_ZONE, LGT_COND, WEATHER, INT_HWY, UNDEROVERRIDE, ROLLOVER, DEFORMED, VEH_ALCH, VSPD_LIM, VSURCOND, AGE, SEX, INJ_SEV, EJECTION, DRUGS, BODY_TYP_binned. An example to show binning of INJ_SEV variable.



D NUMOCCS
D HIT_RUN
D MOD_YEAR
D MAKE
D BODY_TYP (37 bins defined, append new)
D UNDEROVERRIDE (4 bins defined, replace this)
D ROLLOVER (4 bins defined, replace this)
D DEFORMED (5 bins defined, replace this)
D VEH_ALCH (3 bins defined, replace this)
D DR_PRES
D SPEEDREL
D VSPD_LIM (4 bins defined, replace this)
D VSURCOND (12 bins defined, replace this)
D AGE (6 bins defined, replace this)
D SEX (4 bins defined, replace this)
D PER_TYP
D INJ_SEV (4 bins defined, replace this)
D SEAT_POS
D REST_USE
D AIR_BAG

```
No Injuries : ] -∞ ... 0.5 [
Low Injury : [ 0.5 ... 2.5 [
High Injury : [ 2.5 ... 4.5 [
Unknown : [ 4.5 ... ∞ [
```

:] -∞ .. ∞ [

☐ Append new column INJ_SEV_binned

Figure 3: Numeric Binning INJ_SEV Variable.

Row Filter:

The Row Filter node in Knime assumes a crucial role in the removal of specific or undesirable rows associated with a particular variable. In the case of the INJ_SEV variable, which is of particular interest, this node is applied to focus on rows categorized as 1, 2 (designated as Low Injury) and 3, 4 (designated as High Injury), excluding the Unknown and No

Injuries bins within INJ_SEV. This targeted filtration is achieved by configuring the Row Filter to perform a case-sensitive string match.

Rule Engine:

The Rule Engine node in KNIME is a powerful tool that allows users to define and apply rules to their data within a workflow. It enables rule-based data transformation, allowing users to specify conditions and actions for the dataset. To differentiate the driver in the SEAT_POS variable rule engine node is used where SEAT_POS = 11.

Missing Value Node:

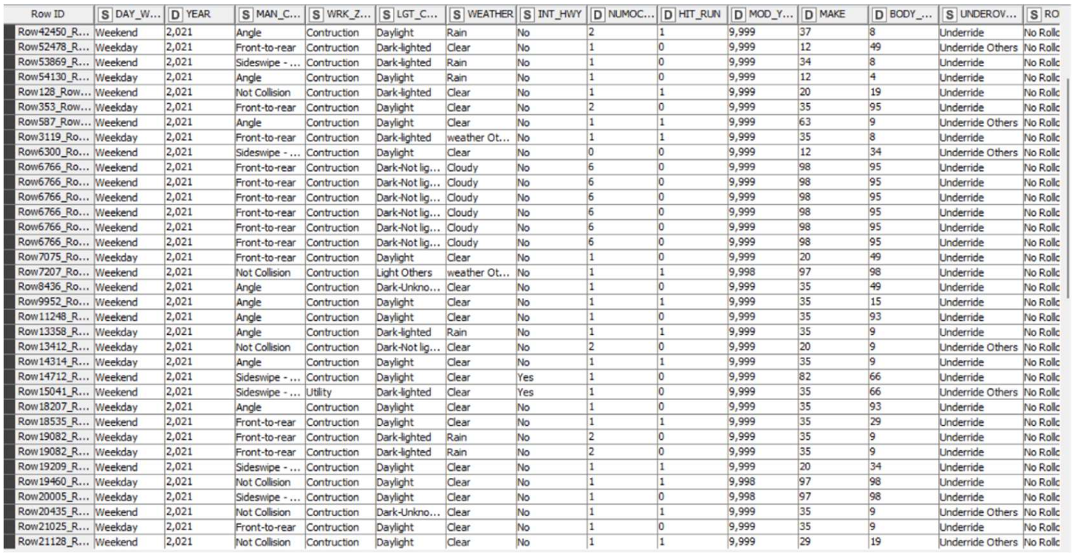
We replaced all the Unknown Values with the median of the respective variable. The missing value node in KNIME is used to manage missing values present in the dataset. To address this issue, numerical attributes are replaced with the rounded mean value. Alternatively, string attributes are substituted with a most frequent value known as mode.

Rule Based Row Engine:

This node is also similar to the row filter in filtering the unwanted rows and including the desired ones. BODY_TYP variable is used in the rule-based row filter to filter the auto mobiles and exclude the other vehicles like tractors, motorcycles, and farm equipement.

Math Formula:

Utilized for executing mathematical operations on variables, derived a distinct variable called CAR_AGE by making the difference of the YEAR and MOD_YEAR variables. Although the inclusion of this variable enhanced the overall model, a few instances with atypical ages in a small subset of cars necessitated the application of an outlier removal filter to rectify the issue.



Row ID	S DAY_W…	D YEAR	S MAN_C…	S WRK_Z…	S LGT_C…	S WEATHER	S INT_HWY	D NUMOC…	D HIT_RUN	D MOD_Y…	D MAKE	D BODY_…	S UNDEROV…	S RO
Row42450_R…	Weekend	2,021	Angle	Contruction	Daylight	Rain	No	2	1	9,999	37	8	Underride	No Rollc
Row52478_R…	Weekday	2,021	Front-to-rear	Contruction	Dark-lighted	Clear	No	1	0	9,999	12	49	Underride Others	No Rollc
Row53869_R…	Weekend	2,021	Sideswipe - …	Contruction	Dark-lighted	Rain	No	1	0	9,999	34	8	Underride	No Rollc
Row54130_R…	Weekday	2,021	Angle	Contruction	Daylight	Rain	No	1	0	9,999	12	4	Underride	No Rollc
Row128_Row…	Weekend	2,021	Not Collision	Contruction	Dark-lighted	Clear	No	1	1	9,999	20	19	Underride	No Rollc
Row353_Row…	Weekday	2,021	Front-to-rear	Contruction	Daylight	Clear	No	2	0	9,999	35	95	Underride	No Rollc
Row587_Row…	Weekend	2,021	Angle	Contruction	Daylight	Clear	No	1	1	9,999	63	9	Underride Others	No Rollc
Row3119_Ro…	Weekday	2,021	Front-to-rear	Contruction	Dark-lighted	weather Ot…	No	1	1	9,999	35	8	Underride	No Rollc
Row6300_Ro…	Weekend	2,021	Sideswipe - …	Contruction	Daylight	Clear	No	0	0	9,999	12	34	Underride Others	No Rollc
Row6766_Ro…	Weekend	2,021	Front-to-rear	Contruction	Dark-Not lig…	Cloudy	No	6	0	9,999	98	95	Underride	No Rollc
Row6766_Ro…	Weekend	2,021	Front-to-rear	Contruction	Dark-Not lig…	Cloudy	No	6	0	9,999	98	95	Underride	No Rollc
Row6766_Ro…	Weekend	2,021	Front-to-rear	Contruction	Dark-Not lig…	Cloudy	No	6	0	9,999	98	95	Underride	No Rollc
Row6766_Ro…	Weekend	2,021	Front-to-rear	Contruction	Dark-Not lig…	Cloudy	No	6	0	9,999	98	95	Underride	No Rollc
Row6766_Ro…	Weekend	2,021	Front-to-rear	Contruction	Dark-Not lig…	Cloudy	No	6	0	9,999	98	95	Underride	No Rollc
Row6766_Ro…	Weekend	2,021	Front-to-rear	Contruction	Dark-Not lig…	Cloudy	No	6	0	9,999	98	95	Underride	No Rollc
Row7075_Ro…	Weekday	2,021	Front-to-rear	Contruction	Daylight	Clear	No	1	0	9,999	20	49	Underride	No Rollc
Row7207_Ro…	Weekend	2,021	Not Collision	Contruction	Light Others	weather Ot…	No	1	1	9,998	97	98	Underride	No Rollc
Row8436_Ro…	Weekend	2,021	Angle	Contruction	Dark-Unkno…	Clear	No	1	0	9,999	35	49	Underride	No Rollc
Row9952_Ro…	Weekend	2,021	Angle	Contruction	Daylight	Clear	No	1	1	9,999	35	15	Underride	No Rollc
Row11248_R…	Weekend	2,021	Angle	Contruction	Daylight	Clear	No	1	0	9,999	35	93	Underride	No Rollc
Row13358_R…	Weekday	2,021	Angle	Contruction	Dark-lighted	Rain	No	1	1	9,999	35	9	Underride	No Rollc
Row13412_R…	Weekday	2,021	Not Collision	Contruction	Dark-Not lig…	Clear	No	2	0	9,999	20	9	Underride Others	No Rollc
Row14314_R…	Weekday	2,021	Angle	Contruction	Daylight	Clear	No	1	1	9,999	35	9	Underride	No Rollc
Row14712_R…	Weekend	2,021	Sideswipe - …	Contruction	Daylight	Clear	Yes	1	0	9,999	82	66	Underride	No Rollc
Row15041_R…	Weekend	2,021	Sideswipe - …	Utility	Dark-lighted	Clear	Yes	1	0	9,999	35	66	Underride Others	No Rollc
Row18207_R…	Weekday	2,021	Angle	Contruction	Daylight	Clear	No	1	0	9,999	35	93	Underride	No Rollc
Row18535_R…	Weekend	2,021	Front-to-rear	Contruction	Daylight	Clear	No	1	1	9,999	35	29	Underride	No Rollc
Row19082_R…	Weekday	2,021	Front-to-rear	Contruction	Dark-lighted	Rain	No	2	0	9,999	35	9	Underride	No Rollc
Row19082_R…	Weekday	2,021	Front-to-rear	Contruction	Dark-lighted	Rain	No	2	0	9,999	35	9	Underride	No Rollc
Row19209_R…	Weekend	2,021	Sideswipe - …	Contruction	Daylight	Clear	No	1	1	9,999	20	34	Underride	No Rollc
Row19460_R…	Weekday	2,021	Not Collision	Contruction	Daylight	Clear	No	1	1	9,998	97	98	Underride	No Rollc
Row20005_R…	Weekday	2,021	Sideswipe - …	Contruction	Daylight	Clear	No	1	0	9,998	97	98	Underride	No Rollc
Row20435_R…	Weekend	2,021	Not Collision	Contruction	Dark-Unkno…	Clear	No	1	1	9,999	35	9	Underride Others	No Rollc
Row21025_R…	Weekday	2,021	Front-to-rear	Contruction	Daylight	Clear	No	1	0	9,999	35	9	Underride	No Rollc
Row21128_R…	Weekend	2,021	Not Collision	Contruction	Daylight	Clear	No	1	1	9,999	29	19	Underride Others	No Rollc

Figure 4: Filtered data before designing the models

Modeling Techniques:

When we were creating our model, we used different methods to try to answer the main question of predicting and understanding how the injury severity affects the driver in a car accidents, and followed with the preparation of the dataset. We looked at the variable that shows if the injury is low or high i.e., INJ_SEV (dependent variable), and we treated it like a category with two options. Because this is a type of problem where we sort things into categories, there isn't just one model that works well with the dataset. So, we made six different models (DT, RF, ANN, LR, NB, GBT) to get the best and fair results. We also used logistic regression, a common statistical method, to handle the sorting challenge. This helped us get a good understanding of injury severity of driver in a car accident.

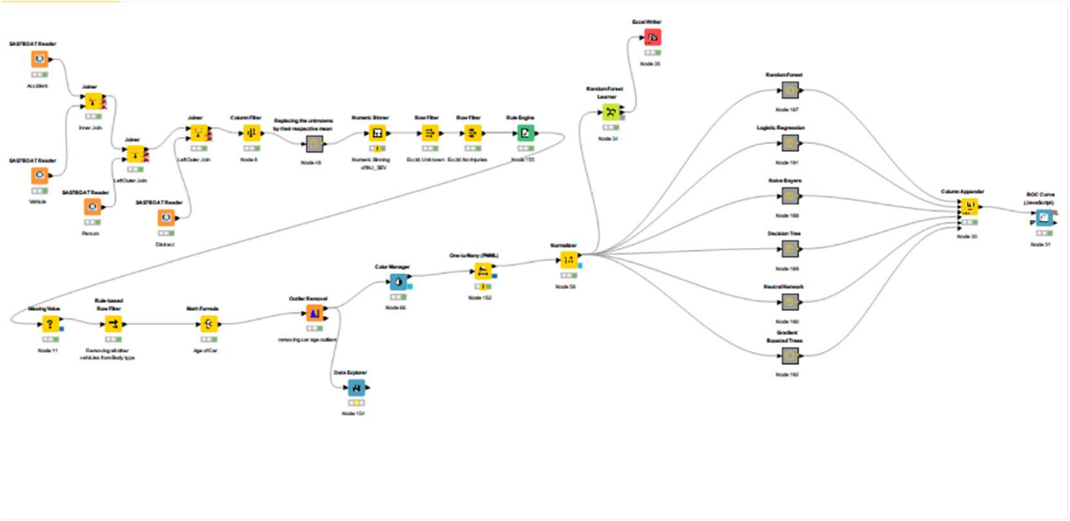


Figure 4: Overview of workflow of all six models

Artificial Neural Networks(ANN):

ANN is a modeling technique inspired from the human nervous system. The unique feature for this model is to establish a relationship between the dependent variables and independent variables that extracts complex knowledge from the data sets. It has three parts: an input layer, hidden layers (which can be more than one), and an output layer (also called a visible layer). The information moves from the input layer through the hidden layers to the output layer. The hidden layers make the input information more abstract before giving the final prediction from the output layer. This kind of neural network is often used when we want to put things into different groups or categories, like sorting inputs into different classes or labels.

In creating this model, we employed a one-to-many node, which altered values in a chosen column and generated a new column. Following this, a normalization node was utilized to standardize the input data for the neural networks, enhancing the model's performance. Subsequently, the Artificial Neural Network learner was employed for training the data, and testing was conducted using the predictor node. Additionally, a suffix variable was created for a specific value. Running this model enables us to achieve a high degree of accuracy in predicting the severity of injuries sustained by drivers in car accidents.

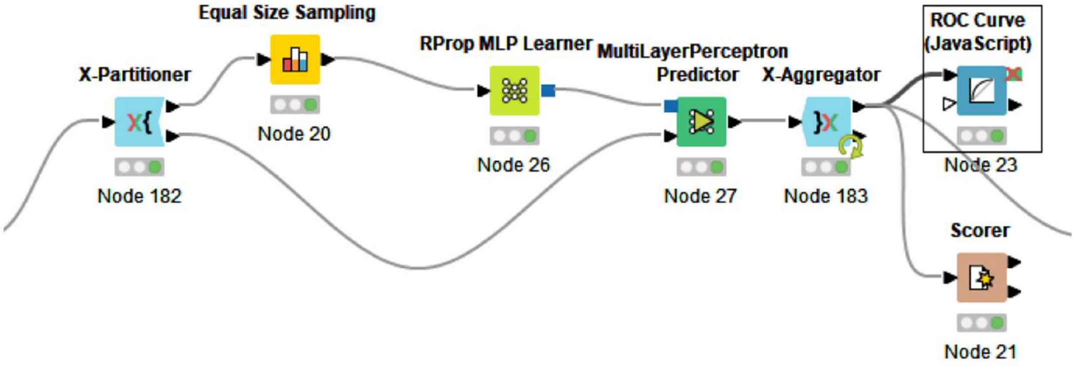


Figure 5: Designing model for ANN

Decision Tree:

This algorithm is most widely used for classification and regression models. It splits the data into subsets based on the features of the data set and creates a tree-like structure to predict the target variable. This process is repeated recursively for each subset until the criteria is met. This characteristic makes it a widely favored choice for constructing predictive models, serving as an accessible and easily interpretable starting point. Furthermore, the decision tree offers a way to pinpoint the most crucial variables within the classification framework.

In the decision tree implementation, steps were taken to enhance the model robustness by introducing the x-partitioner node at the beginning of a cross-validation loop. This step aims to address the issues of bias and overfitting of the dataset. Implementation of cross-validation techniques is significant and increases the model's ability to generalize the unseen data.

Furthermore, equal-sized sampling was implemented. This is a process that balances the dataset by eliminating the rows and ensures an even distribution of categorical columns. This preprocessing step plays a vital role in the removing the potential biases by contributing the overall accuracy.

After the preprocessing steps, we proceeded to train the decision tree learner and predictor node. Here the data is divided into training and testing sets. Finally, we collected all the data from the predictor node by using the x-aggregator node which allows us to compare the predicted class outcomes with the actual ones.

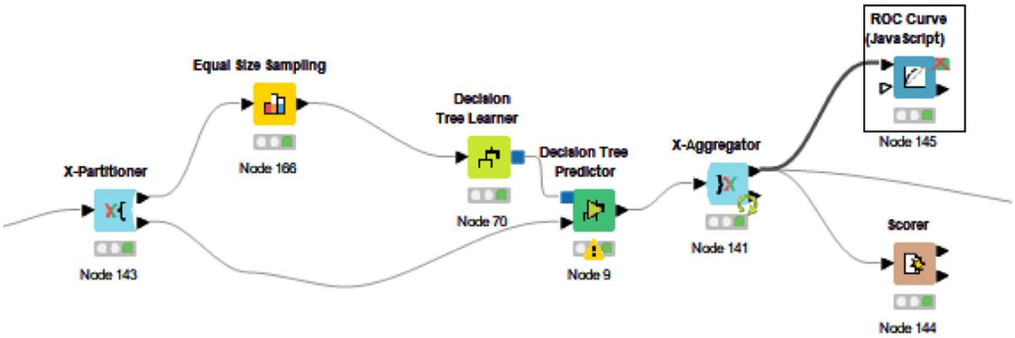
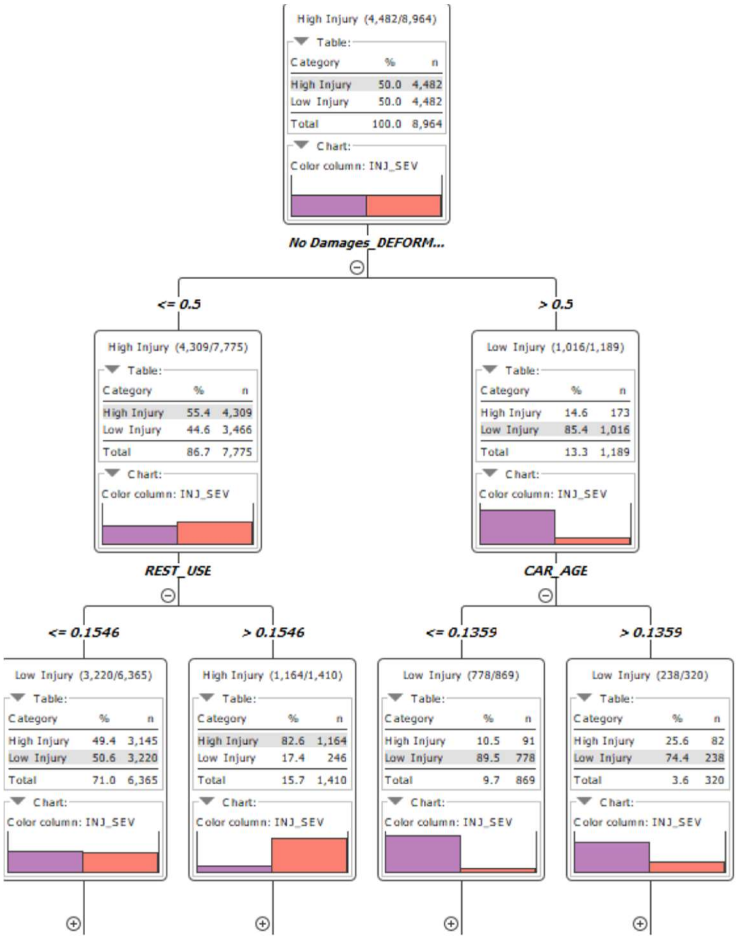


Figure 6: Designing Model for Decision Tree



High Injury (4,482/8,964)

Category	%	n
High Injury	50.0	4,482
Low Injury	50.0	4,482
Total	100.0	8,964

Color column: INJ_SEV

High Injury (4,309/7,775)

Category	%	n
High Injury	55.4	4,309
Low Injury	44.6	3,466
Total	86.7	7,775

Color column: INJ_SEV

Low Injury (1,016/1,189)

Category	%	n
High Injury	14.6	173
Low Injury	85.4	1,016
Total	13.3	1,189

Color column: INJ_SEV

Low Injury (3,220/6,365)

Category	%	n
High Injury	49.4	3,145
Low Injury	50.6	3,220
Total	71.0	6,365

Color column: INJ_SEV

High Injury (1,164/1,410)

Category	%	n
High Injury	82.6	1,164
Low Injury	17.4	246
Total	15.7	1,410

Color column: INJ_SEV

Low Injury (778/869)

Category	%	n
High Injury	10.5	91
Low Injury	89.5	778
Total	9.7	869

Color column: INJ_SEV

Low Injury (238/320)

Category	%	n
High Injury	25.6	82
Low Injury	74.4	238
Total	3.6	320

Color column: INJ_SEV

Figure 7: Shows the 3 level split of Decision Tree.

The three-level decision tree demonstrates that it is trying to predict the driver's injury severity at different stages of the accident, from the initial impact to the outcome. It also shows the percentage of people who sustained high and low injuries. It depicts that a higher percentage of people sustain high injuries at the initial impact stage, while a higher percentage of people sustain low injuries at the hospitalization and discharge stages.

This information can be used to develop strategies to reduce the number and severity of injuries in car accidents. For example, it may be beneficial to focus on improving safety features in vehicles and developing better protocols for responding to car accidents.

Random Forests:

This approach combines multiple decision trees to enhance the accuracy and robustness of predictions, constituting an ensemble method. It is constructed on different subsets of datasets. During the training process, each decision tree is trained independently with its own subset of the data set. Once all the trees are trained, they are combined into the final prediction.

The set-based random forest algorithm proves highly effective in addressing the challenges associated with the decision tree algorithm, including issues such as overfitting and sensitivity to outliers. This algorithm exhibits versatility in handling diverse datasets and stands out as an excellent option for tackling intricate classification problems. It surpasses individual decision trees by mitigating overfitting concerns and enhancing robustness, making it particularly suitable for datasets with varying complexities.

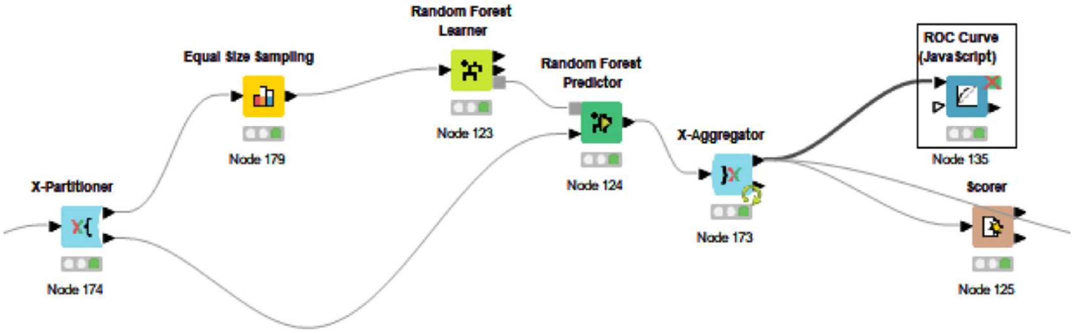


Figure 8: Designing model of Random Forest

Moreover, the set-based random forest algorithm is adept at capturing intricate interactions between variables, offering a nuanced understanding of the factors influencing the outcomes. Its ability to provide insights into the key determinants makes it an ideal choice for unraveling complex relationships within data, contributing to more accurate and reliable classification results.

Boosted Trees:

Boosted trees are a technique that assembles numerous weaker models to form a robust model. The fundamental concept behind boosting trees is to progressively introduce new models to the ensemble, with each one aimed at rectifying the mistakes of the preceding models. The ultimate model is a weighted combination of all the individual models. This procedure is carried out iteratively, with each subsequent model concentrating on addressing the errors of the prior models.

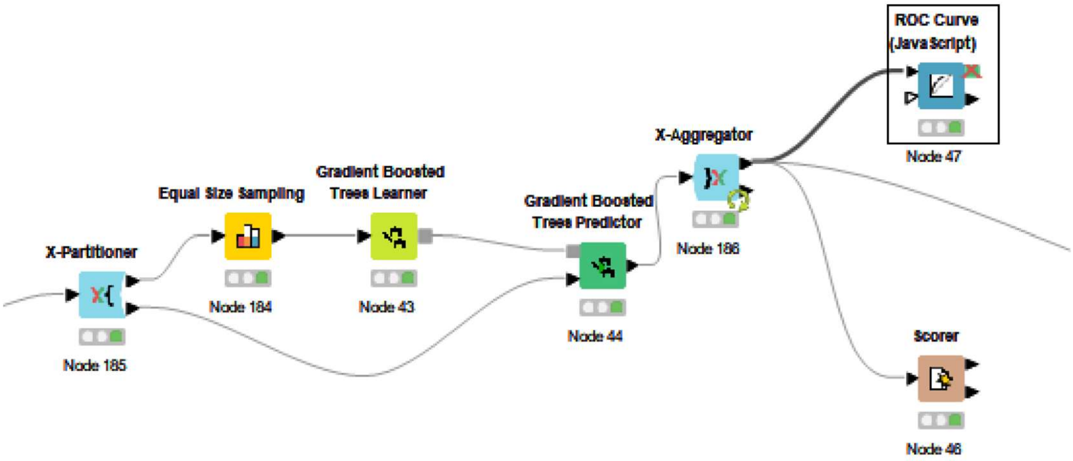


Figure 9: Designing model for Boosted Trees

Logistic Regression:

Logistic regression is a statistical method employed to predict binary outcomes, such as a positive or negative outcome, using past observations from a dataset. This model assesses the connection between one or more independent variables within the dataset to anticipate dependent variable.

In addressing our binomial classification problem, where the goal was to explore the factors influencing injury severity in car accidents, logistic regression emerged as the most widely employed statistical technique. This method was utilized to predict the binary dependent variable, which is injury severity.

However, to integrate logistic regression into our workflow, we introduced a one-to-many node before the normalizer. This node transformed values in a chosen column, creating a new column that held the processed data in multiple transformed columns as the output. Logistic regression is known for efficiently uncovering the relationship between independent variables and the dependent variable. Its ability to predict the probability of an outcome makes it well-suited for our classification problem. Moreover, logistic regression aids in identifying significant predictors, offering valuable insights for decision-making purposes.

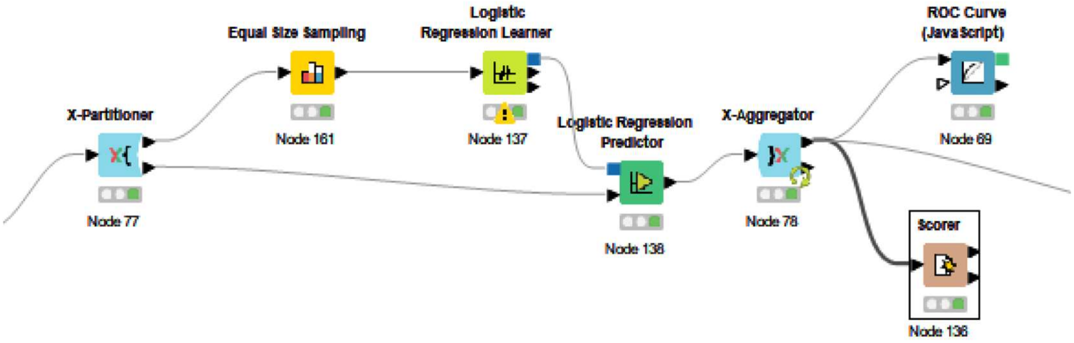


Figure 10: Designing model for Logistic Regression.

Naive Bayes:

This classification approach aimed to create a model for how inputs are distributed in a specific class and assign them to instances of a problem. Unlike the models we used earlier, the naive Bayes classifier makes a strong assumption of attribute independence. This means it assumes that the value of a certain attribute doesn't depend on other attributes in the given context. This assumption makes it relatively easy to put into practice and computationally efficient.

A significant advantage of the naive Bayes classifier is that it doesn't need a large amount of training data to estimate parameters and perform classification. This makes it particularly useful in situations where there's only a limited amount of available data. The process of building this classifier followed a workflow similar to the previous two models, the MLP and logistic regression.

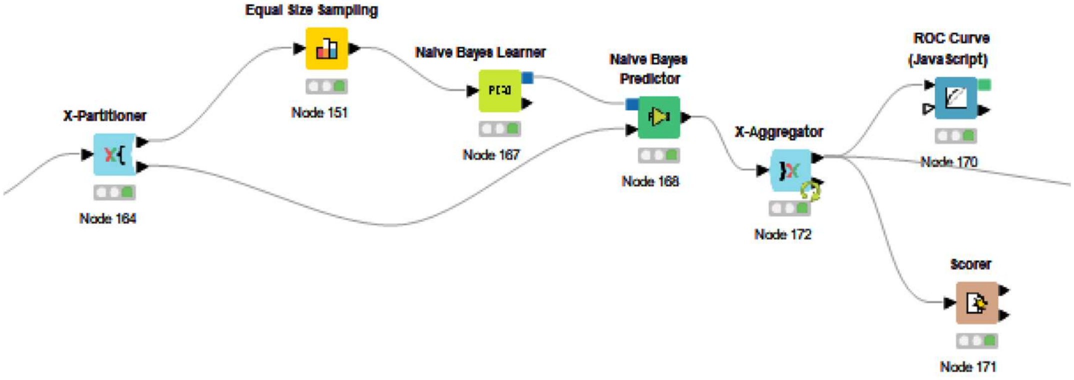


Figure 11: Designing model for Naïve Bayer.

Evaluating the Models:

For every model we have some particular terms to evaluate the models they are the accuracy, Sensitivity, specificity and roc on curve.

Model	Accuracy %	Sensitivity	Specificity	ROC Curve
ANN	70.64	0.697	0.709	0.777
DT	64.67	0.653	0.645	0.661
RF	73.67	0.699	0.745	0.797
BT	72.43	0.715	0.726	0.797
LR	71.31	0.698	0.717	0.779
NB	76.13	0.818	0.824	0.725

Table 3: Showing Accuracy, Sensitivity, Specificity and ROC Value.

After running all the models, the resulting matrix illustrates how accurately each model predicted the outcome. On the lower end of the spectrum, our Decision Tree classifier achieved a respectable accuracy of up to 64.67%. On the high end, the Naïve Bayer model demonstrated outstanding performance with an accuracy of 76.13%. This model focuses on regression-based classification. The second highest is Random Forests with 73.67%. The Random Forest model specializes in regression-based classification, creating a network of decision trees that operate within the context of regression.

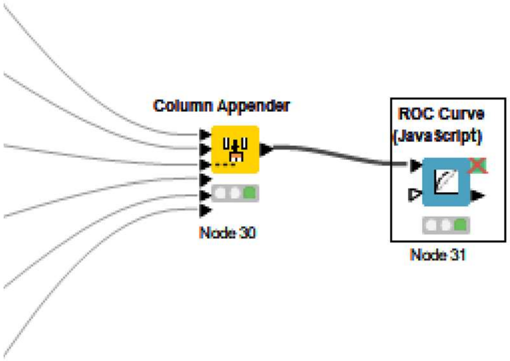


Figure 12: Inputs to Column Appender and ROC Curve.

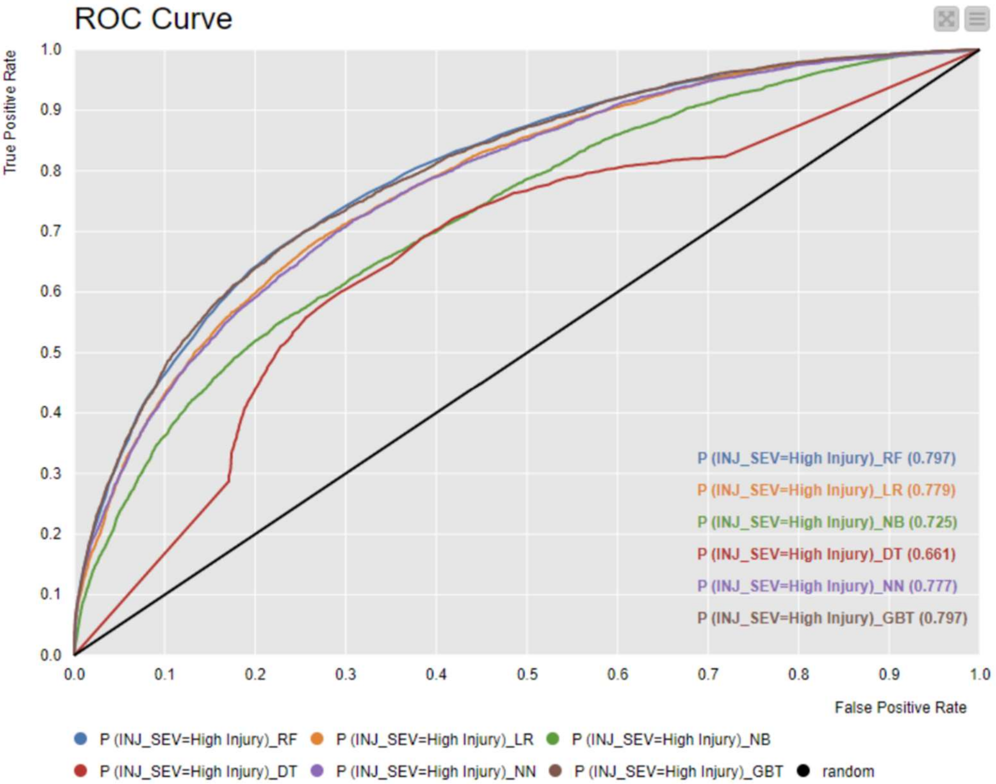


Figure 13: ROC Curve of all models

Figure 14 illustrates the variables' significance concerning whether a crash resulted in severe injury. Unsurprisingly, factors like EJECTION and REST_USE rank high in importance, aligning with common public awareness about the importance of seat belts and proper child seating. Although not explicitly stated, the chart below suggests that individuals who are smaller or lighter may face an increased risk of ejection or airbag-related injuries, or both. This implies a potential positive correlation among these top variables.

However, there's a surprising drop in variable importance for factors like drug usage and hit-and-run compared to the previously mentioned variables. According to our model, Drugs are slightly less important than age in determining injury severity. Notably, the seat position variable ranked surprisingly low on the importance chart. It was anticipated that, similar to other factor-related variables such as car age and airbags, seat position would play a significant role in the data skew. Contrarily, it was found to have minimal impact on the results. This lack of importance could be attributed to the physics of a car crash, which might overshadow the influence of variables like weather and time of day on crash severity.

Variable Importance:

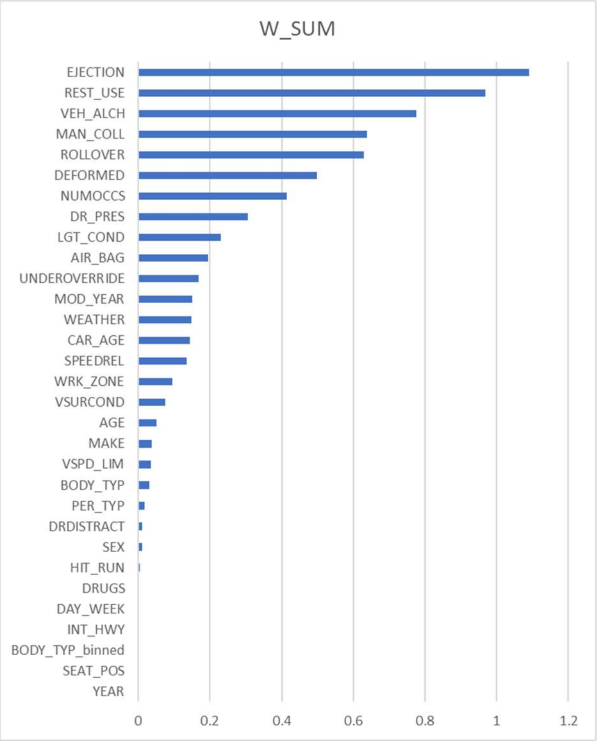


Figure 14: Graph shows the variables with the most impact on injury severity

Deployment:

In the final step of the data mining process, we organized and shared what we learned so that important people can make smart decisions in the future. In our case, the information we found will be part of a system that helps decision-making. This way, the decision makers can keep improving and collecting important information to deal with safety issues in cars and human injuries.

Our main goal in this project was to find out what makes car accidents more dangerous and share that information with the car industry and important people like manufacturers, government officials, and insurance companies. We want to help reduce the clear risks to people's lives and property that come from these accidents.

Additionally, we suggest that car makers use our most successful model, the Naive Bayes, at least once a year to adjust to changes in how safe cars need to be. Once they do this, they should include our findings in their yearly updates, future plans for investing money, and projects for developing new technology. This will help reduce the risks linked to dangerous car accidents. Government officials are also crucial in making a big impact in local areas and across the whole country. They can use what we found to create or change laws about traffic, making them more effective and encouraging vehicle safety. Even though car crashes and how severe injuries can be is a big problem, the important people involved, both in the government and in companies, can actively help solve it by making new rules and changes based on what we found. This will make roads safer in the future.

Of the variables that demonstrated significance, restraint use stands out as one that can be controlled prior to the crash. Understandably governments have attempted to increase seatbelt use through campaigns such as “click it or ticket.” Additionally, car manufacturers enhanced seatbelt reminder (ESBR) systems that flash or audibly signal when an occupant is not wearing a seatbelt. These systems increase the likelihood of individuals wearing a seatbelt.[iii]

Conclusion:

The ongoing and important problem of road safety, both in other countries and at home, has become a major focus for study among scholars, government officials, car makers, and insurance companies. From a scholarly point of view, as we got real data from the National Highway Traffic Safety Administration, a part of the US Department of Transportation, we realized how crucial it is to spend enough time understanding and preparing the data before we could start making any predictions. How well we worked with the data played a big role in how accurate the six different models we created were. These models were all about answering the main question: figuring out what factors make injuries more severe in car accidents.

We made six different models, both with numbers and sets, we discovered that using equal-size sampling to deal with uneven data, along with K-fold cross-validation, was crucial in getting a well-balanced and optimal result for all the models. Out of all the outcomes, the Naive Bayes model gave the best result with 76.13% accuracy, 82.8% sensitivity, 81.4% specificity, and 72.5% on the ROC curve. This means it was not only the most accurate model, but it also had the best balance overall. Moving on to different model specifications, it was crucial to find out which factors had the most impact on our study about road safety and reducing human injury in various levels of car crashes. When thinking about a car crash, there are many things to consider, both inside and outside the car, but our study focused on the main factors: whether someone was thrown from the car (EJECTION) and if they had a seatbelt or some form of body restraint (REST_USE). Other important factors included details about how the person was thrown, if an airbag was used, and the damage to the car. However, these factors didn't play a big role in the outcome compared to our main three variables.

Finally, it's important for us to use what we learned by sharing the information with important people who can use our results to make better decisions in their areas of work. To achieve the goal of making public roads safer and reducing injuries, we hope that the impact of this project can help in making future decisions that improve rules and bring innovation for a safer future on the road.

References

i. Dong, X., Xie, K., & Yang, H. (2022). How did COVID-19 impact driving behaviors and crash Severity? A multigroup structural equation modeling. *Accident; analysis and prevention*, *172*, 106687. Retrieved from: [https://doi.org/10.1016/j.aap.2022.106687]

ii. National Center for Statistics and Analysis. (2023, April). *Early estimate of motor vehicle traffic fatalities in 2022* (Crash Stats Brief Statistical Summary. Report No. DOT HS 813 428). National Highway Traffic Safety Administration.

iii. National Highway Traffic Safety Administration. (2009, February). *Effectiveness and Acceptance of Enhanced Seat Belt Reminder Systems: Characteristics of Optimal Reminder Systems*. Retrieved from [https://www.nhtsa.gov/sites/nhtsa.gov/files/811097.pdf]

iv. National Center for Statistics and Analysis. (2023, April). Crash Report Sampling System analytical user's manual, 2016-2021 (Report No. DOT HS 813 436). National Highway Traffic Safety Administration.