Injury Severity Prediction:

Decision Tree, Random Forest, Neural Network, and Logistic Regression

By

Ayman Amer Abdulhafed Alkubati

In this project, I will apply the industry-standard CRISP-DM methodology for data mining to unlock insights from the extensive vehicle crash dataset provided by NHTSA. My analysis will develop predictive models using four techniques - Decision Trees, Random Forests, Neural Networks and Logistic Regression. The KNIME analytics platform will enable implementing this modeling workflow through its user-friendly graphical interface.

My analysis will leverage this rich data to uncover patterns related to injury severity arising from crashes. I plan to apply predictive modeling techniques to estimate the likelihood of severe injuries resulting from various crash types and configurations. This could spotlight key risk factors amenable for safety interventions, from drunk driving to hazardous road segments. The insights can assist NHTSA and other agencies in quantifying the harm potential of different crash domains and allocating countermeasures appropriately for maximal life and cost savings.

Business Understanding

One of the primary objectives of the National Highway Traffic Safety Administration (NHTSA) is to mitigate the impact of motor vehicle crashes, which result in thousands of fatalities and injuries annually alongside substantial economic costs. Accurate crash data is imperative for developing and assessing highway safety initiatives aimed at reducing this burden. The CRSS dataset I will utilize represents a national probability sample of over 6 million police-reported crashes spanning minor fender-benders to those with devastating outcomes.

The core analytical focus involves modeling personal injury severity arising from the diverse crash conditions represented in the dataset. By applying techniques like neural networks, we can estimate the likelihood of severe injuries based on attributes of the crash, vehicles and drivers. Statistics on crash factors leading to certain injury types can also guide education or enforcement initiatives targeting accident prevention.

Data Understanding

The extensive dataset captures over 20 distinct dimensions spanning driver behavior like speeding, vehicle traits including age and type, road conditions during the crash, collision dynamics and resulting harm. Both numeric metrics like vehicle deformation as well as categorical inputs like crash manner and location are covered.

Careful inspection is needed to quantify missing values and anomalies that could undermine modeling. Assessing feature distributions can also inform appropriate preprocessing and transformation to enable effective application of the planned analytical techniques. The breadth of real-world data will facilitate building generalizable and policy-relevant injury severity models.

Data Preparation:

For this term project focused on analyzing and modeling injury severity in motor vehicle crashes, comprehensive data processing represented an indispensable first step. I was provided extensive data spanning over 20 attributes of crash incidents, with variables like driver alcohol impairment, speeding involvement, weather conditions, vehicle type, and many more factors that could potentially correlate with crash severity outcomes.

However, the majority of these variables were encoded in a raw format that is not suited for direct analysis. My data wrangling aimed to transform the data into a more manageable format for statistical analysis and predictive modeling.

For example, the "Weather_binned" field originally contained over 15 distinct weather states ranging from clear conditions to rain, snow and fog among others. I judiciously aggregated these into "favorable", "unfavorable" and "unknown" buckets to facilitate assessing hypotheses around weather playing a role in crash injury outcomes. A list of some of the other variables categorized:

- Day_week_binned: Categorized day of week into weekdays and weekends. This allows assessing if crash patterns differ on weekdays versus weekends.
- Man_coll_binned: Grouped manner of collision into front, side, rear, and other impacts. This can help understand if certain crash orientations lead to more or less severe injuries.
- Speedrel_binned: Binned speeding related variable into yes, no, and unknown categories. This will allow analysis of whether speeding increases injury severity.
- Month_binned: Separated months into winter, summer, and other seasons. This enables examining if time of the year impacts crash outcomes.
- Region: Converted region to a numeric code to describe broader regions like Northeast, Midwest, etc.
- Sex_binned: Categorized biological sex into male, female, and other. Permits evaluating if injury patterns differ by sex.
- Hour_binned: Divided time of day into daytime and nighttime. Checks if late night crashes lead to more severe injuries for instance due to fatigue.

- Drinking_binned: Classified police-reported alcohol involvement into yes, no, and unknown. Essential for analyzing if alcohol use increases injury severity.

And over 10 other variables similarly binned. This extensive preprocessing of categorical data will enable more impactful statistical and predictive modeling during my term project. By judiciously binning variables, I have retained enough detail for insightful analysis while simplifying unwieldy raw data. Descriptive definitions for each variable have also been recorded for reference. In summary, thoughtful data wrangling represents an indispensable first step towards a successful data science application.

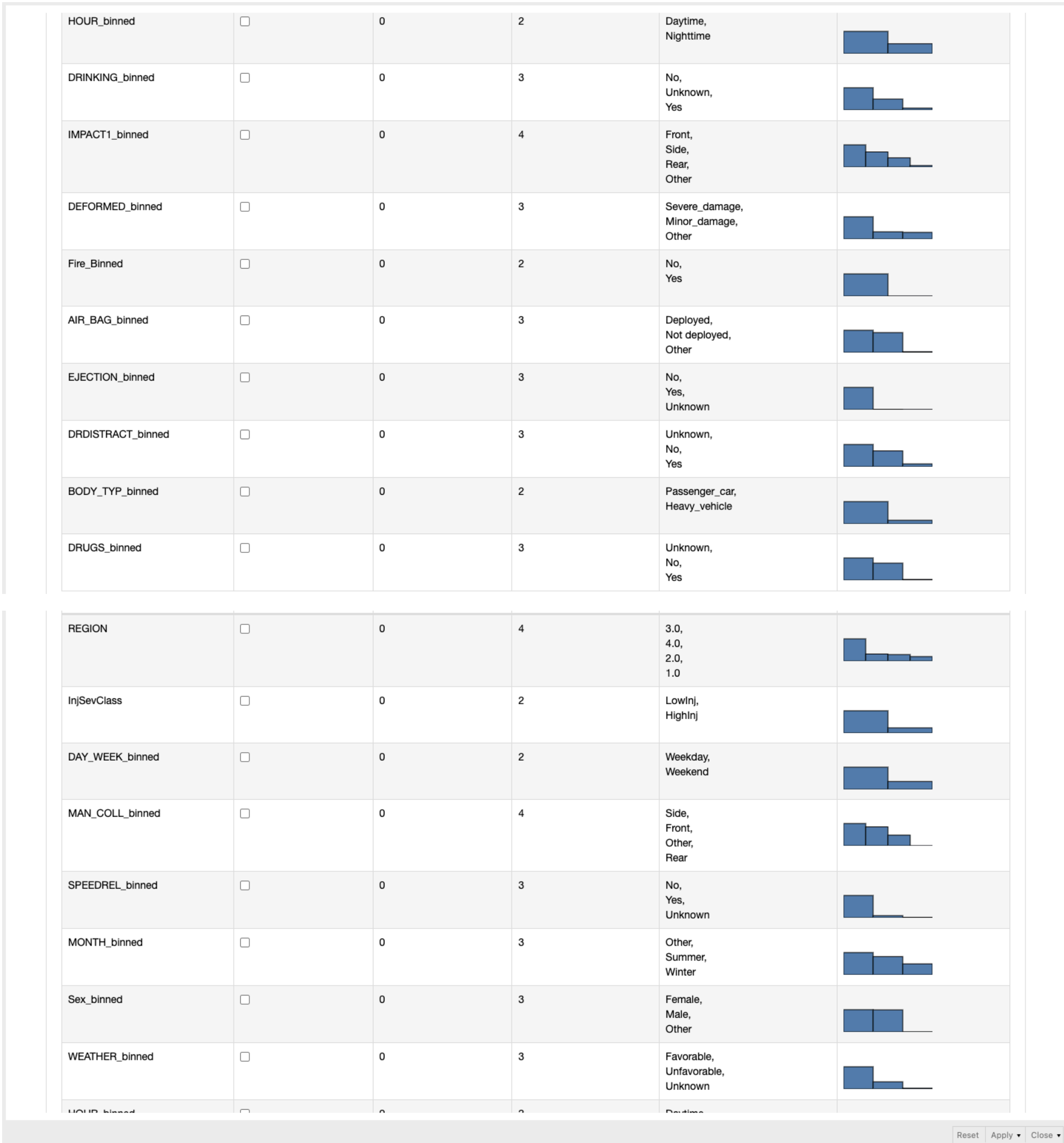
Numeric Nominal Data Preview

Search:

Column	Exclude Column	Minimum	Maximum	Mean	Standard Deviation	Variance	Skewness	Kurtosis	Overall Sum	No. zeros	No. missings	No. NaN	No. +∞
Age_Miss	☐	12	99	40.823	17.520	306.954	0.609	-0.515	953579	0	0	0	0
Age_VEH	☐	0	92	9.207	6.955	48.377	1.174	2.625	215059	98	0	0	0
OverSpeeding	☐	-75	85	-14.980	18.479	341.457	-0.228	0.649	-349916	3516	0	0	0

Showing 1 to 3 of 3 entries

Figure 1: Numeric data on Data explorer.



HOUR_binned	☐	0	2	Daytime, Nighttime
DRINKING_binned	☐	0	3	No, Unknown, Yes
IMPACT1_binned	☐	0	4	Front, Side, Rear, Other
DEFORMED_binned	☐	0	3	Severe_damage, Minor_damage, Other
Fire_Binned	☐	0	2	No, Yes
AIR_BAG_binned	☐	0	3	Deployed, Not deployed, Other
EJECTION_binned	☐	0	3	No, Yes, Unknown
DRDISTRACT_binned	☐	0	3	Unknown, No, Yes
BODY_TYP_binned	☐	0	2	Passenger_car, Heavy_vehicle
DRUGS_binned	☐	0	3	Unknown, No, Yes
REGION	☐	0	4	3.0, 4.0, 2.0, 1.0
InjSevClass	☐	0	2	LowInj, HighInj
DAY_WEEK_binned	☐	0	2	Weekday, Weekend
MAN_COLL_binned	☐	0	4	Side, Front, Other, Rear
SPEEDREL_binned	☐	0	3	No, Yes, Unknown
MONTH_binned	☐	0	3	Other, Summer, Winter
Sex_binned	☐	0	3	Female, Male, Other
WEATHER_binned	☐	0	3	Favorable, Unfavorable, Unknown

Figure 2: Nominal data on the Data explorer node.

Numeric | Nominal | Data Preview

Search:

Row ID	REGION	InjSevClass	Age_Miss	Age_VEH	DAY_WEEK_binned	MAN_COLL_binned	SPEEDREL_binned
Row4_Row6_Row6_Row6	1.0	LowInj	59	5	Weekend	Other	Yes
Row6_Row8_Row8_Row8	2.0	HighInj	18	14	Weekend	Other	No
Row7_Row9_Row9_Row9	3.0	HighInj	21	3	Weekday	Other	Yes
Row9_Row12_Row12_Row12	3.0	LowInj	57	16	Weekend	Other	No
Row12_Row17_Row17_Row17	3.0	HighInj	72	4	Weekend	Other	No
Row14_Row21_Row21_Row21	3.0	LowInj	71	24	Weekday	Front	No
Row17_Row26_Row26_Row26	3.0	HighInj	19	16	Weekend	Other	Unknown
Row18_Row27_Row27_Row27	3.0	HighInj	56	19	Weekend	Front	No
Row19_Row29_Row29_Row29	3.0	LowInj	22	25	Weekday	Other	Yes
Row20_Row31_Row32_Row31	3.0	LowInj	18	18	Weekday	Front	No
Row22_Row33_Row36_Row33	3.0	HighInj	23	9	Weekday	Other	No
Row30_Row46_Row59_Row46	4.0	LowInj	30	5	Weekday	Other	No
Row32_Row50_Row63_Row50	4.0	LowInj	46	28	Weekday	Front	No
Row32_Row51_Row64_Row51	4.0	LowInj	78	8	Weekday	Front	No
Row36_Row56_Row73_Row56	3.0	LowInj	21	25	Weekend	Front	No
Row36_Row57_Row74_Row57	3.0	LowInj	39	12	Weekend	Front	No
Row39_Row61_Row81_Row61	3.0	LowInj	38	2	Weekend	Side	No
Row40_Row62_Row82_Row62	3.0	LowInj	30	2	Weekend	Front	No

Reset | Apply | Close

Figure 3: Data Preview on the Data explorer node.

Table with Color information - 3:28 - Color Manager

File Edit Hilite Navigation View

Table "default" - Rows: 23359 | Spec - Columns: 21 | Properties | Flow Variables

Row ID	S REGION	S InjSev...	D Age_...	D Age_V...	S DAY_...	S MAN_...	S SPEED...	D OverS...	S MONT...	S Sex_bi...	S WEAT...
Row4_Row...	1.0	LowInj	59	5	Weekend	Other	Yes	-15	Winter	Male	Unfavorable
Row6_Row...	2.0	HighInj	18	14	Weekend	Other	No	5	Winter	Male	Favorable
Row7_Row...	3.0	HighInj	21	3	Weekday	Other	Yes	30	Winter	Male	Unfavorable
Row9_Row...	3.0	LowInj	57	16	Weekend	Other	No	0	Winter	Male	Favorable
Row12_Ro...	3.0	HighInj	72	4	Weekend	Other	No	-15	Winter	Male	Favorable
Row14_Ro...	3.0	LowInj	71	24	Weekday	Front	No	-45	Winter	Female	Favorable
Row17_Ro...	3.0	HighInj	19	16	Weekend	Other	Unknown	-15	Winter	Male	Favorable
Row18_Ro...	3.0	HighInj	56	19	Weekend	Front	No	-25	Winter	Male	Unfavorable
Row19_Ro...	3.0	LowInj	22	25	Weekday	Other	Yes	-5	Winter	Male	Unknown
Row20_Ro...	3.0	LowInj	18	18	Weekday	Front	No	-5	Winter	Female	Favorable
Row22_Ro...	3.0	HighInj	23	9	Weekday	Other	No	-10	Winter	Male	Unfavorable
Row30_Ro...	4.0	LowInj	30	5	Weekday	Other	No	15	Winter	Male	Favorable
Row32_Ro...	4.0	LowInj	46	28	Weekday	Front	No	-10	Winter	Female	Unfavorable
Row32_Ro...	4.0	LowInj	78	8	Weekday	Front	No	-35	Winter	Male	Unfavorable
Row36_Ro...	3.0	LowInj	21	25	Weekend	Front	No	10	Winter	Male	Favorable
Row36_Ro...	3.0	LowInj	39	12	Weekend	Front	No	0	Winter	Female	Favorable
Row39_Ro...	3.0	LowInj	38	2	Weekend	Side	No	-15	Winter	Female	Favorable
Row40_Ro...	3.0	LowInj	30	2	Weekend	Front	No	0	Winter	Male	Unfavorable
Row40_Ro...	3.0	LowInj	35	9	Weekend	Front	No	-10	Winter	Female	Unfavorable
Row42_Ro...	3.0	LowInj	28	32	Weekend	Side	No	-58	Winter	Male	Favorable
Row48_Ro...	3.0	HighInj	37	10	Weekend	Front	No	0	Winter	Male	Favorable
Row48_Ro...	3.0	HighInj	46	4	Weekend	Front	No	0	Winter	Female	Favorable
Row53_Ro...	2.0	LowInj	38	4	Weekday	Side	No	5	Winter	Female	Unfavorable
Row55_Ro...	1.0	LowInj	65	17	Weekday	Other	Yes	-25	Winter	Male	Unfavorable
Row62_Ro...	3.0	LowInj	26	5	Weekday	Front	No	-35	Winter	Male	Favorable
Row64_Ro...	2.0	HighInj	58	16	Weekday	Other	Yes	-25	Winter	Male	Unfavorable
Row65_Ro...	2.0	LowInj	40	27	Weekday	Other	No	-25	Winter	Male	Favorable

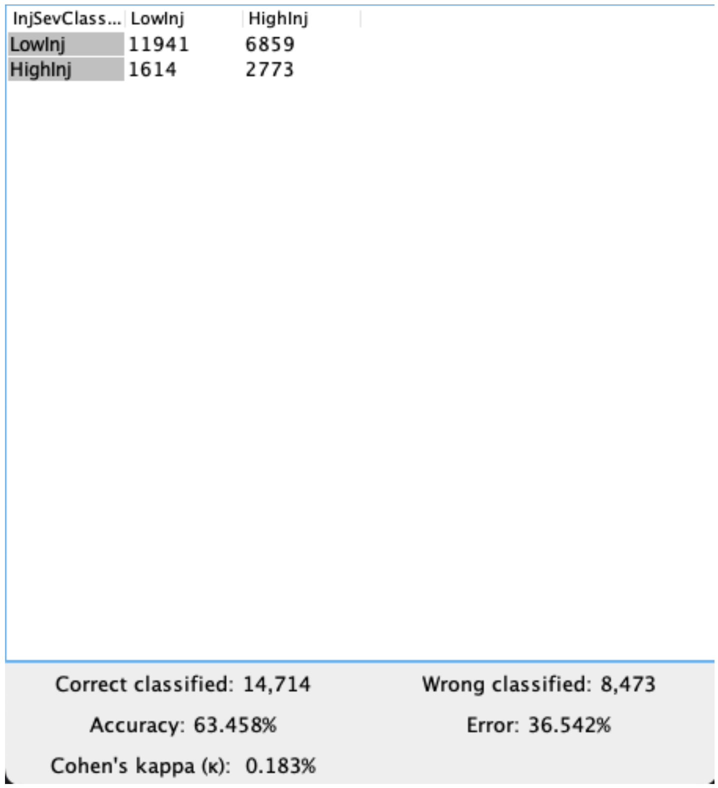
Figure 4: Colored Rows based on InjurySevClass values.

Model Building and Testing:

A- Decision Tree Model:

After understanding the business problem, exploring and transforming the data, according to the CRISP-DM methodology, we proceed to the fourth step, which involves building our models. To accomplish this, we employed a 10-fold partitioning technique with equal size sampling. The data was divided into ten groups, with each group containing an equal proportion of the dataset. This approach ensures that the training and testing sets are representative of the overall data distribution. Additionally, the partitioning was stratified based on the InjSevClass column to maintain the proportional representation of each class in both the training and testing sets. To ensure consistent results, we used a random seed value of 12345 during the partitioning process

In the next step, we built a Decision Tree model, having the target variable being the column named InjSevClass. To evaluate the model and check its accuracy, Knime offers various tools, in this analysis we can use a tool called Scorer and also plotting it on the ROC curve. For the low specificity where the minority class of values get predicted with a lower success rate of around 60%, we decided to choose equal size sampling to help balancing the training data, and we can see the new accuracy rate in Figure 5. While Figure 6 shows the plotted ROC curve derived from this decision tree model. Lastly, the Decision Tree graphical model of the first two levels is shown in Figure 7.



InjSevClass...	LowInj	HighInj
LowInj	11941	6859
HighInj	1614	2773

Correct classified: 14,714 Wrong classified: 8,473

Accuracy: 63.458% Error: 36.542%

Cohen's kappa (κ): 0.183%

Figure 5: Decision Tree Scorer.

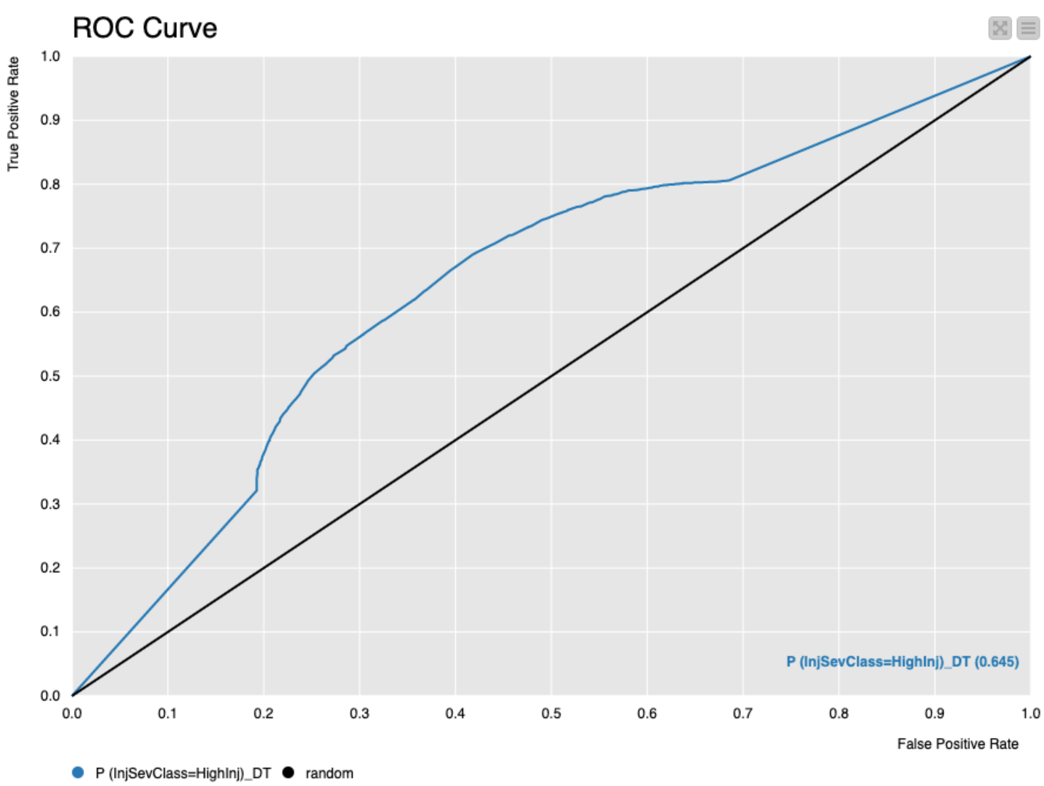


Figure 6: Decision Tree ROC Curve.

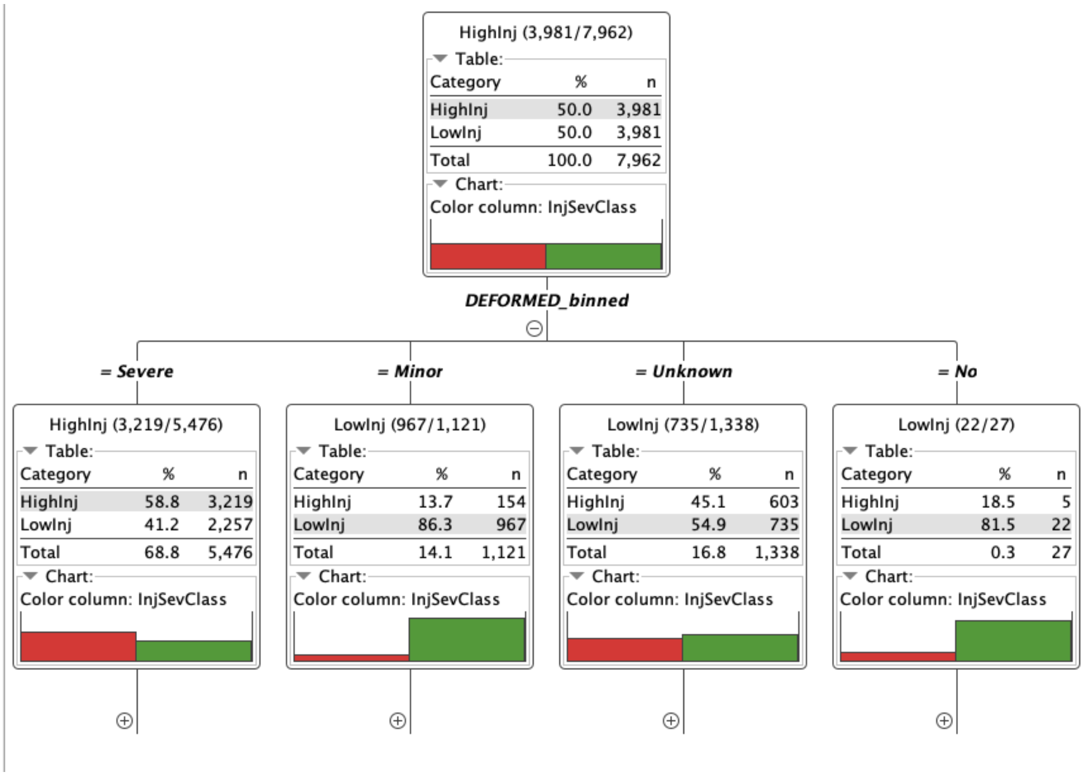


Figure 7: Two levels Decision Tree.

B- Random Forest Model:

Similar to how we created the Decision Tree model, we can use the exact same settings we used to create the Random Forest model to be able to compare the accuracy that each model produce. And as the fifth step of the CRISP-DM methodology focuses on testing and evaluating the output of the newly built models, we will use the two tools we used previously, the Scorer and ROC Curve, displayed in Figure 8 and 9 respectively.

InjSevClass...	LowInj	HighInj
LowInj	11941	6859
HighInj	1614	2773

Correct classified: 14,714 Wrong classified: 8,473

Accuracy: 63.458% Error: 36.542%

Cohen's kappa (κ): 0.183%

Figure 8: Random Forest Scorer.

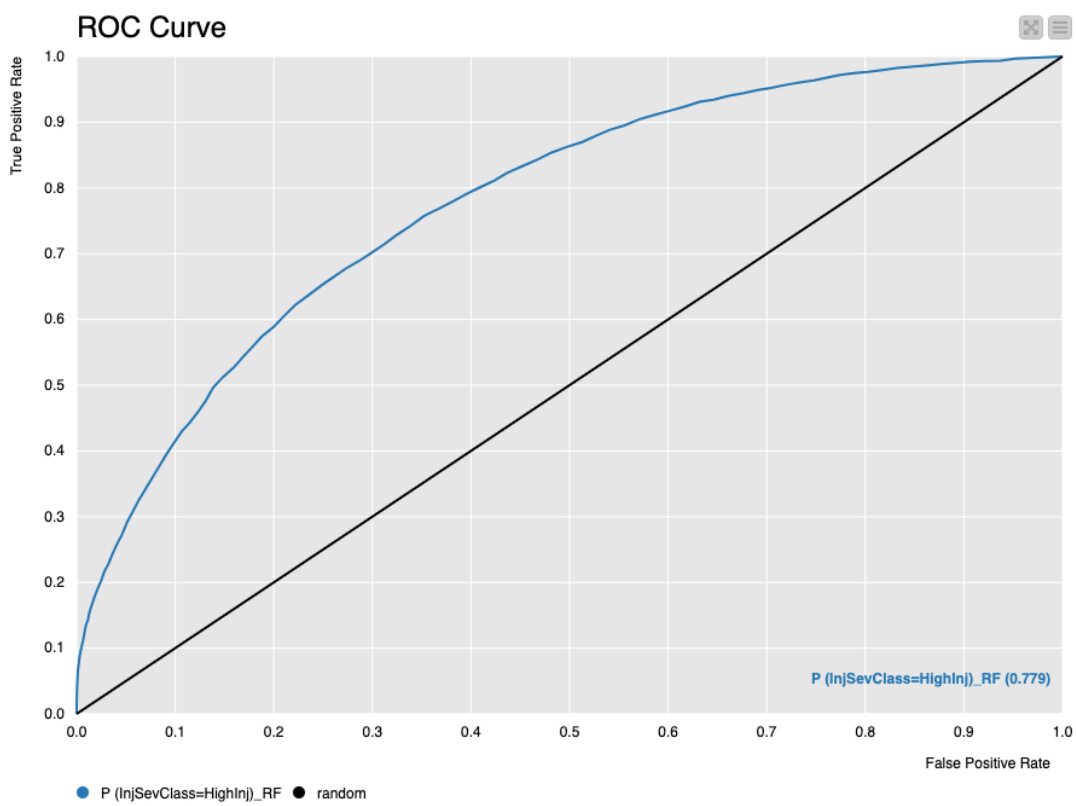
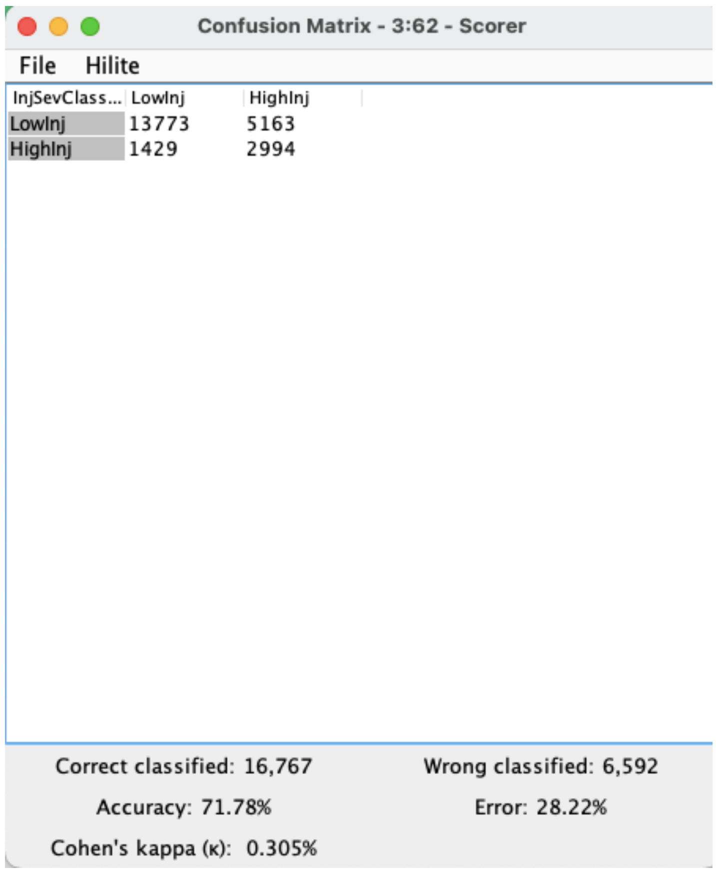


Figure 9: Random Forest ROC Curve.

C- Neural Network (MLP type) Model:

The third model utilizes a neural network (MLP) structured with multiple connected nodes to preprocess the data before feeding it into the machine learning algorithm. Specifically, the variables and data are converted into numerical categories via a one-to-many node. The data then passes through a normalization node to standardize the data range. After preprocessing, the data goes into partition and MLP learner nodes, which have the InjSevClass column set as the target variable. Accuracy metrics are generated by ROC Curve and Scorer nodes in KNIME. The full workflow with intermediate outputs is visualized in the accompanying figures.


Confusion Matrix - 3:62 - Scorer

File Hilite

InjSevClass...	LowInj	HighInj
LowInj	13773	5163
HighInj	1429	2994

Correct classified: 16,767 Wrong classified: 6,592

Accuracy: 71.78% Error: 28.22%

Cohen's kappa (κ): 0.305%

Figure 10: Neural Network Scorer.

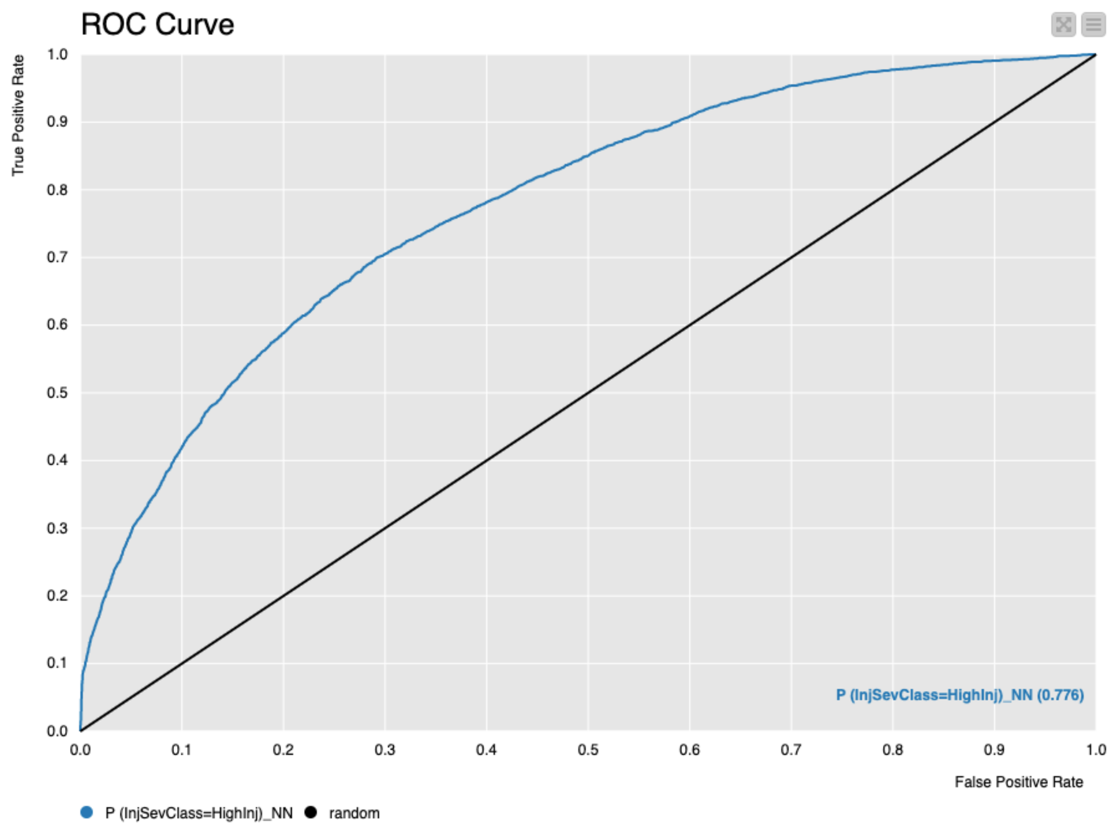
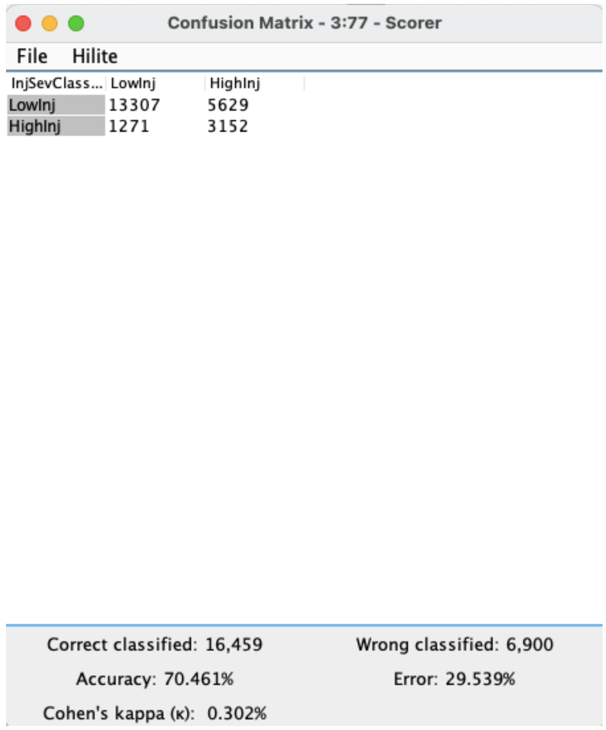


Figure 11: Neural Network ROC Curve.

D- Logistic Regression (LR) Model:

The final model employs logistic regression (LR), structured in KNIME with a series of nodes for data preprocessing before the machine learning algorithm. As in the prior neural network model, the raw data is converted to numerical categories via a one-to-many node, then normalized across variables through a normalization node. The processed data is partitioned and fed into the LR learner, with the InjSevClass set again as target output variable. Model accuracy is evaluated by routing the LP model predictions into ROC Curve and Scorer nodes native to KNIME. These provide numeric metrics and visual evaluation of how well injury severity level is predicted. The complete workflow - data transformations, LP model building, accuracy checking - is visualized in the accompanying figures within the KNIME interface.



Confusion Matrix - 3:77 - Scorer

File Hilite

InjSevClass...	LowInj	HighInj
LowInj	13307	5629
HighInj	1271	3152

Correct classified: 16,459 Wrong classified: 6,900

Accuracy: 70.461% Error: 29.539%

Cohen's kappa (κ): 0.302%

Figure 12: Logistic Regression Scorer.

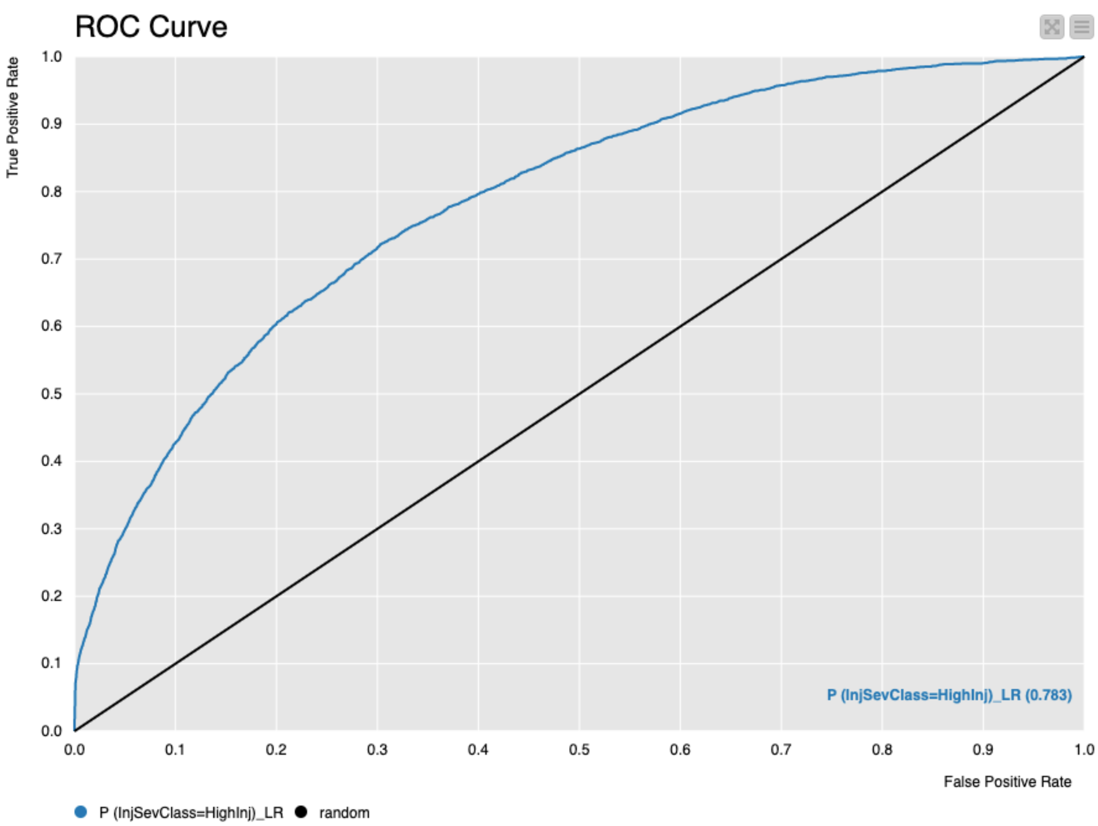


Figure 13: Logistic Regression ROC Curve.

Testing and Evaluation:

To compare the performances of all the models we used in this report, Decision Tree, Random Forest, Neural Network, and Logistic Regression, we created an ROC curve that combines the output of each into one chart, Figure 15. The Logistic Regression model performed better than the other models with 0.783 accuracy. On the other hand, the model that yielded the lowest prediction results is Decision Tree model with 0.645 accuracy. Therefore, the Logistic Regression model is preferred to the other models we built for having more accuracy and less error rate.

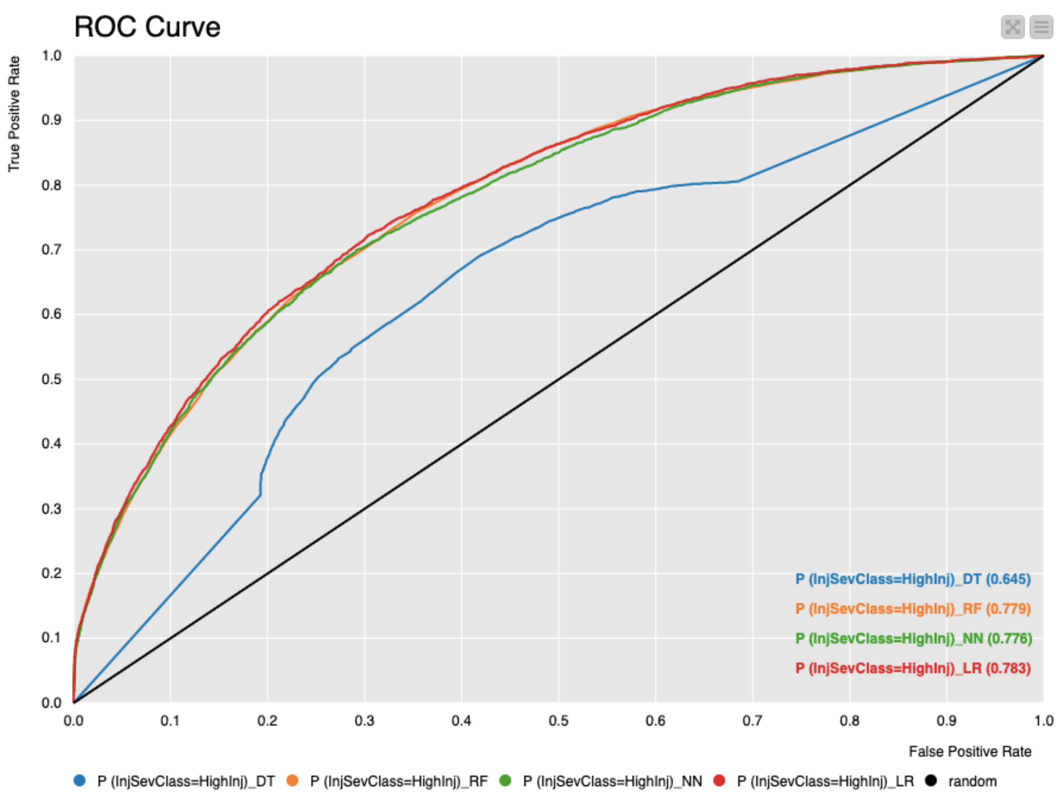


Figure 14: ROC Curve combining all the models.

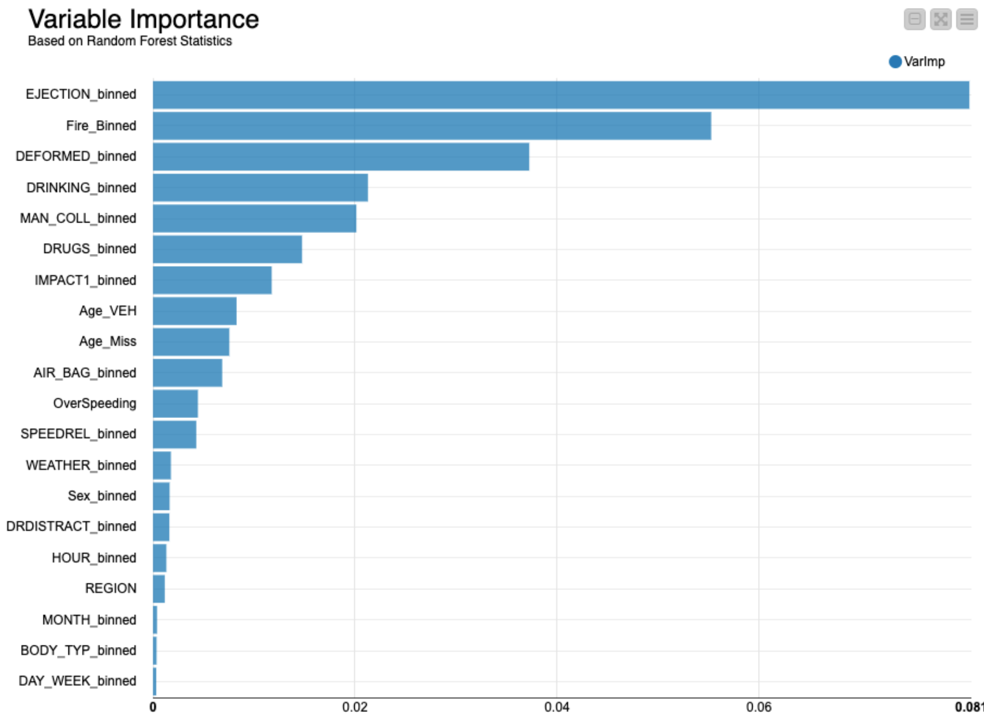


Figure 15: Random Forest Importance Graph

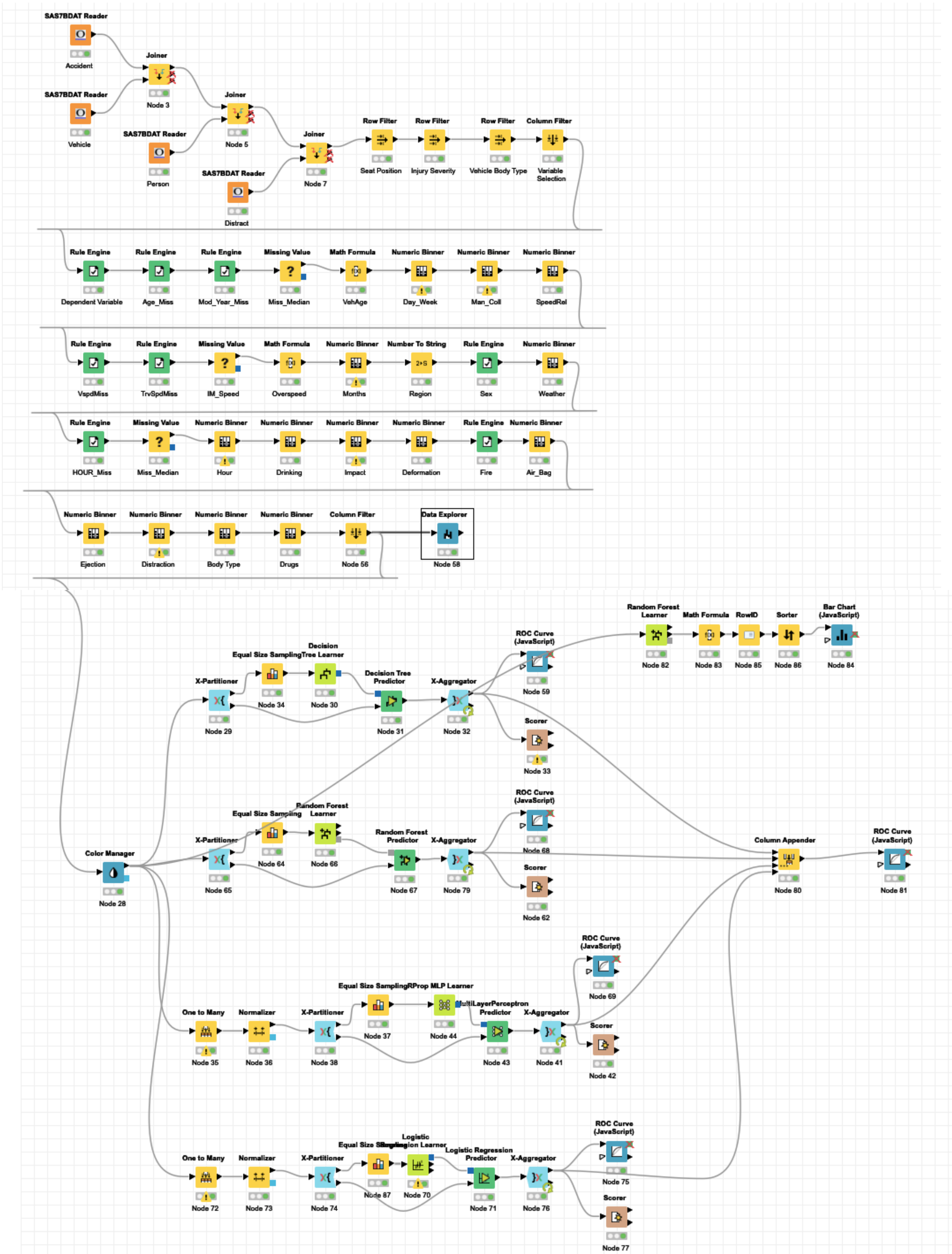


Figure 16: Knime Workflow.