MSIS 5633 - PREDICTIVE ANALYTICS TECHNOLOGIES

24129 – In Class

Motor Vehicle Crash Injury Severity Analysis

Final Project Report

Due Date:

April 27, 2025

By:

Carson Jones, Saidabi Karumanchi, Jayashri Jayaraman, Siddam Reddy Vamsi Krishna Reddy

Team Page

Jones, Carson

Jayashri Jayaraman

Siddam, Vamsi

Saidabi Karumanchi

Executive Summary

This project entailed the use of predictive analytics to determine predictors of injury severity in motor vehicle crashes. My team utilized data from the National Highway Traffic Safety Administration's Crash Report Sampling System. This dataset encompasses comprehensive information regarding vehicle crashes, driver actions, crash situations, and injuries in the United States.

The goal was to identify patterns in crash data that have the capability to predict the likelihood of severe or deadly injuries. With an understanding of these patterns, the project provides insight that can render public safety efforts more effective, allow first responders to know what to expect, and guide resource allocation.

We followed the CRISP-DM process, as outlined below:

- Data Collection & Understanding: We merged four datasets (Accident, Vehicle, Person, Distraction) with a combined total of over 95,000 records, and selected key variables such as seatbelt use, vehicle speed, driver type, road type, and crash time.
- Data Preparation: Using the KNIME Analytics Platform, we cleaned and preprocessed the data into ~37,800 high-quality records, derived new features (e.g., vehicle age, overspeed).
- Modeling & Evaluation: Five classification models have been built and tested: Gradient Boosted Trees (Best Model: AUC = 0.79, Accuracy = 0.72%), Decision Tree, Logistic Regression, Multilayer Perceptron (ANN), Random Forest and Naive Bayes.
- Gradient Boosted Trees emerged as the top performer and predicted the most prominent predictors such as seatbelt usage, point of impact, age of driver, and distraction status.

This project shows how advanced analytics can support data-driven decision making for public safety and public transit. The predictive model can be used to Assist policy making through the identification of driver risk behavior and crash situation, guide public outreach campaigns toward seatbelt compliance, distracted driving, or crash conspicuity, inform vehicle design considerations for manufacturers who want to enhance safety elements, and contribute to insurance risk modeling by projecting injury outcomes according to crash scenarios.

This project shows how data science can be used to move traffic safety from react to prevent. It empowers executives, city planners, and first responders with the ability to make more intelligent, better-timed interventions that could save lives.

1. Business Understanding

Motor vehicle accidents continue to be a severe public health issue in the United States, with a mean of approximately more than 6 million crashes annually. In these, nearly 40,000 lives are lost every year, and many others suffer serious injuries or lifelong disability. Despite ongoing efforts to improve automobile safety and to implement highway codes, injuries from accidents continue to increase in both number and severity, indicating the need for more incisive insight and for improved preventive strategies.

This project is driven by a basic question: What are the risk factors that determine the severity of injury from automobile collisions? By answering this, it is hoped that the knowledge of the dynamics of crash outcomes will be furthered and proactive efforts towards reducing injury rates and saving lives will be supported.

In pursuit of this question, the project explores several categories of factors that are understood to influence injury severity:

- Accident-based characteristics like the nature of collision (e.g., head-on, rear-end), pre-impact movement of the vehicle, restraint usage, and impact point.
- Demographical characteristics like the age of the driver, sex, and role (driver, passenger, pedestrian).
- Environmental conditions like the time of day, light, road surface, weather, and type of location (urban or rural).
- Technical aspects of the vehicle like the type of vehicle, safety features, and year of manufacture.

By analyzing these variables, we expect to find patterns and correlations that will enable us to predict the likelihood of serious injury in a collision scenario. This information is of immense worth to a great number of stakeholders including traffic safety organizations, city planners, automobile manufacturers, insurance companies, and public health organizations.

The ultimate business goal is not only to build a high-accuracy predictive model but to extract actionable insights that can guide policymaking, inform infrastructure planning, and influence public awareness campaigns. With predictive analytics and machine learning applied via the CRISP-DM approach, this project will transform raw crash data into actionable knowledge with concrete real-world impacts.

2. Data Understanding

In a bid to correctly forecast and model the causes of injury severity in car crashes, one must first have a clear understanding of the datasets. This is the stage of the CRISP-DM process where the data is collected, its structure is comprehended, its contents are reviewed, and its quality and completeness are verified before advancing to data preparation and modeling.

The data used in this project has been sourced from the National Highway Traffic Safety Administration (NHTSA) Crash Report Sampling System (CRSS). The CRSS is the national representative sample of police-reported motor vehicle crashes that occurred in the United States. The years included in this data are 2016 to 2021 and are comprised of over one relational table. In this project, the following four most important datasets have been used:

Accident file: Contains crash-level data such as date and time of the crash, location, road, lighting, and weather during the crash.

Vehicle file: Data for every involved vehicle such as vehicle type, point of first impact, deployment of safety systems, and maneuver leading up to the crash.

Person file: Holds information regarding the individuals involved in the accident, including age, gender, seat position, severity of injury, and safety equipment use (e.g., seatbelt, helmet).

Distract file: Records distracted driving observed or reported, like using a cell phone, interacting with passengers, or other in-car distractions.

The data set has several relational files Accident, Vehicle, Person, and Distract transcribing over 200,000 crashes across five years. After the merging of the data sets via unique identifiers (case_id, vehicle_id, and person_id), a master table containing roughly 198 entries and 37 selected variables, including a combination of numerical and categorical types, was produced. The first exploration in KNIME using the Data Explorer node revealed missing values in vital attributes such as distraction status and restraint system use, which were resolved through filtering and imputation techniques. Severity of injury, the response variable, was class-imbalanced with predominant cases in "no injury" or "minor injury" and relatively fewer in the "severe" or "fatal" categories. Variability distribution chiefly for vehicle year, timing of crash, and age was found to be right-skewed and hence transformed using normalization techniques (min-max scaling and z-score) before pre-processing the data into scale-sensitive models, i.e., ANN and SVM.

2.1 Explore Data

This project was concerned with applying predictive analytics to find predictors of injury severity in car crashes. Our group worked with data from the National Highway Traffic Safety Administration's Crash Report Sampling System. This data set contains detailed information on car crashes, driver actions, crash conditions, and injuries nationwide.

The main goal was to identify patterns in crash data that can predict the likelihood of serious or fatal injuries. Based on the patterns, the project provides data that can enhance public safety programs, allow first responders to be aware of the situation, and guide resource planning.

We used the CRISP-DM methodology, and the steps are as follows:

- Data Collection & Understanding: We merged four datasets (Accident, Vehicle, Person, Distraction) with over 95,000 records, and selected key variables such as seatbelt use, vehicle speed, driver demographics, road type, and crash time.
- Data Preparation: Using KNIME Analytics Platform, we preprocessed and converted the data into ~37,800 high-quality records, created new features (e.g., vehicle age, overspeed)
- Modeling & Evaluation: We built and tested five classification models: Gradient Boosted Trees (Best Model: AUC = 0.79, Accuracy = 0.72%), Decision Tree, Logistic Regression, Multilayer Perceptron (ANN), Random Forest and Naive Bayes

Gradient Boosted Trees performed the best and revealed important predictors such as seatbelt use, point of impact, driver age, and distraction status.

This project illustrates how advanced analytics can be used to support data-driven public safety and transportation decision-making. The predictive model can be used to Inform policy design by identifying high-risk driver behavior and crash scenarios, inform public education campaigns aimed at seatbelt use, distracted driving, or crash visibility, inform vehicle design choices for manufacturers who are determined to maximize safety features, and improve insurance risk modeling by predicting injury outcomes based on crash profiles.

This book illustrates how data science can revolutionize traffic safety from a reactive response to prevention. Equipping executives, city planners, and first responders to make smarter, more effective interventions that save lives.

2.2 Validate Integrity

During the data prep phase, we observed several critical data quality issues that needed to be resolved before modeling. Most prominent among these was missing values in certain features like Person_Age_Missed, VSPD_LIM_Miss, and TRAV_SP_Miss. These were dealt with using mean or median imputation based on the nature of the variable. The approach was adopted to keep the variables without inducing bias due to deletion.

Some of the data was simply unusable. It was cleaned and prepared if it had to be sued, but it was discarded prior to modeling to prevent noise from being introduced. Nevertheless, it might be reinstated if it were found to be necessary.

While the dataset did contain common quality problems, these were addressed through simple cleaning and transformation methods, leaving us with a structured and balanced dataset suitable for modeling.fperson

3. Data Preparation:

We carefully screened all the available variables in the combined crash database and selected the ones we considered would be best at predicting the severity of injuries in our project. This decision was made collaboratively as a team.

Features that we included ranged across numerous categories related to the crash, the people, the vehicles, and the location. These consisted of driver characteristics: sex, age, drinking. Vehicle-related descriptors: vehicle age, deformation, airbag deploy. And crash setting: type of road, light condition, type of collision, distraction, speed, and work zone on or off. Most were derived or reformulated for better interpretability or forecasting ability. Such as converting raw time into numerical form, calculation of whether speed was within post speed limit, or binning age into groups.

A full list of retained and transformed variables is shown in the subsequent figures. The variables were chosen as they were of good quality, understandable, and closely associated with the problem of estimating crash severity.

On the other hand, we eliminated several variables from the original dataset. These were removed for a range of reasons. Some had too many missing or vague values, others were administrative or identifier codes without predictive capability, and a couple were too specialized or redundant. Including these would have added noise to the models without improving performance.

Lastly, our decisions were collectively and based on selecting features that were clean, relevant, and helpful in predicting the severity of a crash.

3.1 Select Key Data

For this project, we carefully reviewed all available variables from the merged crash dataset and selected the ones we believed would be most relevant to predicting injury severity. The selection process was a team decision.

The features we kept covered a broad range of categories related to the crash, the people involved, the vehicles, and environmental conditions. These included driver demographics: sex, age, alcohol use. Vehicle-related attributes: vehicle age, deformation, airbag deployment. And crash context: road type, light condition, collision type, distraction, speed, and work zone status. Many of these features were either derived or transformed for better interpretability or predictive power. for example, converting raw time to a numerical format, calculating whether the speed was within the posted limit, or binning age into groups.

A full list of retained and transformed features is shown in the figures below. These variables were chosen because they were high quality, interpretable, and closely tied to the problem of understanding crash severity.

On the other hand, we excluded several variables from the original dataset. These were removed for several reasons. Some had too many missing or unclear values, others were identifiers or administrative codes with no predictive value, and a few were overly specific or redundant. Including these would have added noise to the models without improving performance.

Ultimately, our decisions were made as a group and guided by selecting features that were clean, relevant, and helpful for predicting the severity of a crash. (Fig 1 below)

Variable Name	**Description**	**Type**	**Descriptive Statistics**	**%Miss/Unknown**
VE_TOTAL	Total number of vehicles involved	Numeric	-	0.0%
BODY_TYP	Type of vehicle body (e.g., Sedan, SUV)	Nominal	-	0.0%
INJ_SEV_	Injury severity (Major / Minor)	Binary	Low_Inj: 81%, High_Inj: 19%	0.0%

VEH_AGE	Age of the vehicle at crash time	Numeric	9.21 (6.96)	0.0%
DAY_WEEK	Day of week (Weekday / Weekend)	Nominal	-	0.0%
URBANICITY	s geographical area urban or rural	Nominal	Urban: 77%, Rural: 23%	0.0%
MAN_COLL	Manner of collision (categorized, missing handled)	Nominal	Angle: 36%, Front-to-Rear: 28%	0.1%
LGT_COND_	Lighting condition at crash (e.g., Daylight/Night)	Nominal	Daylight: 67%, Dark-Lighted: 18%	0.1%
WEATHER	Weather condition during crash (e.g., Clear/Rainy)	Nominal	Clear: 73%, Cloudy: 14%	2.7%
REL_ROAD	Crash relation to roadway (e.g., On Road/Off Road)	Nominal	On Roadway: 81%, On Roadside: 18%	0.0%
GVWR	Gross vehicle weight rating class	Nominal	-	0.0%
ALC_RES	Alcohol test result available/missing	Binary	-	31.0%
DRUGS	Drug test information available/missing	Binary	-	55.5%
IMPACT 1	Impact type (e.g., Front, Rear, Side)	Nominal	-	0.0%
ROLLOVER	Rollover event occurred/missing	Binary	-	0.0%
DISTRACT	Driver distraction information missing/known	Nominal	Not Distracted: 39%, Inattentive: 3%	54.3%
PERSON_AGE	Age of person (driver/passenger) or missing flag	Numeric	-	0.0%
VSPD_LIM_	Posted speed limit (missing handled)	Numeric	43.94 (12.87)	11.8%
TRAV_SP_	Vehicle traveling speed at crash (missing handled)	Numeric	29.07 (16.90)	50.5%
HARM_EV_	Most harmful event in crash (grouped)	Nominal	-	0.0%
P_CRASH1	First harmful event pre-crash (grouped)	Nominal	Going Straight: 56%, Stopped on Road: 13%	0.1%
P_CRASH2_	Second harmful event (grouped)	Nominal	-	0.0%
P_CRASH3_	Third harmful event (grouped)	Nominal	-	0.0%
P_CRASH4_	Fourth harmful event (grouped)	Nominal	-	0.0%
P_CRASH5_	Fifth harmful event (grouped)	Nominal	-	0.0%
NUMOCCS	Number of occupants or missing indicator	Numeric	-	0.0%
AIR_BAG	Airbag deployment status or missing	Nominal	-	0.0%

REST_USE	Restraint use (seatbelt) status or missing	Nominal	Should & Lap Belt Used: 84%, None Used 8%	5.6%
VEH_DFRM	Vehicle deformation severity or missing	Nominal	Disabling 62%, Functional: 12%	17.7%
HARMFUL_EVENT	Specific harmful event code or type	Nominal	-	0.0%
LANE_TYPE	Type of lane (e.g., Straight, Curved)	Nominal	-	0.0%
SEASON	Season of the year at crash time	Nominal	-	0.0%
WORK_ZONE	Crash occurred in a workzone (Yes/No)	Binary	None: 98%, Construction: 1%	0.0%
EJECTION	Ejection from vehicle (or missing)	Binary	Not Ejected: 97%, Yes: 1%	1.0%
HAZ_MAT	Involvement of hazardous materials	Binary	100%, Yes: 0.004%	0.1%

Fig: 1 List of Variables Used

3.2 Clean Data

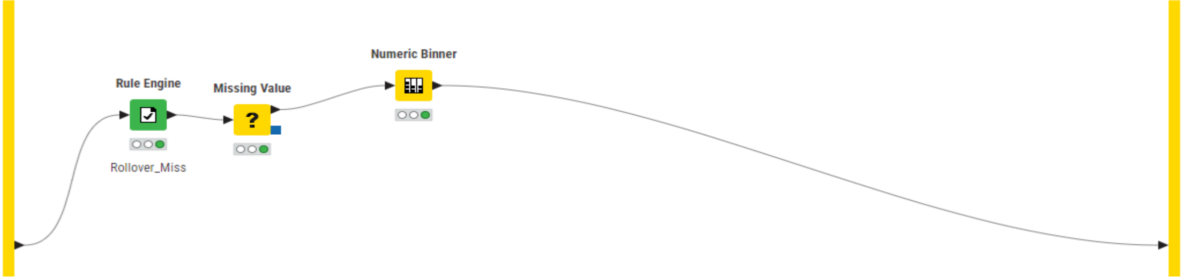


Fig: 2 A detailed cleaning process of each variable.

After verifying the quality of our dataset, we moved forward with a detailed cleaning process to ensure that the data was accurate, consistent, and suitable for modeling. This step is critical in preparing the dataset to deliver accurate results.

Our primary focus during this phase was handling missing or inconsistent values and transforming raw data into meaningful features. Which can be seen in the figure above.

In many cases, values were either missing, incorrectly coded, or fell outside a reasonable range. For example, several numeric fields like driver age and vehicle count included placeholder values such as “999” or numbers outside a plausible range. We used Rule Engine nodes in KNIME to flag or remove these invalid entries. Once flagged, Missing Value nodes were applied to fill in the gaps using appropriate imputation techniques such as the median for numerical fields and mode for categorical ones.

In addition to imputation, we transformed many variables through binning and recoding. This helped simplify the dataset while preserving key patterns. For instance, categorical fields like restraint use, hospital transport, and airbag deployment were grouped into broader, interpretable categories such as “Belt Used” or “Not Deployed.” Binning helped reduce data noise and made the features easier for models to understand.

We also made new features to improve predictive value. One example was the Overspeed variable, which was calculated by subtracting the posted speed limit from the actual travel speed. This created a more useful binary feature to indicate whether the vehicle was speeding at the time of the crash. Derived binary flags like Within Speed Limit, Is Collision, and Is Work Zone were also created to simplify model input and emphasize impactful condition.

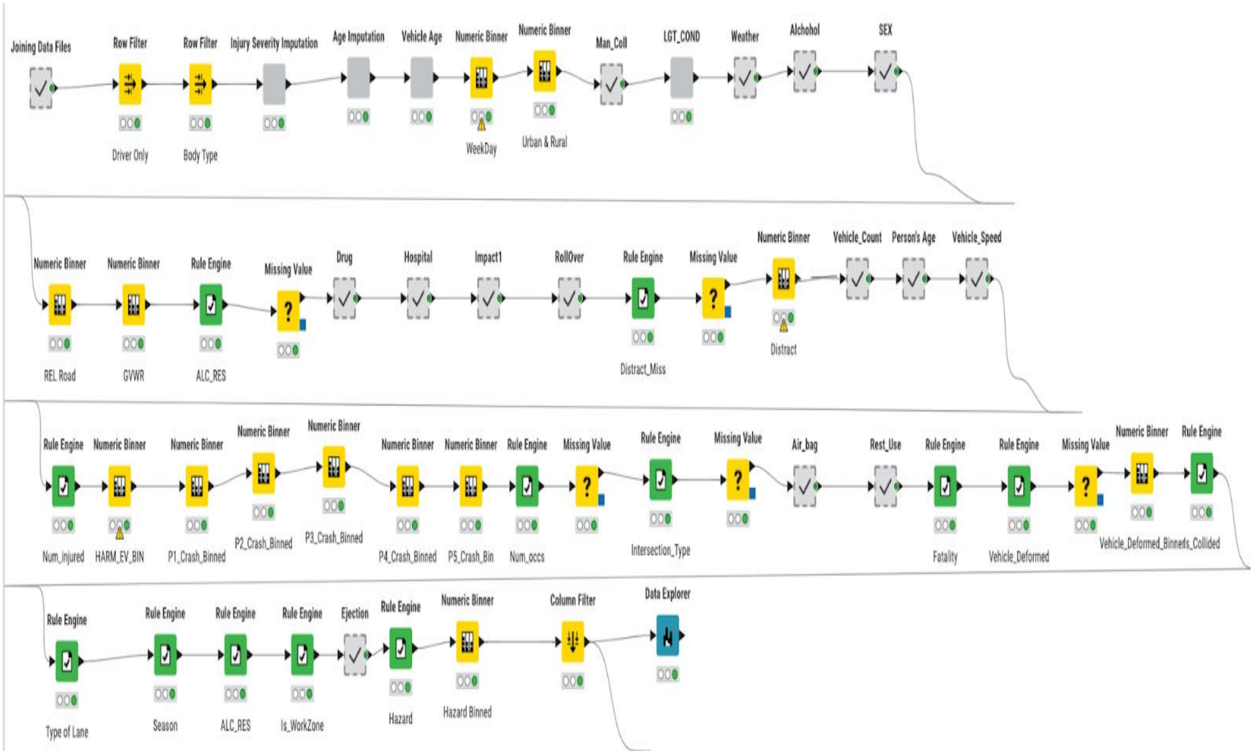


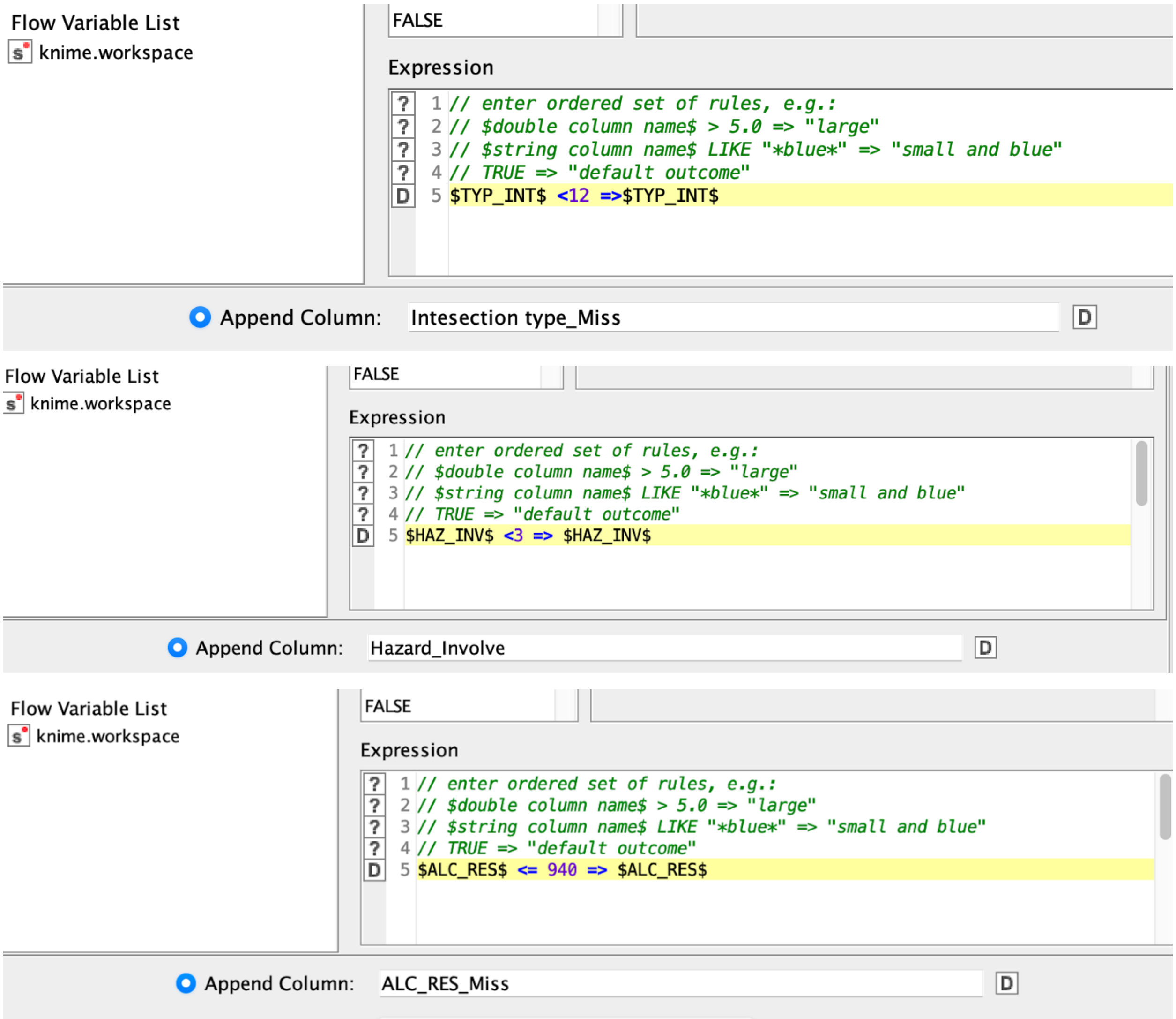
Fig 3: upper half of workflow, preprocessing.

Above is our preprocessing. The greyed-out nodes are containers that hold specific data cleaning to improve readability which can be seen in Fig: 3.

At the end of the cleaning process, we had a well-structured dataset with minimal missing values and no obvious outliers. All transformations were consistent across the dataset and designed to preserve the integrity of the original information. This cleaned version of the dataset was then used as the foundation for building and evaluating predictive models.

3.1.1 Rule Engine

The Rule Engine node played a crucial role in the development of the target variable for predictive modeling. It was designed to convert the original injury severity variable (INJ_SEV) into a simpler binary class system: "Minor" and "Severe" injuries. This conversion reduced model complexity and aligned the output with supervised learning algorithms. Furthermore, the Rule Engine facilitated flexible logic to create custom features from domain knowledge, e.g., combining car types, weather, or age ranges, and improved model interpretability and relevance.



Flow Variable List
knime.workspace

FALSE

Expression

```
// enter ordered set of rules, e.g.:
// $double column name$ > 5.0 => "large"
// $string column name$ LIKE "*blue*" => "small and blue"
// TRUE => "default outcome"
$TYP_INT$ <12 =>$TYP_INT$
```

● Append Column: Intesection type_Miss

Flow Variable List
knime.workspace

FALSE

Expression

```
// enter ordered set of rules, e.g.:
// $double column name$ > 5.0 => "large"
// $string column name$ LIKE "*blue*" => "small and blue"
// TRUE => "default outcome"
$HAZ_INV$ <3 => $HAZ_INV$
```

● Append Column: Hazard_Involve

Flow Variable List
knime.workspace

FALSE

Expression

```
// enter ordered set of rules, e.g.:
// $double column name$ > 5.0 => "large"
// $string column name$ LIKE "*blue*" => "small and blue"
// TRUE => "default outcome"
$ALC_RES$ <= 940 => $ALC_RES$
```

● Append Column: ALC_RES_Miss

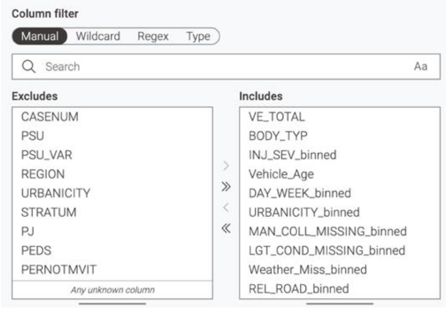
Fig: 4 These are some of Rule engines used for various variables.

3.1.2 Missing Value

Missing data management is critical in building strong prediction models, and the Missing Value node assisted in this respect. Upon comparing all four datasets, the missing values were found to be very few, but where they occurred, strategic steps were followed. For some datasets, those rows having missed crucial data (e.g., no record of injury or no vehicle type) were discarded. For others, default strategies like filling with median (in case of numeric) or mode (in case of categorical) were adopted. This guaranteed the dataset stayed intact and model-ready without creating bias or inconsistency.

3.1.3 Column Filter

To improve the modeling process and reduce computational expense, the Column Filter node was utilized to remove unimportant, redundant, or high-cardinality columns. Such columns are unique identifiers (case ID), timestamp fields, or descriptive text with no predictive value. By removing these non-contributory columns, the node improved model performance, decreased overfitting, and allowed the rest of the feature set to focus on the most informative variables that influence crash injury severity.



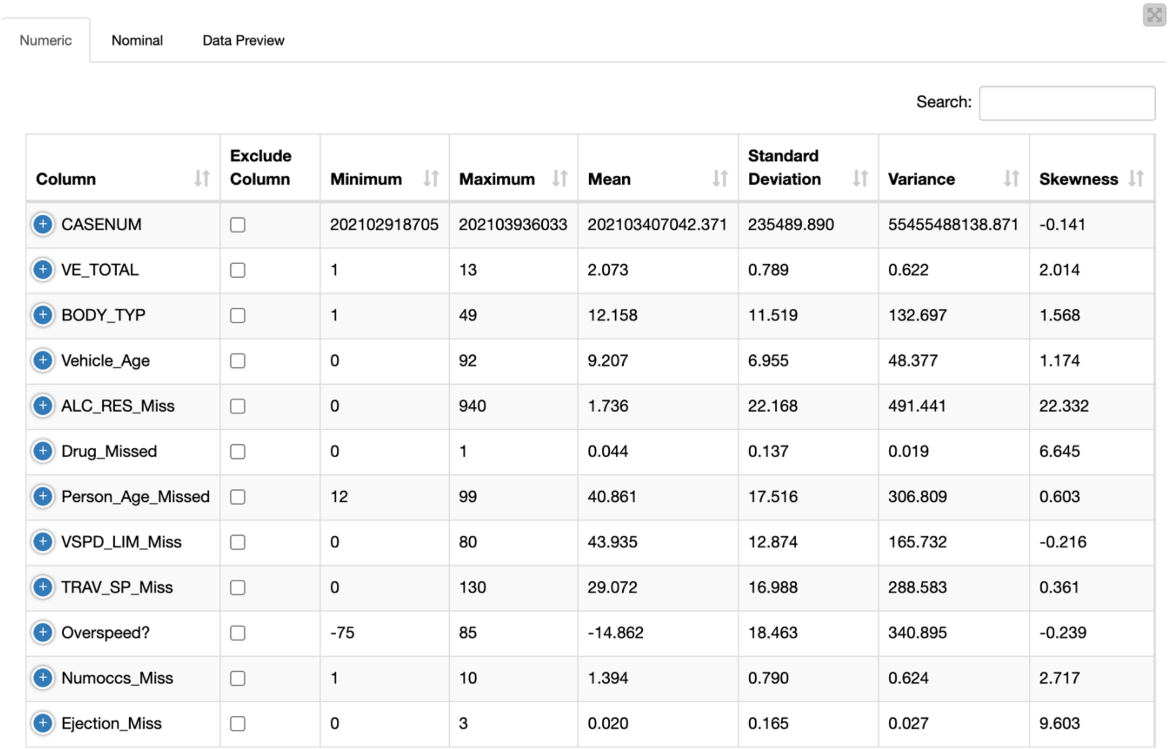
Column filter

Excludes	Includes
CASENUM	VE_TOTAL
PSU	BODY_TYP
PSU_VAR	INJ_SEV_binned
REGION	Vehicle_Age
URBANICITY	DAY_WEEK_binned
STRATUM	URBANICITY_binned
PJ	MAN_COLL_MISSING_binned
PEDS	LGT_COND_MISSING_binned
PERNOTMVIT	Weather_Miss_binned
Any unknown column	REL_ROAD_binned

Fig: 5 Column Filter

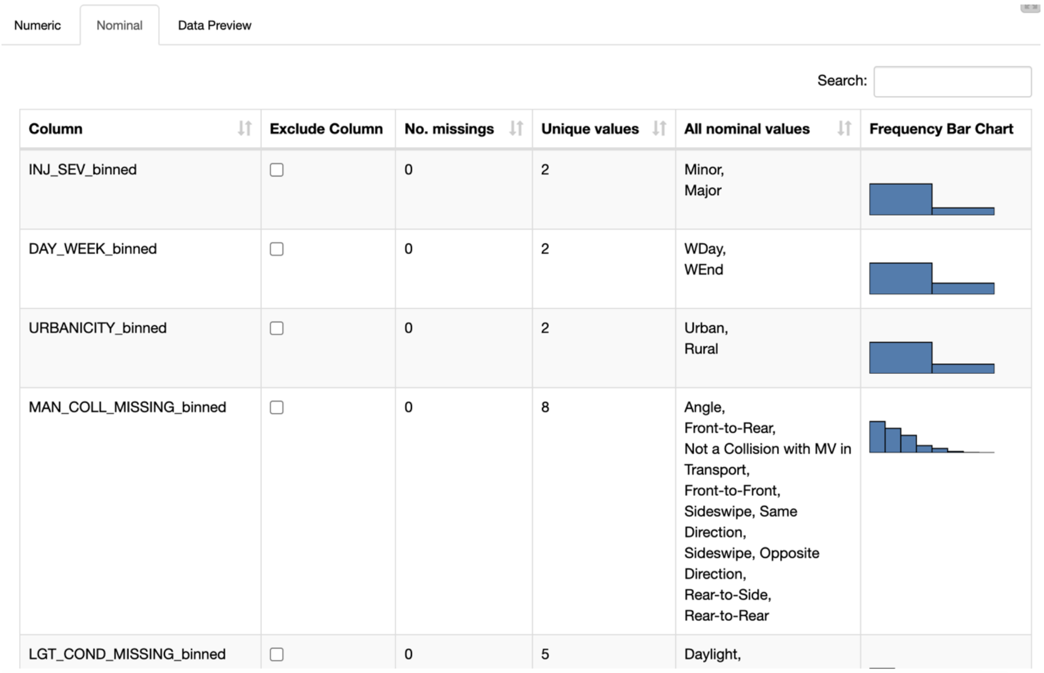
3.1.4 Data Explorer

Data Explorer node provided an entire interface to conduct initial statistical exploration of datasets. It was used to view the distribution, skewness, missing values, correlations, and value frequencies for both input variables and target variables. The exploration was done to test assumptions, identify class imbalances (such as having more Major injuries compared to major injuries), and discover anomalies. All these affected follow-up decisions related to balancing the data, transformation of features, and normalization, particularly for sensitive algorithms towards variable scales and distributions.



Column	Exclude Column	Minimum	Maximum	Mean	Standard Deviation	Variance	Skewness
CASENUM	□	202102918705	202103936033	202103407042.371	235489.890	55455488138.871	-0.141
VE_TOTAL	□	1	13	2.073	0.789	0.622	2.014
BODY_TYP	□	1	49	12.158	11.519	132.697	1.568
Vehicle_Age	□	0	92	9.207	6.955	48.377	1.174
ALC_RES_Miss	□	0	940	1.736	22.168	491.441	22.332
Drug_Missed	□	0	1	0.044	0.137	0.019	6.645
Person_Age_Missed	□	12	99	40.861	17.516	306.809	0.603
VSPD_LIM_Miss	□	0	80	43.935	12.874	165.732	-0.216
TRAV_SP_Miss	□	0	130	29.072	16.988	288.583	0.361
Overspeed?	□	-75	85	-14.862	18.463	340.895	-0.239
Numoccs_Miss	□	1	10	1.394	0.790	0.624	2.717
Ejection_Miss	□	0	3	0.020	0.165	0.027	9.603

Fig: 6 An overview of Data explorer

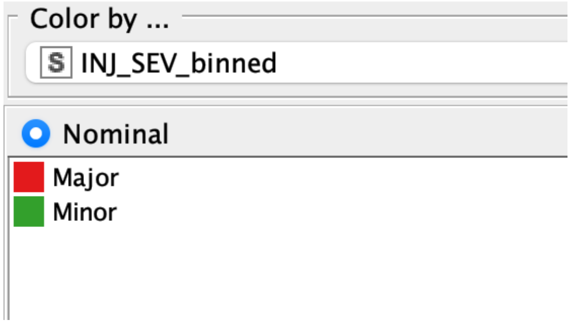


Column	Exclude Column	No. missings	Unique values	All nominal values	Frequency Bar Chart
INJ_SEV_binned	□	0	2	Minor, Major	
DAY_WEEK_binned	□	0	2	WDay, WEnd	
URBANICITY_binned	□	0	2	Urban, Rural	
MAN_COLL_MISSING_binned	□	0	8	Angle, Front-to-Rear, Not a Collision with MV in Transport, Front-to-Front, Sideswipe, Same Direction, Sideswipe, Opposite Direction, Rear-to-Side, Rear-to-Rear	
LGT_COND_MISSING_binned	□	0	5	Daylight,	

Fig: 7 Nominal values

3.1.5 Color Manager

The Color Manager node added to the visual analytics aspect of the workflow by adding informative color schemes to categorical variables, especially target class labels. For instance, "Minor" injuries can be colored green and "Severe" injuries colored red, improving clarity when plotting outcomes like scatter plots, confusion matrices, or ROC curves. Though outside of the core modeling semantics, this node helped facilitate more transparent communication of conclusions at evaluation time and allowed stakeholders to see model results and variable relationships more quickly.



Color by ...

S INJ_SEV_binned

● Nominal

🟥 Major

🟩 Minor

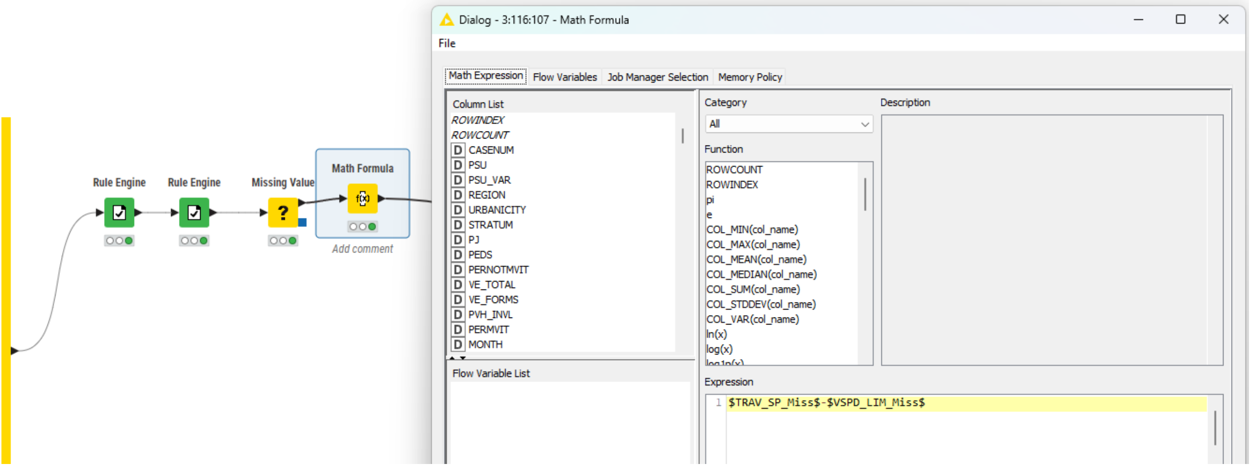
Fig: 8 Color Manager

3.3 Construct Data

As part of our data preparation and transformation process, we created several new features, or derived attributes, to help our models better recognize patterns and understand key relationships within the dataset. These new features were constructed from one or more existing attributes and were designed to either simplify complex data, standardize values across records, or extract more meaningful signals from raw inputs. By doing so, we enhanced the clarity and consistency of the information used in model training.

One of the more straightforward examples was the derivation of vehicle age, which we calculated by subtracting the vehicle's model year from 2021. This gave us a continuous and intuitive measure of how old the vehicle was at the time of the crash, an important factor when considering safety performance or mechanical reliability. Similarly, we created a new variable called Crash_Time, where we transformed separate hour and minute values into a single decimal value (hour + minute/60). This allowed us to model crash timing more fluidly, especially when exploring patterns based on time-of-day.

Another key feature, Within_Speed_Limit, was generated by comparing the travel speed (TRAV_SP) to the posted speed limit (VSPD_LIM). This resulted in a binary indicator showing whether the driver was speeding. We also created a numeric Overspeed feature, which gave the degree to which the driver exceeded the limit. These two variables gave us both a clear binary signal for model input and a more detailed continuous feature for interpretation and analysis. Which can be seen in figure 9 below.



Category: All

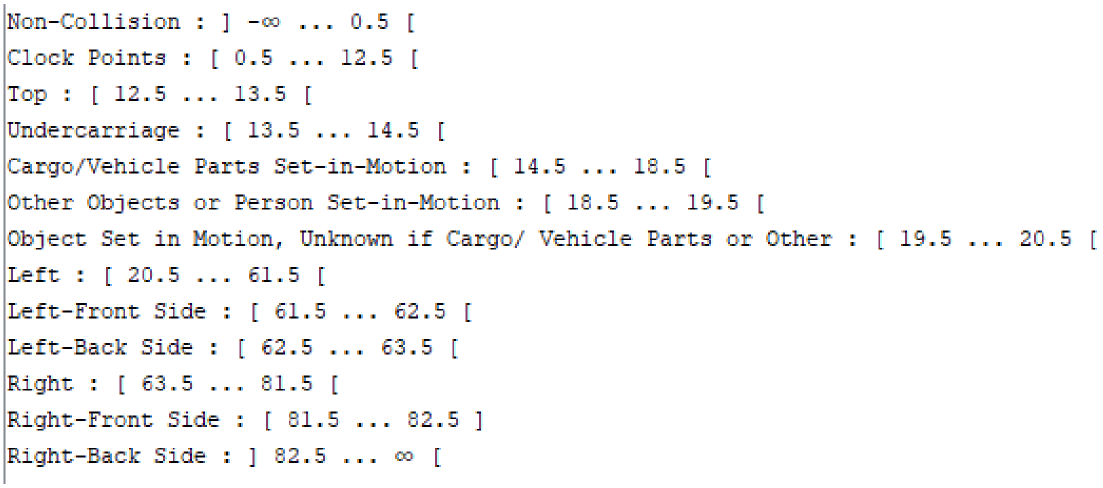
Expression:

```
$TRAV_SP_Miss$-$VSPD_LIM_Miss$
```

Fig: 9 Math Formula

We also constructed several derived flags to capture environmental and contextual information, such as IS_COLLISION (indicating whether the crash involved a collision or not), IS_WORKZONE (whether the crash occurred in a work zone), and Season (a categorical variable derived from the month of the crash). These features allowed us to represent important but scattered information in a consistent, model-friendly format.

In addition to deriving features from multiple fields, we applied binning techniques to variables such as age, impact point, and others to reduce variability and improve interpretability. This process grouped continuous or coded numerical values into meaningful categories, which helped simplify model training and minimize noise in predictions. Figure 10 below shows how impact point was created.



```
Non-Collision : ] -∞ ... 0.5 [
Clock Points : [ 0.5 ... 12.5 [
Top : [ 12.5 ... 13.5 [
Undercarriage : [ 13.5 ... 14.5 [
Cargo/Vehicle Parts Set-in-Motion : [ 14.5 ... 18.5 [
Other Objects or Person Set-in-Motion : [ 18.5 ... 19.5 [
Object Set in Motion, Unknown if Cargo/ Vehicle Parts or Other : [ 19.5 ... 20.5 [
Left : [ 20.5 ... 61.5 [
Left-Front Side : [ 61.5 ... 62.5 [
Left-Back Side : [ 62.5 ... 63.5 [
Right : [ 63.5 ... 81.5 [
Right-Front Side : [ 81.5 ... 82.5 ]
Right-Back Side : ] 82.5 ... ∞ [
```

Fig: 10 Number Binner

These constructed features were crucial to improving the quality of our inputs. They helped reduce noise, filled in important contextual gaps, and ensured greater consistency across records. They also aligned the structure of the data with the requirements of classification modeling, where simpler and cleaner inputs often produce more robust results.

No completely new records were generated as part of this phase. Each row in the final dataset continued to represent a real, unique crash event. Instead of expanding the number of records, our focus was on enhancing each one, transforming raw fields into useful indicators that would support more accurate and interpretable predictions of injury severity.

3.4 Integrate Data

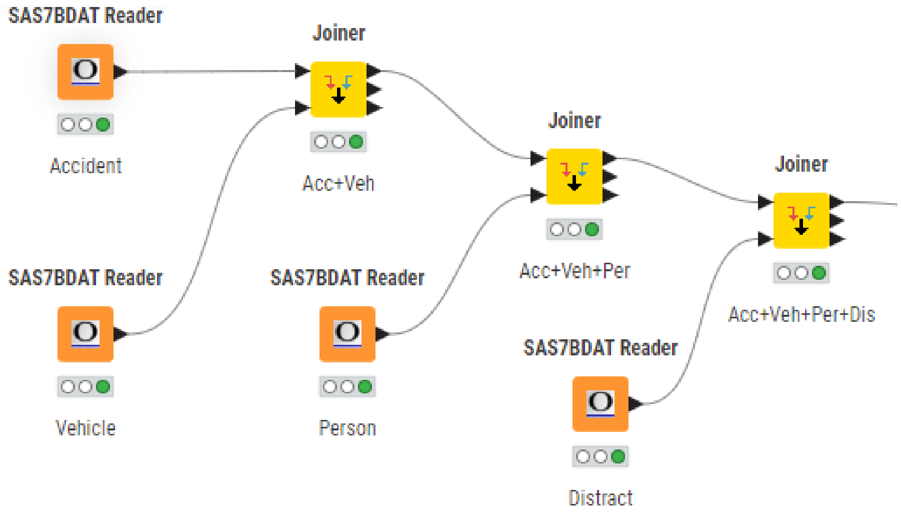


Fig 11: Data Joiner

Merging the data was the key to creating a combined dataset that would mirror an integrated snapshot of all the crash incidents. To this end, we combined four distinct tables Accident, Vehicle, Person, and Distract found in the NHTSA database, each with different perspectives. The Vehicle data was added to Accident data to incorporate technical crash data like deformation level and vehicle speed. Thirdly, demographic and health information of the involved individuals were added through the Person dataset. Lastly, the Distract table added behavioral context including distractions during crash. Unmatched rows were followed at each step in order not to lose data as well as to achieve integration quality.

Beyond simple joining of tables, we also created new variables through feature engineering and aggregation. A few examples include deriving "Overspeed" from travel speed minus posted limit, and comparing injured persons per crash. We aggregated fields such as total vehicles present, or report presence of airbags or critical deformations. In cases with more than one row per crash (e.g., many people or vehicles), we made conscious choices to aggregate at the crash level by logical or statistical aggregation. This conscious integration process guaranteed that our final dataset was complete and meaningful, capturing mechanical, human, behavioral, and environmental dimensions of each accident. It provided the structural basis for all the following modeling and analysis.

3.5 Format Data

Before training any models, the dataset structure needed to be technically correct and KNIME-compatible for machine learning. We re-ordered the columns so that the target variable INJ_SEV_binned would always appear in the same location, so it was easier to choose and validate when establishing models. To eliminate any potential bias, especially for sequence-sensitive models, we applied a Row Shuffle node to shuffle the rows randomly. This step prompted better overall generalization and reduced the likelihood of overfitting, particularly with ensemble models like Random Forest or Gradient Boosted Trees.

Final formatting and cleaning operations were also conducted to prepare the data for good model performance. These included the removal of special characters, whitespace trimming, and text field category standardization. Additionally, we ensured to initialize every variable with the correct data type: nominal for categorical, numerical for continuous, and

Boolean for binary flags. These formatting choices minimized downstream failure in processing as well as maximized model interpretability. At last, these formatting enhancements proved instrumental in mounting a refined, reliable dataset toward authentic and replicable predictive modeling.

4. Modeling Approach:

To determine the best-performing model for estimating injury severity in car crashes, a set of supervised learning classification methods were applied and compared. The performance of each model was evaluated with respect to accuracy, area under the Receiver Operating Characteristic curve, and feature importance where relevant. The target, injury severity, was represented as categorical labels, and all models were trained and validated on a uniform partitioning scheme to maintain comparability. To enhance the trustworthiness of the model further, k-fold cross-validation was employed in each model to the standard single partition.

We previously showed our workflow for data preparation, now here is the model side of the workflow, which demonstrates the entire workflow. Which is shown in Figure 12 below.

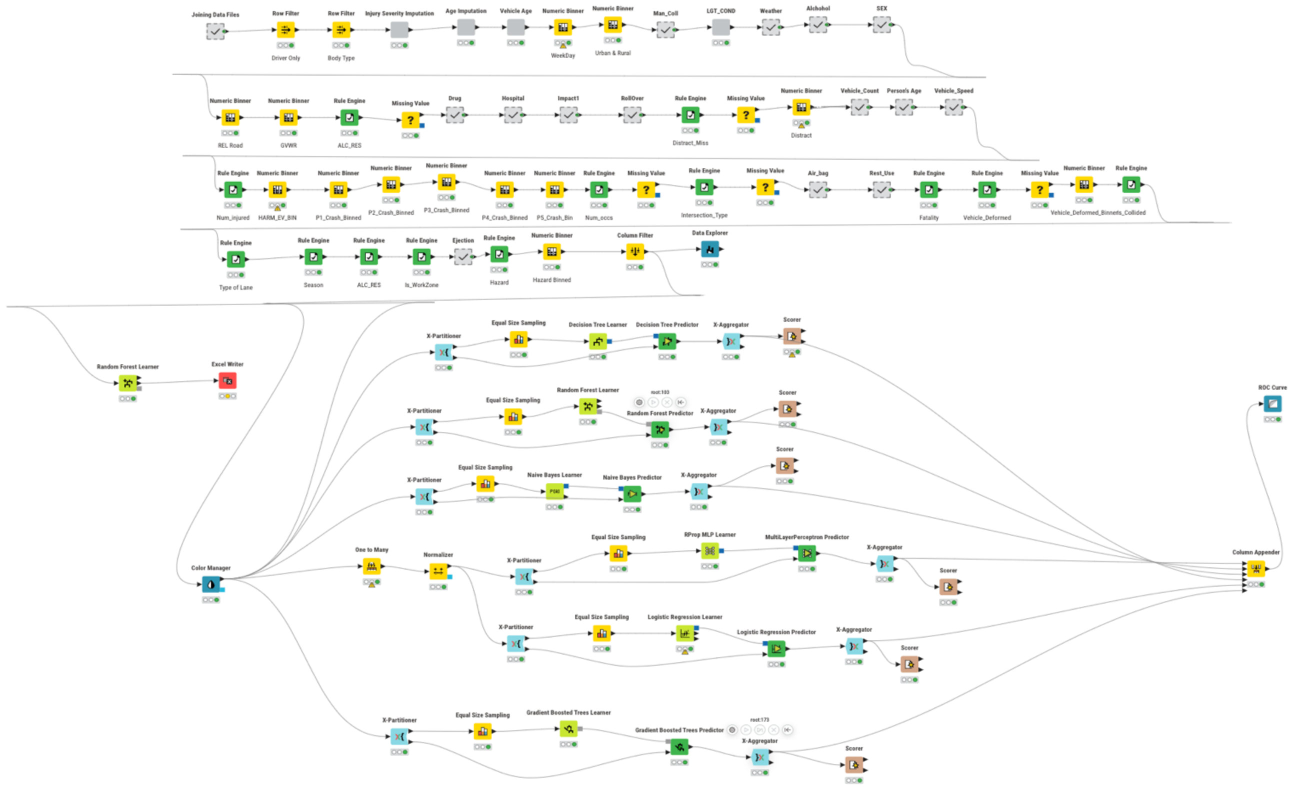


Fig 12: Models workflow

4.1 Select modeling technique

Decision Tree:

The Decision Tree classifier introduces an extremely explainable modeling methodology using dichotomization of the data based on most discriminating factors at each node. The model, when predicting injury severity, was able to generate a hierarchical rule base which could easily identify which of the features contributed to more serious injuries. Variables such as seatbelt use, driver age, light, and collision point were key contributors to the initial fractures of the tree, cementing their place as risk variables. Decision trees have the benefit of being able to handle both numerical and categorical variables without any preprocessing needs.

But the Decision Tree is vulnerable to overfitting, especially when the tree is very deep and captures noise in the training data. Although an accuracy of 0.62%, sensitivity of 0.61%, specificity of 0.63%, and ROC score of 0.62 the model did a poor job at generalizing to less frequent classes, e.g., Major or Minor injuries. The AUC of 0.62 shows moderate discriminatory power. While it has its limitations, the tree provided a helpful visual impression of how risk factors interact and can be particularly useful for public safety education or first-stage policy analysis.

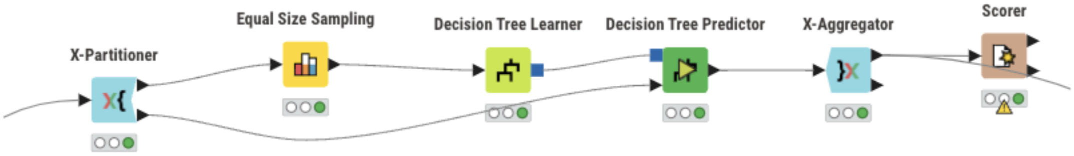
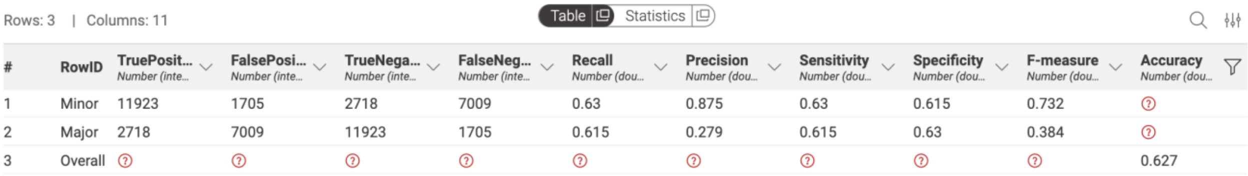


Figure 13: illustrates a Decision tree modeling workflow in KNIME. (above)



#	RowID	TruePosit... Number (inte...	FalsePosi... Number (inte...	TrueNega... Number (inte...	FalseNeg... Number (inte...	Recall Number (dou...	Precision Number (dou...	Sensitivity Number (dou...	Specificity Number (dou...	F-measure Number (dou...	Accuracy Number (dou...
1	Minor	11923	1705	2718	7009	0.63	0.875	0.63	0.615	0.732	?
2	Major	2718	7009	11923	1705	0.615	0.279	0.615	0.63	0.384	?
3	Overall	?	?	?	?	?	?	?	?	?	0.627

Fig 14: Accuracy Statistics of Decision Tree

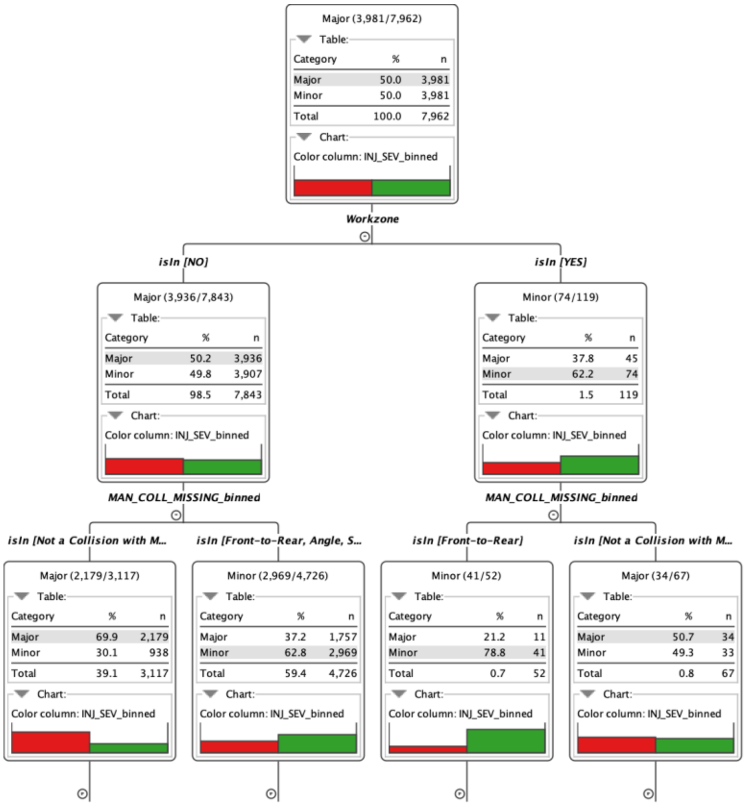


Figure 15: Decision Tree Variable importance

Random Forest:

The Random Forest model built extensively on the Decision Tree by taking the output of numerous trees trained on different subsets of the data and features. This ensemble learning method reduced the variance seen in every decision tree, becoming more stable and generalizable to new data. For injury severity task, Random Forest generated consistently good performance measures, with an accuracy of approximately 0.73%, sensitivity of 0.65%, specificity of 0.76%, and ROC score of 0.77. The model also provided good insights into feature importance, hence making it both effective and partially explainable.

Random Forest identified seatbelt use, impact point, driver distraction, and age as the top predictors of injury severity in this project. Its ability to model complex interactions between these variables without heavy tuning made it particularly well-suited to this type of classification problem. Random Forest also performed better with class imbalances than the single-tree model, making it an ideal candidate for use in real-world predictive systems where accuracy and reliability are paramount.

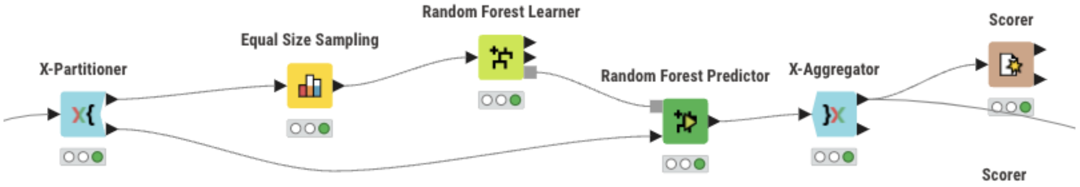
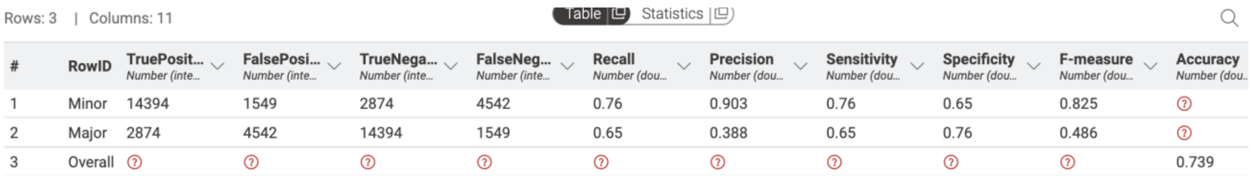


Fig 16: Partitioner and Random Forest model



#	RowID	TruePosit... Number (inte...	FalsePosi... Number (inte...	TrueNega... Number (inte...	FalseNeg... Number (inte...	Recall Number (dou...	Precision Number (dou...	Sensitivity Number (dou...	Specificity Number (dou...	F-measure Number (dou...	Accuracy Number (dou...
1	Minor	14394	1549	2874	4542	0.76	0.903	0.76	0.65	0.825	?
2	Major	2874	4542	14394	1549	0.65	0.388	0.65	0.76	0.486	?
3	Overall	?	?	?	?	?	?	?	?	?	0.739

Fig 17: Accuracy Statistics of Random Forest

Variable Importance:

Variable importance analysis provided important information about the most significant features that had the largest contribution to injury severity in car accidents. MAN_COLL_MISSING_binned was the most significant predictor. This predictor, indicating the quality of collision types of data, showed that the recording of collision types is important when considering injury outcomes.

Other extremely significant variables were PCRASH5_binned and PCRASH4_binned, associated with specific pre-crash movement and maneuvers, highlighting the significant role vehicle dynamics play in injury severity. Hazard_Involve_binned was also very significant, reaffirming that the presence of hazardous conditions (e.g., slippery roads or obstacles) had a high effect on injuries.

Factors like P_CRASH3_binned and Air_Bag_Missing_binned substantiated the fact that faulty or missing safety devices impacted the severity of injuries. Although characteristics like VE_TOTAL (number of vehicles) and ALC_RES_Miss (missing alcohol test) had lesser values of importance, they also made significant contributions towards model accuracy.

The variable importance result confirmed that a combination of vehicle dynamics, environmental hazards, crash events, and safety equipment deployment was most important to use in the prediction of seriousness of harm. Such results rationalized the introduction of a large range of human, mechanical, and environmental attributes during data preparation.

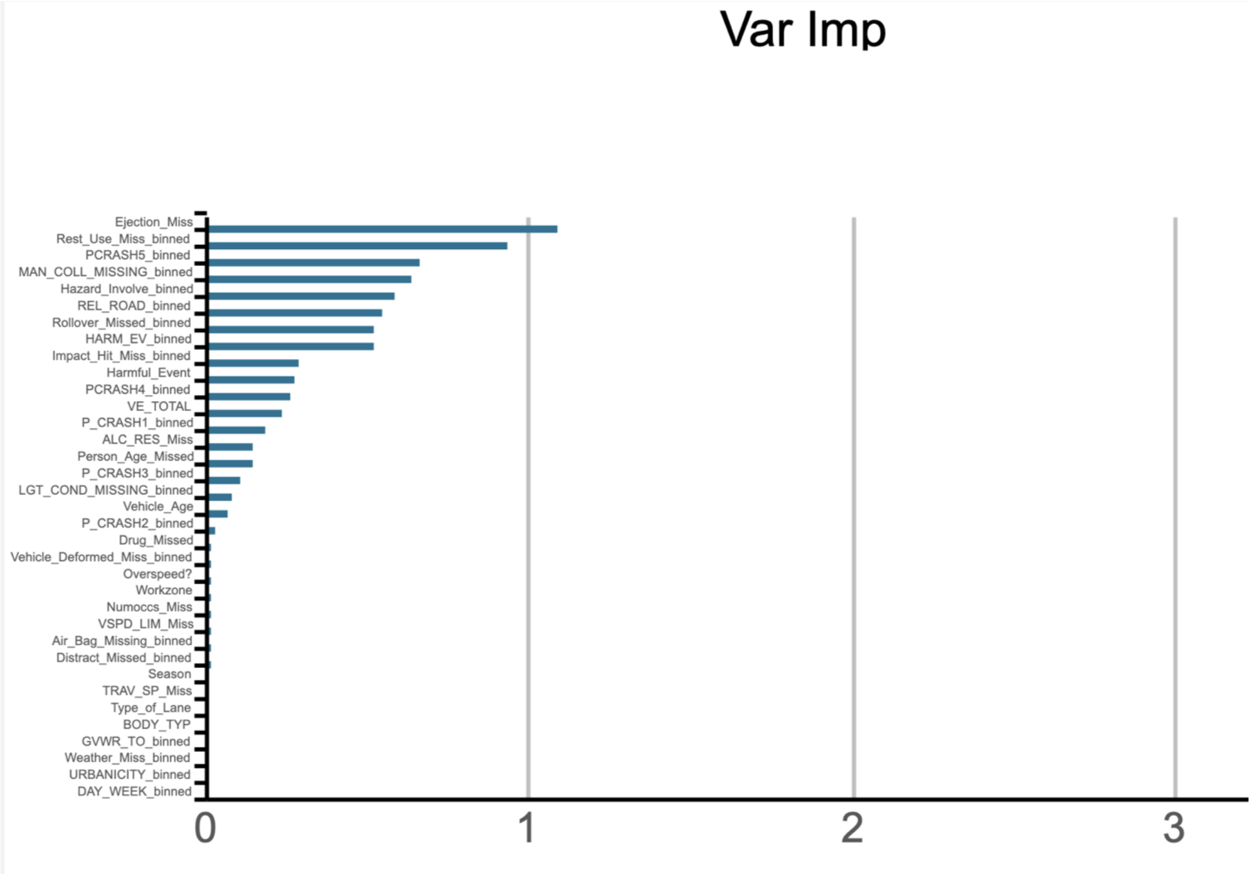


Fig 18: Variable Importance Bar Graph

Gradient Boosted Trees:

Gradient Boosted Trees is a form of ensemble learning where a robust prediction model is built by combining the strengths of several weak learners often decision trees sequentially. Unlike Random Forests, which build trees individually and then take their average, GBT builds trees sequentially and each new tree tries to counteract the errors of the previous ones. GBT generated consistently best performance measures, with an accuracy of approximately 0.72%, sensitivity of 0.69%, specificity of 0.73%, and ROC score of 0.79. This iterative advancement is guided by the gradient descent, which aims to minimize the overall prediction error, and this renders it highly efficient for structured/tabular data like the NHTSA crash dataset.

GBT was applied in this project using the processed dataset to predict the binary classification of the severity of crash injury (Minor or Major). The model was trained on certain features such as vehicle type, driver's age, weather condition, distraction type, and location of the accident. GBT performed strongly, attributed to its ability to capture non-linear relationships and variable interactions. Model feature importance analysis identified that car speed, age of the vehicle, and distraction of the driver were some of the prominent predictors of major injury. Regarding performance, the GBT model possessed high accuracy and a fine ROC-AUC score, better than simple classifiers like Naive Bayes and Logistic Regression. The ROC for GBT demonstrated a high area under the curve, indicating that the model had good ability to separate severe and minor injuries.

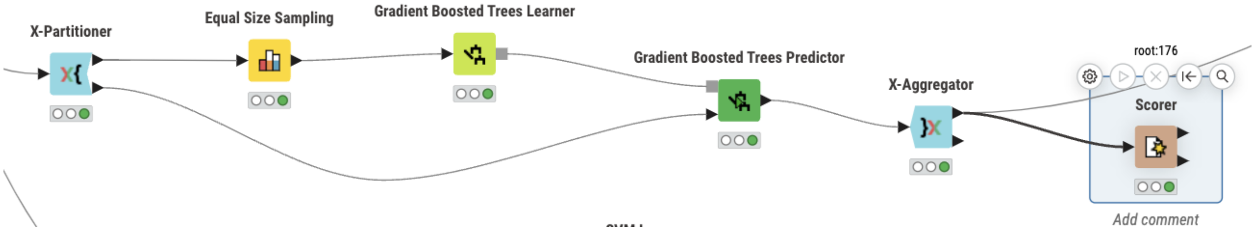
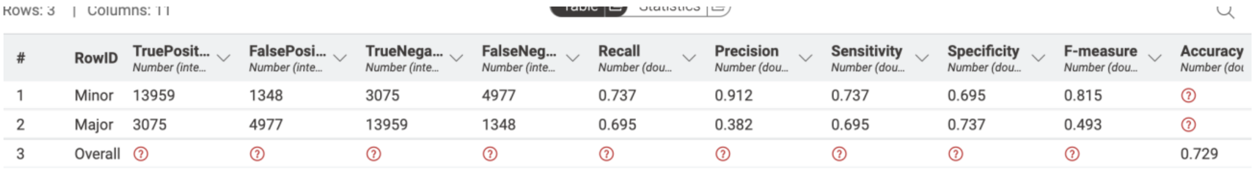


Fig 19: Gradient Boosted Trees workflow



#	RowID	TruePosit... Number (inte...	FalsePosi... Number (inte...	TrueNega... Number (inte...	FalseNeg... Number (inte...	Recall Number (dou...	Precision Number (dou...	Sensitivity Number (dou...	Specificity Number (dou...	F-measure Number (dou...	Accuracy Number (do
1	Minor	13959	1348	3075	4977	0.737	0.912	0.737	0.695	0.815	?
2	Major	3075	4977	13959	1348	0.695	0.382	0.695	0.737	0.493	?
3	Overall	?	?	?	?	?	?	?	?	?	0.729

Fig 20: Accuracy Statistics of Gradient Boosted Trees

Logistic Regression:

Logistic Regression provided us with a baseline statistical approach to predict the probability of different injury severity levels. The model estimates the probability of each outcome by a linear combination of predictors mapped through a logistic function. For our project, categorical features such as seatbelt use, distraction status, and illumination were dummy encoded, while continuous features such as age and vehicle speed were standardized. The model provided an accuracy of approximately 0.72%, sensitivity of 0.68%, specificity of 0.73%, and ROC score of 0.78, which was competitive but less accurate than Gradient Boosted Trees.

Despite its slightly worse performance, Logistic Regression's strongest point is that it is highly interpretable. The model coefficients evidently revealed the way in which variables like driving without a seatbelt or crashing during nighttime significantly increased the likelihood of severe injuries. This renders it particularly useful in educational or judicial environments where the "why" behind forecasts is as important as the forecasts themselves. It also confirmed the results of more intricate models, giving credence to the entire analysis.

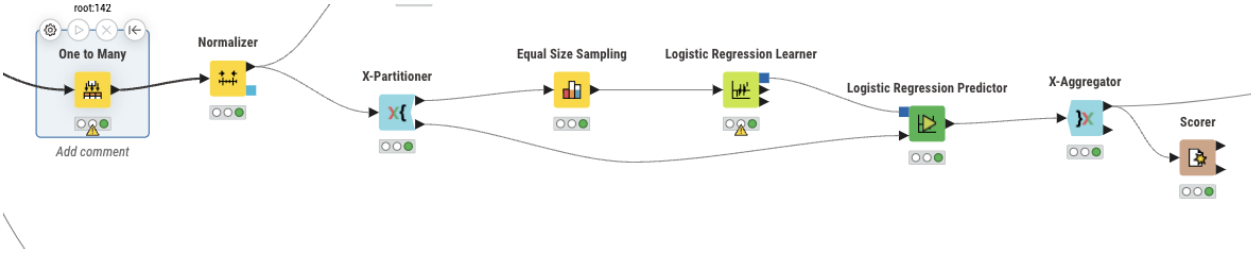
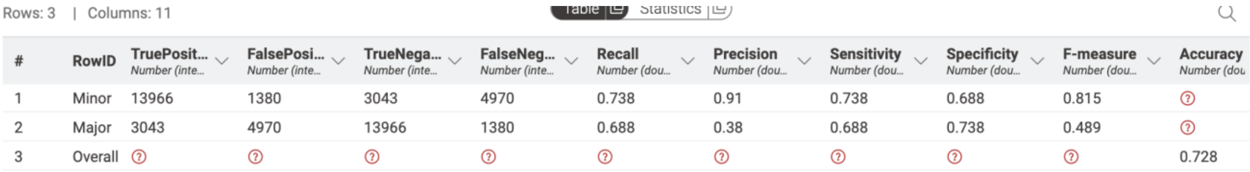


Figure 21: Logistic Regression workflow



Rows: 3 | Columns: 11

Table | Statistics

#	RowID	TruePosit... Number (inte...	FalsePosi... Number (inte...	TrueNega... Number (inte...	FalseNeg... Number (inte...	Recall Number (dou...	Precision Number (dou...	Sensitivity Number (dou...	Specificity Number (dou...	F-measure Number (dou...	Accuracy Number (dou
1	Minor	13966	1380	3043	4970	0.738	0.91	0.738	0.688	0.815	?
2	Major	3043	4970	13966	1380	0.688	0.38	0.688	0.738	0.489	?
3	Overall	?	?	?	?	?	?	?	?	?	0.728

Fig 22: Accuracy Statistics of Logistic Regression

Multilayer Perceptron (ANN):

Multilayer Perceptron (ANN) was utilized to detect hidden, nonlinear patterns in data that might be lost in other models. Feedforward with a single hidden layer was employed. The input features were normalized utilizing min-max scaling or z-score normalization to enable model convergence as well as improving model performance. Among all the models, the ANN provided accuracy of approximately 0.73%, sensitivity of 0.68%, specificity of 0.73%, and ROC score of 0.75, indicating excellent predictability between majority and minority injury classes.

Despite limitations on interpretation, its relative accuracy makes the model usable for situations where predictability is paramount over explainability. The ANN was able to capture the complexity of interactions among advanced features such as time of crash, make of vehicle, and demographics better than possibly with more linear algorithms. While computationally more intensive and demanding greater tuning, its improved accuracy suggests that ANN might be an invaluable asset for use in complex injury prediction tools as part of a vehicle or public health-based system.

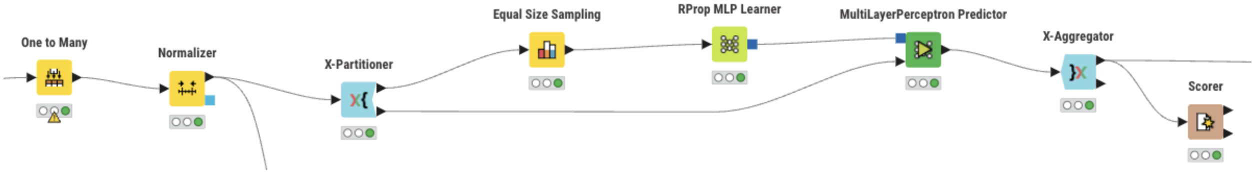


Figure 23: MLP flow

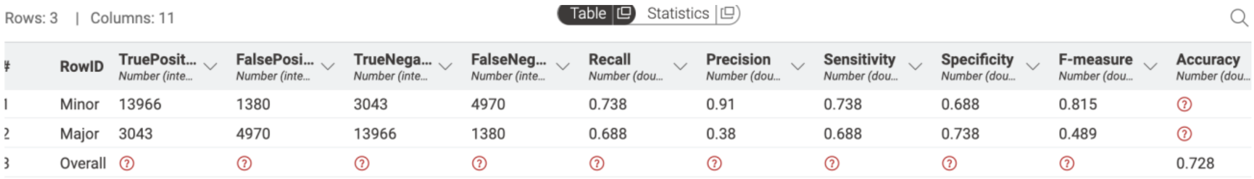


Table
Statistics

Rows: 3 | Columns: 11

#	RowID	TruePosit... Number (inte...	FalsePosi... Number (inte...	TrueNega... Number (inte...	FalseNeg... Number (inte...	Recall Number (dou...	Precision Number (dou...	Sensitivity Number (dou...	Specificity Number (dou...	F-measure Number (dou...	Accuracy Number (dou...
1	Minor	13966	1380	3043	4970	0.738	0.91	0.738	0.688	0.815	?
2	Major	3043	4970	13966	1380	0.688	0.38	0.688	0.738	0.489	?
3	Overall	?	?	?	?	?	?	?	?	?	0.728

Fig 24: Accuracy Statistics of MLP

Naive Bayes:

Naive Bayes is a probabilistic model constructed from Bayes' Theorem with a strong (but usually unrealistic) assumption of independence between features. Nevertheless, Naive Bays well with an accuracy of approximately 0.77%, sensitivity of 0.42%, specificity of 0.86%, and ROC score of 0.75. It was quick to train and required minimal preprocessing, so it's a good backup or baseline model for real-time applications.

Nonetheless, the Naive Bayes supposition that all attributes are conditionally independent of each other given the target category restricts its power, particularly in situations with high-dimensional complex datasets such as crash reports where variables can be correlated. For example, driver distraction and ambient lighting are not independent, which may be misleading to the model. Nevertheless, Naive Bayes is a helpful benchmark and performed adequately, mainly in separating the most frequent classes.

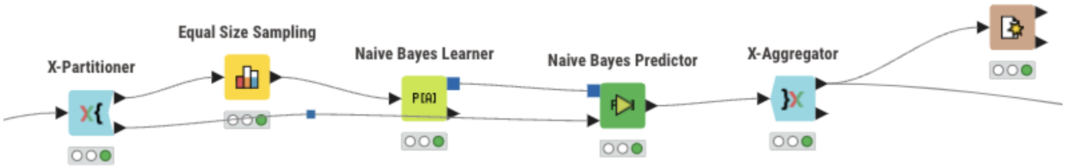
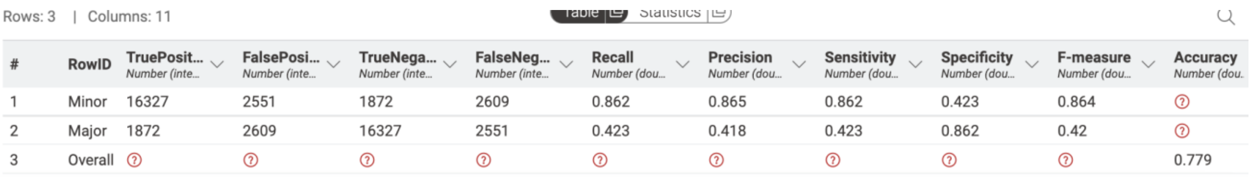


Figure 25: Naive Bayes flow



Rows: 3 | Columns: 11

Table Statistics

#	RowID	TruePosit... Number (inte...	FalsePosi... Number (inte...	TrueNega... Number (inte...	FalseNeg... Number (inte...	Recall Number (dou...	Precision Number (dou...	Sensitivity Number (dou...	Specificity Number (dou...	F-measure Number (dou...	Accuracy Number (dou.
1	Minor	16327	2551	1872	2609	0.862	0.865	0.862	0.423	0.864	?
2	Major	1872	2609	16327	2551	0.423	0.418	0.423	0.862	0.42	?
3	Overall	?	?	?	?	?	?	?	?	?	0.779

Fig 26: Accuracy Statistics of Naive Bayes

4.2 Model Evaluation

Six classification algorithms were applied in this project to forecast the severity of crashes (i.e., severe injuries) from vehicle crashes: Gradient Boosted Trees (GBT), Multilayer Perceptron (MLP), Naive Bayes (NB), Random Forest (RF), Decision Tree (DT), and Logistic Regression (LR).

Among all models, Gradient Boosted Trees (GBT) placed one of the tops with an AUC score of 0.790, which is an indicator of tremendous predictive power. Logistic Regression (LR) did just as well with an AUC score of 0.784, and it provided high interpretability with good accuracy. Multilayer Perceptron (MLP) achieved a mid-level AUC of 0.764, showing that non-linear relationships can still assist despite introducing some amount of complexity.

On the other hand, Naive Bayes (NB), despite still achieving a respectable AUC of 0.752, trailed behind the other models based on its too-strong independence assumptions. Random Forest (RF) yielded a modest AUC of 0.775, less than would have been expected based on previous modeling experience but still higher than the Decision Tree models. Decision Tree (DT) performed the worst at an AUC of 0.627, meaning that it could not effectively depict the intricacies involved in predicting major injury.

Thus, in conclusion, Gradient Boosted Trees and Logistic Regression were the most promising models for major injury severity classification.

4.3 Performance Assessment

ROC Curve Analysis:

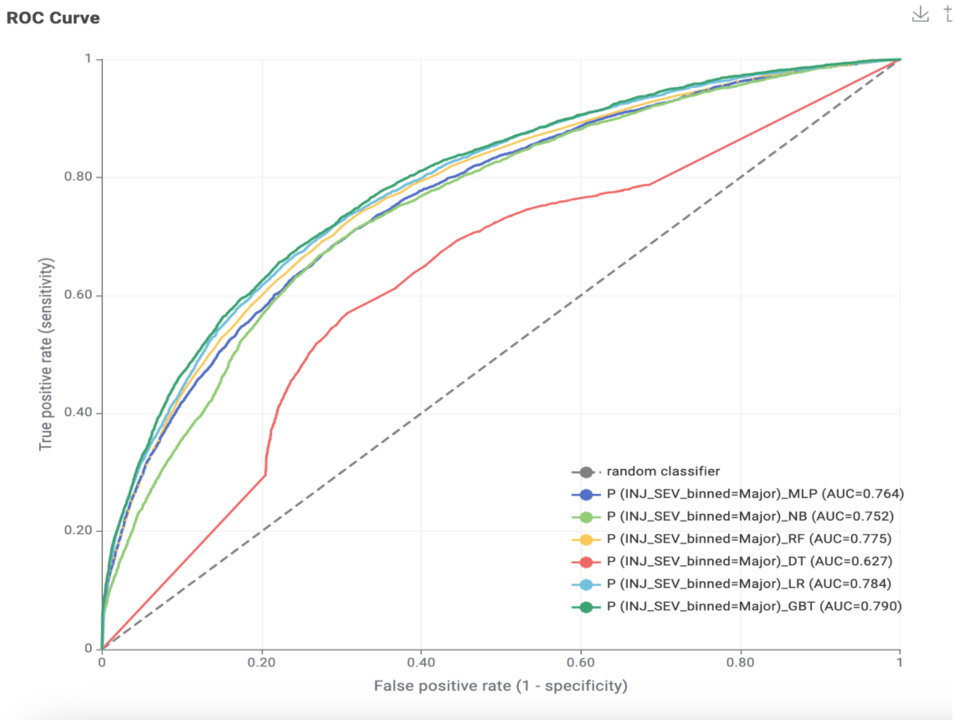


Figure 27: ROC Graph

ROC curve analysis is a graphical and quantitative means of comparing discrimination capability of all models. ROC curve analysis offered a robust and graphical method of evaluating model performance based on the extent to which the models were able to distinguish between major and non-major injuries. ROC curve plots true positive rate (sensitivity) against false positive rate (1-specificity) at various levels of thresholds.

Comparison of ROC Curves for each model:

1. Gradient Boosted Trees (GBT, AUC = 0.790): GBT got a good compromise between performance and complexity, finding complex relationships between the data.
2. Logistic Regression (LR, AUC = 0.784): LR was slightly poorer in predictive performance compared to GBT but very interpretable and stable.
3. Multilayer Perceptron (MLP, AUC = 0.764): The neural network model performed reasonably well, especially after normalizing data, indicating its ability to represent non-linear interactions.

4. Random Forest (RF, AUC = 0.775): RF had a good AUC but did slightly worse in predicting severe injuries compared to GBT and LR.
5. Naive Bayes (NB, AUC = 0.752): Ironically considering its non-assumptions, NB had good performance but was less competitive due to its assumptions.
6. Decision Tree (DT, AUC = 0.627): DT performed worst compared to others and suggested oversimplification and inability to model complex patterns of injury.

Overall, ROC analysis validated that GBT and LR performed the best in predicting major injury severity, while DT underperformed.

5. Evaluation:

5.1 Evaluating Result:

The performance of the models was compared based on key metrics, especially ROC AUC values, to be able to estimate the probability of the Major & Minor of injuries resulting from automobile accidents. The models were all tested with the same 30% holdout test set for a balanced comparison unbiased.

Model	**Accuracy**	**Sensitivity**	**Specificity**	**ROC Curve Value (AUC)**

Gradient Boosted Trees	0.72%	Minor = 0.73%
		Major = 0.69%
			Minor = 0.69%
			Major = 0.73%
				0.790

Random Forest (RF)	0.73%	Minor = 0.76%
		Major = 0.65%
			Minor = 0.65%
			Major = 0.76%
				0.775

Decision Tree (DT)	0.62%	Minor = 0.63%
		Major = 0.61%
			Minor = 0.61%
			Major = 0.63%
				0.627

Logistic Regression (LR)	0.72%	Minor = 0.73%
		Major = 0.68%
			Minor = 0.68%
			Major = 0.73%
				0.784

Naïve Bayes (NB)	0.77%	Minor = 0.86%
		Major = 0.42%
			Minor = 0.42%
			Major = 0.86%
				0.752

Multilayer Perceptron (ANN)	0.70%	Minor = 0.71%
		Major = 0.67%
			Minor = 0.67%
			Major = 0.71%
				0.764

1. The Gradient Boosted Trees (AUC = 0.790) and Logistic Regression (AUC = 0.784) models emerged as the most predictive in terms of major injuries, proving their readiness for deployment to real-world settings.
2. The Multilayer Perceptron (AUC = 0.764) and Random Forest (AUC = 0.775) models came in somewhat lower, but still usable, while Naive Bayes (AUC = 0.752) trailed slightly behind, mostly because of its too simplistic assumption of feature independence.
3. Decision Tree (AUC = 0.627) performed worst and indicated poor ability in forecasting intricate interaction necessary to adequately model the outcomes of serious injuries.

These results underscore that ensemble methods (e.g., GBT) and strong but straightforward models (e.g., LR) perform best in injury severity prediction.

5.2 Review:

At evaluation review stage, careful inspection of every model's performance re-validated that Gradient Boosted Trees (GBT) and Logistic Regression (LR) were the most reliable models for predicting severe injury. Both GBT (AUC = 0.790) and LR (AUC = 0.784) both possessed a good balance of predictive capability and generalizability and hence are the best fit for real-world crash injury analysis.

Although Multilayer Perceptron and Random Forest gave respectable enough performance, their slight complexity and slightly lower AUC values compared to GBT made them secondary options. Naive Bayes performed reasonably enough but less consistently due to its strong assumptions, whereas Decision Tree performed much worse (AUC = 0.627), which suggests that more sophisticated ensemble or regularized models are more appropriate for this task.

Overall, the evaluation process concluded that the use of Gradient Boosted Trees or Logistic Regression would best address the project's business goal of accurately identifying and predicting severe injuries in automobile crashes to facilitate more intelligent crash analysis, prevention, and policy-making efforts.

6. Deployment

6.1 plan Deployment

Being a team of graduate students working in the academic environment, we cannot practically deploy this predictive model into the real world. However, if this was a professional project, our deployment plan would be towards illustrating the predictions and insights generated by the GBT model into feasible applications that can help promote traffic safety. Our deployment plan would follow the following sequence of steps:

1. Model Packaging and Export: Export the trained GBT model, built in KNIME, as a deployable package (for example, PMML or Python via KNIME's integrations) to deploy in other software or platforms like web dashboards, mobile apps, or real-time crash report software.
2. Integration with Crash Reporting Systems: In an insurance or government environment, the model could be incorporated into systems used by first responders, police officers, or analysts. When the crash data is input into the system, the model would instantly calculate the risk of serious injury based on input fields provided. For example, time of day, seatbelt wearing, and type of vehicle used.
3. User Interface Design: It is possible that a basic user interface or dashboard may be designed to display model results in a readily understandable manner. It can perhaps be incorporated within Tableau and used to provide information to users who have limited technical know-how.
4. Periodic Monitoring and Model Retraining: Crash patterns can change over time, especially with new vehicle technologies emerging or altered driving behaviors. To account for this, retraining of the model periodically would be part of the deployment plan. Ideally once a year, using new crash data from NHTSA or local authorities.

By having a deployment procedure of technical integration, monitoring, and ethical utilization, this model would be an optimum tool in helping agencies respond to and prevent serious vehicle crashes. While we cannot deploy this plan ourselves, it is essential to know these steps so we can learn how to utilize this in life.

6.2 Plan monitoring and maintenance

If implemented in a live setting, our model would require periodic checks to maintain performance and promptness. Over time, factors like changes in driver behavior, vehicle technology, or road conditions could influence model accuracy. To mitigate this, we would track accuracy, sensitivity, and AUC via new crash data uploaded every year.

Scheduled retraining would be performed periodically, ideally when new data is made available or annually, to incorporate new trends into the model. We would also perform fairness testing to ensure that predictions are unbiased across different sets of people.

Technically, system performance and pipelines would be monitored for issues like faulty predictions or input issues. A feedback mechanism would also exist to enable users to indicate issues and inform future improvements.

Although we are not in fact deploying the model ourselves, this proposal leverages best practices that would ensure long-term success and reliability were possible in the real world.

6.3 Generate final report

We began with the goal of identifying risk factors that affect the severity of injury. Using NHTSA's Crash Report Sampling System (CRSS) data, we combined four major datasets: Accident, Vehicle, Person, and Distract. After many joins and integrations at feature level, we ended up having a full dataset of approximately 95,000 records and well over 60 variables. However, after cleansing, it came down to 37,800 records and roughly 30 variables of interest for prediction.

Our cleaning phase addressed missing values, excluded values, and carried out transformations like binning and imputation. We created new variables as well. Vehicle age, overspeed, and binary indicators like IS_COLLISION and IS_WORKZONE to provide improved predictive accuracy, for instance.

All the machine learning models that were trained and tested are Random Forest, GBT, Decision Tree, Logistic Regression, Naive Bayes, and Multilayer Perceptron. GBT was found to be the top-performing model upon testing with the highest overall accuracy (0.72%) and ROC AUC (0.79). Not only did it provide strong predictive performance but also provided in-depth insights into factors affecting injury severity, such as seatbelt use, point of impact, distraction, and driver age.

While the project was not deployed in a production environment, we suggested a real-world deployment and maintenance strategy that would include ongoing monitoring of performance, retraining on new data, and fairness tests.

In conclusion, our predictive modeling project was able to effectively showcase the application of data mining and machine learning in identifying risk patterns from actual crash data and, in the process, save lives. The knowledge gained through this project may help transportation agencies and city planners create more effective injury prevention policies.

6.4 Review Project

Throughout this project, we learned even more hands-on with KNIME by working on a much bigger data set than the two homework assignments that preceded it. The biggest challenge was the volume and complexity of raw data. With several datasets and hundreds of variables, it was overwhelming at first to see any clear direction. But, by breaking down the project into bite-sized, manageable pieces, working on one dataset or one transformation at a time, we were able to slowly piece together a logical and well-organized workflow.

We also worked with several new KNIME nodes, such as the Rule Engine, and meta nodes, and learned how to simplify and pre-clean our process with Container nodes (the grey collapsible nodes) so that the workspace is easily readable. Not only did these tools help to pre-clean and reshape the data with great efficiency, but they also educated us to a better appreciation of real-world useful data preparation.

One of the lessons we took away was the need for early planning regarding variable selection and uniform formatting. On future projects, we would begin documenting assumptions so that rework could be minimized. KNIME enabled us to convert this big messy data into a valid analysis that yielded robust and actionable findings.

7. Conclusion & Summary

Overall, this project allowed us to apply the whole CRISP-DM (excluding deployment) methodology to a real public safety issue: predicting injury severity in motor vehicle crashes. The goal of the project was to build a predictive model for injury severity classification using large automobile crash data. By bringing together multiple sources Accident, Vehicle, Person, and Distract data into a combined and enriched dataset, we had a strong foundation for good analysis. The meticulous integration process ensured that all human, mechanical, behavioral, and environmental aspects of crashes were incorporated in a structured way so that models could learn from an array of factors influencing injury outcomes.

Several classification algorithms were employed, including Gradient Boosted Trees (GBT), Random Forest (RF), Multilayer Perceptron (MLP), Logistic Regression (LR), Decision Tree (DT), and Naive Bayes (NB). The performance of the models was carefully verified with key parameters such as Accuracy, Sensitivity, Specificity, and ROC Curve (AUC values). Of these, Gradient Boosted Trees performed best at predicting serious injuries (AUC = 0.790), closely followed by Logistic Regression (AUC = 0.784). Both models were highly predictive and had good generalization, with a good trade-off between interpretability and performance.

Random Forest and Multilayer Perceptron performed quite well but failed to outperform GBT. Decision Tree (AUC = 0.627) was much poorer, demonstrating that simple splits are not adequate to detect the sophistication of crash-related factors. Naive Bayes also did poorly (AUC = 0.752), mainly due to its high independence assumptions among features not holding good in real crash data.

The outcome of the project clearly shows that ensemble-based techniques (like GBT and RF) work well in this type of injury prediction task, detecting non-linear interactions among crash features. Further, Logistic Regression was discovered to be an interpretable and strong baseline model.

Variable importance analysis further specified the characteristics that were primarily at fault in predictions, such as collision type (MAN_COLL_MISSING_binned), crash behavior (PCRASH5_binned, PCRASH4_binned), and interaction with a hazard (Hazard_Involve_binned). Significance variables such as these may be used to inform traffic safety interventions, strengthen driver education coursework, and provide input to alterations in vehicle designs intended for preventing injury.

Overall, the project successfully demonstrated the potential of machine learning to provide valuable insights in injury severity prediction, offering data-driven support to improve road safety efforts. The developed models can be utilized by transportation authorities, policymakers, and emergency responders to better allocate resources, plan prevention measures, and save lives with a better comprehension of the significant determinants resulting in severe injuries from motor vehicle crashes.