Oklahoma State University

PREDICT INJURY SEVERITY USING RISK FACTORS IN AUTOMOBILE CRASHES

Predictive Analytics Technologies

Dr. Dursun Delen

Team 2

Manjusree Paimagham

Rishi Poudyal

Hieu Nghiem

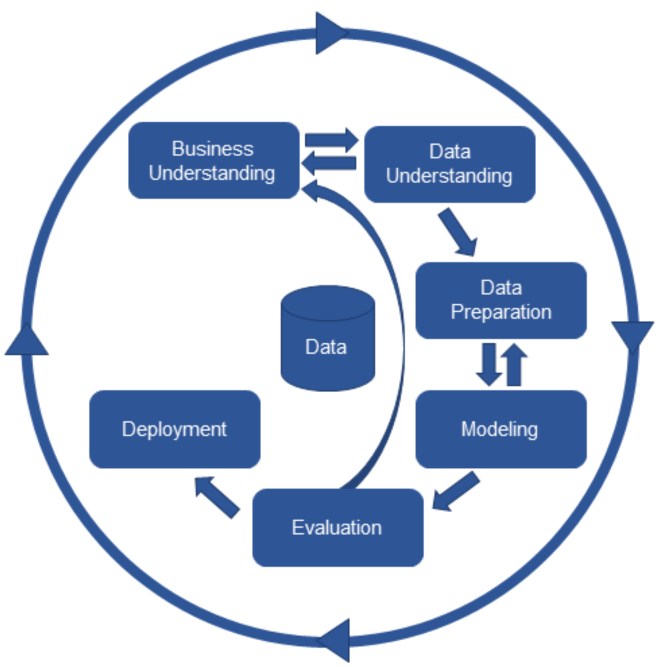
Ahmed Sodeinde

Table of Contents

Executive Summary

The National Highway Traffic Safety Administration has collected crash data since the early 1970s with a view to reduce motor vehicle crashes, injuries, and deaths on US highways. The Crash Report Sampling System (CRSS), an arm of the NHTSA's crash data collection program is a sample of police-reported crashes that includes all types of motor vehicles, pedestrians, and cyclists, ranging from property-damage-only crashes to those that result in fatalities. CRSS is effectively used to project the overall crash picture, identify traffic safety problem areas, observe trends, implement consumer information initiatives, and form the basis for cost/benefit analyses of highway safety initiatives and regulations. The data is from a nationally representative sample of the roughly 6 million police-reported crashes that occur in the country.

Using a sample of car crash data from CRSS, we built and test 6 predictive models which includes Decision Tree, Naïve Bayes, Random Forest, Logistic Regression, Artificial Neural Network, Support Vector Machines to predict the injury severity of drivers in car crashes whether it is high or low. We follow the CRISP-DM process, which is stand for Cross-Industry Standard Process for Data Mining for the completion of the project.



1. Business Understanding

The first step is Business Understanding. We have car crashes data in 2017 from CRSS (Crash Report Sampling System). We will choose to predict injury severity using the variables which are characteristics or risk factors of people, vehicle and environment when crashes happened.

The main objective of this project is to build a predictive model to identify the factors that are responsible for the accidents. Main causes of collisions and crash related injury seriousness are of exceptional worry to overall population, yet particularly to researchers since such examination would be pointed at avoidance of crashes as well as at decrease of their critical consequences, possibly saving numerous lives and money. Notwithstanding lab and experimentation-based research strategies, another approach to address the issue is to recognize the most probable factors that influence injury seriousness by mining the data on vehicle crashes. Close comprehension of the complex conditions where drivers and additionally travelers are bound to experience severe wounds or even be killed in an automobile collision has an incredible potential to reduce the dangers associated with car accidents and subsequently advance the prosperity of individuals involved in these car accidents.

Project Approach

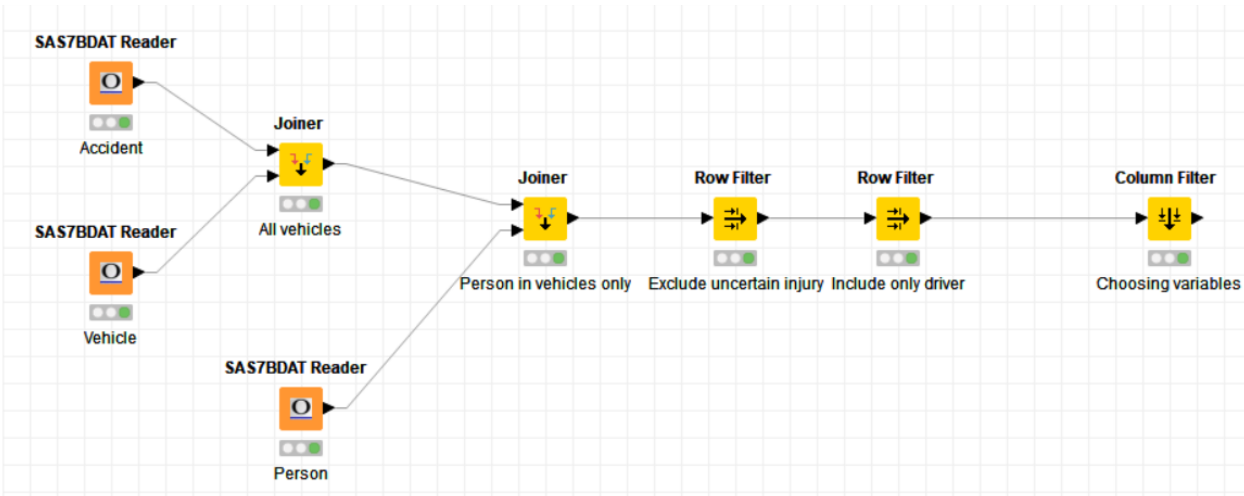
Our team consists of 4 members. First, we analyzed the dataset given to us. To have a better understanding of the dataset, we have gone through domain literatures, articles related to the dataset and gained some domain knowledge. We divided the project into 4 parts and each one of us took care of our respective portion of the project. We had weekly meetings to discuss about the progress of our work. We used KNIME tool to analyze the data workflow, explore data, perform data preprocessing (selecting data, characterizing and aggregating the data) and build predictive models. We used JMP to assess more results of the model we chose.

Target Variable

The dependent variable in our project is "***Injury Severity***". It is a nominal variable, however we converted it to a binary variable with two levels which are low injury severity and high injury severity.

2. Data Understanding

We used the data acquired from NASS GESS. We used the data obtained from three separate datasets – accidents, vehicles and people. We combined these three datasets and chose the variables that is important for our study. Accidents dataset is all about the road accidents, weather conditions and accident related settings. Vehicle dataset contains information about the type, make and model year of vehicles involved in the crashes. The persons dataset is about definite demographics, injury and situational data about the driver and the travelers affected from the car accident. We combined these three datasets into a single dataset using data preprocessing techniques. At this point, the dataset contains nearly 55,000 accidents in 2017 and most of them are categorical variables. Below figure is our workflow for combining data in KNIME – a popular open source data mining tool which we used in this project:



- Inspecting the dataset: vehicle, accident and people data. Reading through the analytical data manuals.
- Right join the Accident dataset with Vehicle dataset to make sure all vehicles information will be included.
- Left join above combination dataset with People dataset to exclude people which are not in a vehicle during crashes.
- We have 8 level of injury severity (INJ_SEV - our target variable). However, level 0, 5, 6, 9 are uncertain so we will exclude them from the dataset. We will regroup the rest levels to: 1, 2 as low injury severity; 3, 4 as high injury severity.

- We will include just the driver of the vehicle only, so we will filter using condition SEAT_POS = 11.
- For consolidation dataset which ready to perform data preprocessing, we will choose 29 most appropriate factors which including risk for car crashed. All missing values are already imputed in the dataset.

3. Data Preparation

As our data is very large and complex, it requires extensive preprocessing. The accident dataset finally has 54969 records. The accident dataset with 51 columns is joined with the Vehicles dataset which has 97625 records and 87 columns so that the resultant table has 137 columns and 97625 records each corresponding to a vehicle. The personal data is left joined to the resultant table with vehicle number (VEH_NO) in order to include the persons who were traveling the vehicle only. Now, our final table has 133608 rows and 196 columns. In KNIME, the row filter nodes were used to include only driver in the vehicles and to exclude uncertain injuries from INJ_SEV which is our target variable.

SAS Name: INJ_SEV

Attribute Codes

2016-Later

0 No Apparent Injury *(O)*
1 Possible Injury *(C)*
2 Suspected Minor Injury *(B)*
3 Suspected Serious Injury *(A)*
4 Fatal Injury *(K)*
5 Injured, Severity Unknown *(U)*
6 Died Prior to Crash
9 Unknown/Not Reported

SAS Name: SEAT_POS

Attribute Codes

2016-Later

0 Not a Motor Vehicle Occupant
11 Front Seat – Left Side *(Driver's Side)*
12 Front Seat – Middle
13 Front Seat – Right Side
18 Front Seat – Other
19 Front Seat – Unknown
21 Second Seat – Left Side
22 Second Seat – Middle
23 Second Seat – Right Side
28 Second Seat – Other
29 Second Seat – Unknown
31 Third Seat – Left Side
32 Third Seat – Middle
33 Third Seat – Right Side
38 Third Seat – Other
39 Third Seat – Unknown
41 Fourth Seat – Left Side
42 Fourth Seat – Middle
43 Fourth Seat – Right Side
48 Fourth Seat – Other
49 Fourth Seat – Unknown
50 Sleeper Section of Cab *(Truck)*
51 Other Passenger in Enclosed Passenger or Cargo Area
52 Other Passenger in Unenclosed Passenger or Cargo Area
53 Other Passenger in Passenger or Cargo Area, Unknown Whether or Not Enclosed
54 Trailing Unit
55 Riding on Exterior of Vehicle
98 Not Reported
99 Unknown

Fig: Different levels in INJ_SEV and SEAT_POS variables

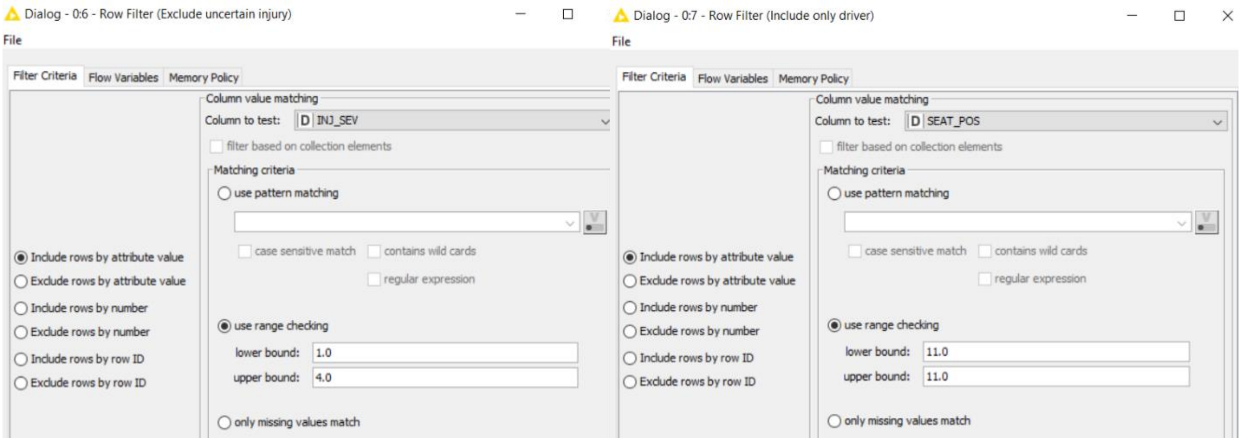


Fig: Excluded uncertain injuries and kept only drivers in our data set using row filters

As we needed only the interested variables from this dataset, we excluded all other variables using column filter in KNIME. Eventually we had 26 variables including CASENUM and INJ_SEV, and 26,809 rows in our final data set.

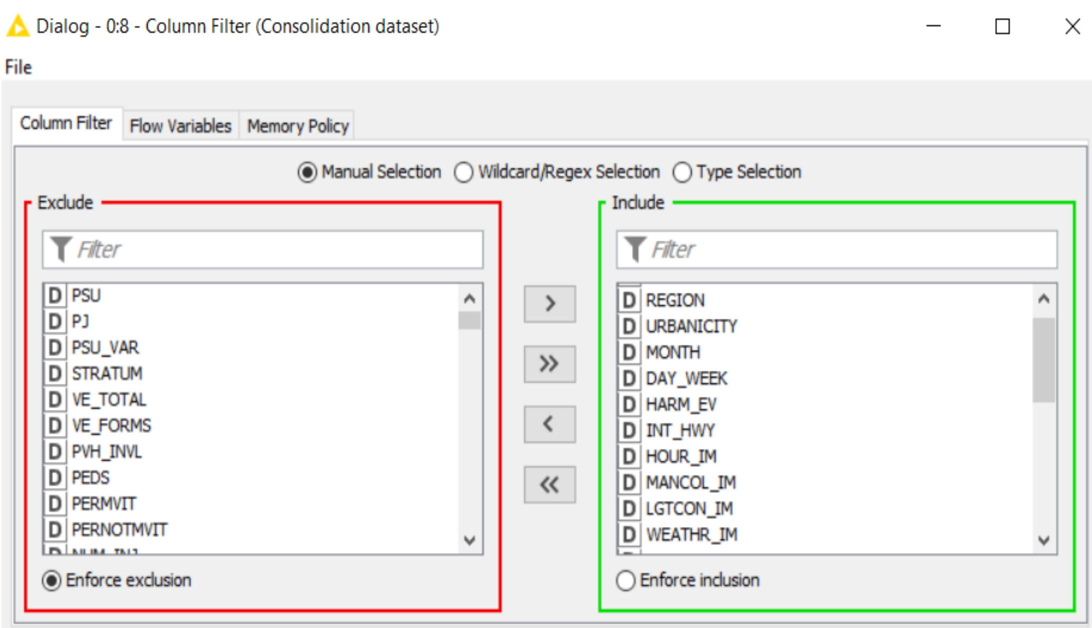


Fig: Extracted 26 variables using column filter

Filtered table - 0:8 - Column Filter (Consolidation dataset)

File Hilite Navigation View

Table "default" - Rows: 26809 | Spec - Columns: 26 | Properties | Flow Variables

Row ID	D CASENUM	D REGION	D URBAN...	D MONTH	D DAY_W...	D HARM_EV	D INT_HWY	D HOUR_IM	D MANCO...	D LGTCO...	D WEATH...	D NUMOC...	D BODY_...	D ROLLO...	D DEFOR...
Row0_Row1_...	201,700,000,...	4	1	3	4	12	0	17	6	1	1	1	89	0	2
Row1_Row2_...	201,700,000,...	4	1	3	5	12	1	7	1	1	1	1	9	0	2
Row1_Row3_...	201,700,000,...	4	1	3	5	12	1	7	1	1	1	1	6	0	4
Row1_Row4_...	201,700,000,...	4	1	3	5	12	1	7	1	1	1	1	34	0	4
Row3_Row8_...	201,700,000,...	4	1	3	5	12	1	20	1	2	10	1	89	0	4
Row6_Row12...	201,700,000,...	3	1	7	1	42	0	4	0	2	1	1	4	0	6
Row11_Row2...	201,700,000,...	3	1	7	3	12	0	8	6	1	1	1	4	0	2
Row12_Row2...	201,700,000,...	3	1	7	7	12	0	15	1	1	10	1	4	0	6
Row16_Row3...	201,700,000,...	3	1	7	7	25	1	15	0	1	2	1	16	0	6
Row18_Row3...	201,700,000,...	3	1	7	5	12	0	13	7	1	10	3	14	0	6
Row19_Row3...	201,700,000,...	3	1	7	3	12	0	16	1	1	1	2	4	0	6
Row21_Row4...	201,700,000,...	3	1	7	3	12	0	12	11	1	1	1	80	0	6
Row24_Row4...	201,700,000,...	3	1	7	5	12	0	14	1	1	1	1	21	0	2
Row24_Row4...	201,700,000,...	3	1	7	5	12	0	14	1	1	1	1	15	0	4
Row26_Row5...	201,700,000,...	3	2	7	1	42	0	6	0	1	1	1	4	0	6
Row27_Row5...	201,700,000,...	3	2	7	5	12	0	11	6	1	10	1	15	0	6
Row28_Row5...	201,700,000,...	3	2	7	4	12	0	17	6	1	1	1	80	0	4
Row30_Row6...	201,700,000,...	3	1	7	2	12	0	6	6	1	1	1	4	0	6
Row33_Row6...	201,700,001,...	3	1	7	5	12	0	3	6	3	1	1	14	0	6
Row34_Row6...	201,700,001,...	3	1	7	3	43	0	16	0	1	2	1	4	0	6
Row38_Row7...	201,700,001,...	4	1	3	5	35	1	1	0	2	1	2	34	9	6

Fig: Final data set

Most of the variables are categorical having large number of levels with large discrimination between them (Imbalanced), so we decided to bin these levels to get better response from the models. For instance, we binned the months 1, 2, 11, 12 into "1" (winter) and rest of the months as not winter ("0"). Similarly, for AIR_BAG, not deployed as "0", deployed as "1" and unknown as "2"; for number of occupants (NUMOCCS) one as "0" and more than one as "1". The summary of the binning process is given in the table below.

No.	Predictors	File	Note	Processing	Type of Variable	KNIME node
1	MONTH	Accident	Convert to 1 and 0 (winter or not) 11, 12, 1, 2: Winter	0: Not winter 1: Winter	Categorical	Rule Eng
2	HOUR_IM	Accident	Morning: 5am to 11; Day: 11 to 7pm; Night: 7pm to 5am	0: Morning 1: Day 2: Night	Categorical	Rule Eng
3	VEH_AGE	Vehicle	2017 - MDLYR_IM. Some 2018 model will have negative age so change the age to 0		Numerical	Math Formula & Rule Engine
4	AGE_IM	People	Keep the same		Numerical	
5	SEX_IM	People	1 Male 2 Female	0 Female 1 Male	Categorical	
6	AIR_BAG	People	Combine 98 and 99: unknown; 20: Not deployed; Others: deployed	0: Not deployed 1: Deployed 2: Unknown	Categorical	Rule Eng
7	BODY_TYP	Vehicle	Categorize as the headers	0: Automobiles (1->10 and 17) 1: Utility Vehicles (14,15,16,19) 2: Truck and Buses (TRUE) 3: Motor Cycles and Others (80->99)	Categorical	Rule Eng
8	VTRAFWAY	Vehicle	8 and 9: Others; Others keep the same	0 Non-Trafficway or Driveway Access 1 Two-Way, Not Divided 2 Two-Way, Divided, Unprotected Median 3 Two-Way, Divided, Positive Median Barrier 4 One-Way Trafficway 5 Two-Way, Not Divided With a Continuous Left-Turn Lane 6 Entrance/Exit Ramp 7 Others	Categorical	Rule Eng
9	DEFORMED	Vehicle	0 and 2; 8 and 9; others keep the same	0: No or Minor damage 1: Functional Damage 2: Disabling Damage 3: Unknown Damage	Categorical	Rule Eng
10	NUMINJ_IM	Vehicle	Keep the same		Numerical	
11	REGION	Accident	1 NE 2 MW 3 S 4 W	0 Northeast (PA, NJ, NY, NH, VT, RI, MA, ME, CT) 1 Midwest (OH, IN, IL, MI, WI, MN, ND, SD, NE, IA, MO, KS) 2 South (MD, DE, DC, WV, VA, KY, TN, NC, SC, GA, FL, AL, MS, LA, AR, OK, TX) 3 West (MT, ID, WA, OR, CA, NV, NM, AZ, UT, CO, WY, AK, HI)	Categorical	Rule Eng
12	URBANICITY	Accident	1 Urban 2 Rural	0 Urban 1 Rural	Categorical	Rule Eng
13	V_ALCH_IM	Vehicle	1 Alcohol 2 No alcohol	0 No alcohol 1 Alcohol	Categorical	Rule Eng
14	WEATHER_IM	Accident	1: clear; 2: rain; 10: cloudy; Combines others: Bad weather	0: Clear 1: Rain 2: Cloudy 3: Bad weather	Categorical	Rule Eng
15	MANCOL_IM	Accident	0: Not collision; 1+2: Front; 7+8+9+10+11: rear and other 6: angle	0: No collision 1: Front collision 2: Angle collision 3: Rear, Side and others	Categorical	Num_bin
16	DAY_WEEK	Accident	Keep the same	1: Sunday 2: Monday => 7: Saturday	Categorical	
17	INT_HWY	Accident	Keep the same	0: No 1: Yes	Categorical	
18	NUMOCCS	Vehicle	Convert to Categorical variables: 1; More than 1 (include 99 - other);	0: One occurrence 1: More than One	Categorical	Num_bin
19	ROLLOVER	Vehicle	0: Rollover, Combine other: No rollover	0: No rollover 1: Rollover	Categorical	Num_bin
20	IMPACT1_IM	Vehicle	0: non collision 11, 12, 13: front 5, 6, 7: back 2,3,4, 8, 9, 10, 61-83: side 13-20: Others	0: No impact 1: Front impact 2: Back impact 3: Side impact 4: Others impact	Categorical	Rule Eng
21	VSURCOND	Vehicle	0; 1; 98 + 99: others; combine the rest as something on the road	0 Non-Trafficway or Driveway Access 1 Dry 2 Something on the road 3 Unknown	Categorical	Num_bin
22	LGTCON_IM	Accident	1,4,5, 7: Daylight 2,3,6: Dark	0 Daylight 1 Dark	Categorical	Rule Eng
23	HARM_EV	Accident	Categorize as the headers 1->7, 16,44,51,72: 0 12+54: 1 TRUE: 2	0 Non Collision 1 collision with motor vehicle in transport 2 collision with other objects	Categorical	Rule Eng
24	SPEEDREL	Vehicle	0; 2 and 3 and 4 and 5: overspeed; 8 and 9: unknown	0 No speeding related 1 Overspeeding 2 Unknown	Categorical	Num_bin

Fig: Summary of the binning process

We wanted to know how the age of the vehicle impacts the accident severity, but this is not in our dataset, so we calculated it by subtracting Model year imputed (MDLYR_IM) from 2017. We found some values as -1 due to the presence of model year 2018, and hence converted it to 0. We performed all these steps in KNIME using "Rule engine", "Numeric binner", and "Math Formula". As we used imputed variables which had no missing values, we did not do any missing value treatment. The workflow of the process in KNIME is shown in the figure below.

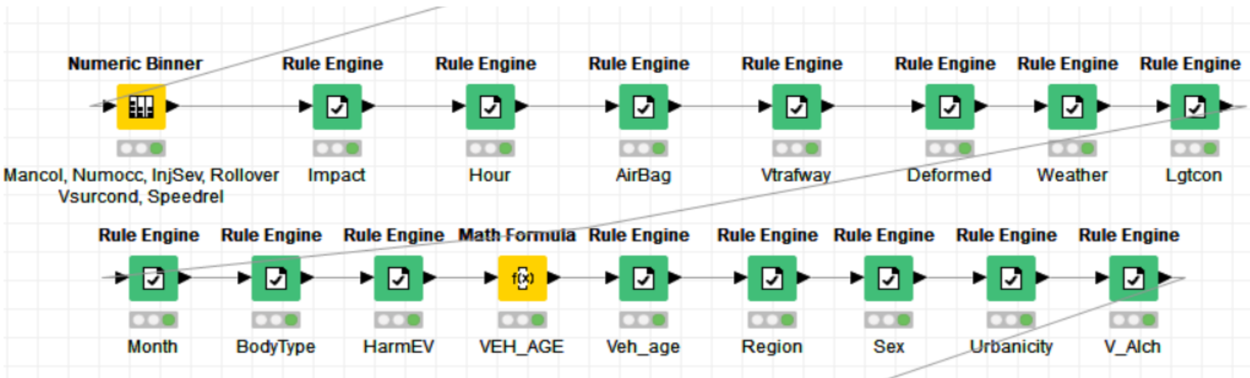


Fig: Workflow of binning in KNIME

The nature of most of the variables is categorical, but they are given in the form of number or integer. We used "Number to String" node in KNIME to change them into string. As we have chosen some number loving models and string loving models like Decision Tree, Naïve Bayes, and Random Forest are string loving models, and Logistic and ANN are number loving models. For number loving models, the independent variables are to be changed back to number which we did in the KNIME by using "Equal Size Sampling" node.

Descriptive Statistics of the Independent Variables:

Descriptive statistics are numbers or values which are used to summarize and describe the data. The mean, median, standard deviation, range, minimum value, maximum value is used to describe numeric variables while the mode is used to describe the categorical variables. Here, the descriptive statistical table for our data that is our target variable and predictors.

Variables	Description	Data type	Number of levels	Descriptive statistics
REGION	Region of country where crash occurred	Nominal	4	South:14126, Northeast:2899

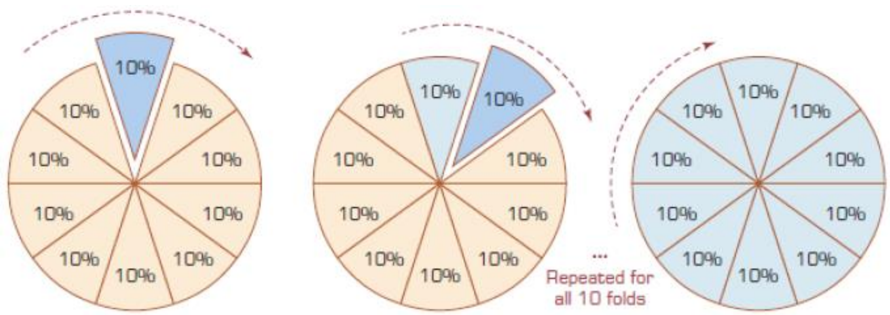
UNBANICITY	Whether Geographical area is urban or rural	Binary	2	Urban:20831, Rural:5923
MONTH	Which month the crash occurred	Binary	2	Winter:8097, Not Winter:18712
DAY_WEEK	Days of a week	Nominal	7	Friday:4405, Sunday:3102
HARM_EV	Damage producing event of the crash	Nominal	3	Collision with Motor vehicle in transport:20802, non-collision:1239
INT_HWY	Interstate Highway	Binary	2	Yes:2618, No:24191
HOUR_IM	Hour of the day	Nominal	3	Day:14496, Night:4850
MANCOL_IM	Manner of collision	Nominal	4	Front collison:10437, Rear side and other:1783
LGTCON_IM	Light condition	Binary	2	Daylight:20006, Dark: 6803
WEATHR_IM	Atmospheric condition	Nominal	4	Clear:19556, Badweather:525
NUMOCCS	Number of occupants	Binary	2	One occupant:19906, More than one occupant:6903
NUMINJ_IM	Number of persons injured in a vehicle	Numeric		Mean:1.24, Median:1, Std Dev: 0.627
VEH_AGE	This is age of vehicle in year	Numeric		Mean:8.43, Median:7, Std Dev:6.79
AGE_IM	Age of driver in years	Numeric		Mean: 40.73, median:38, Std Dev:17.17
ROLLOVER	Vehicle's involvement in	Binary	2	No Rollove:24750 Rollover: 2059

	rollover or turn during crash			
DEFORMED	Extent of Damage sustained by a vehicle	Nominal	4	Disabling damage: 16274 No or Minor Damage: 3055
SPEEDREL	Driver’s speed related to crash or not	Nominal	3	No speeding related: 24285 Unknown: 260
VTRAFWAY	Traffic flow before the crash	Nominal	8	Two way, not divided: 10607 No Trafficway or Driveway access: 405
BODY_TYP	Describe general body configuration	Nominal	3	Auto Mobile:14874 Motor cycle and other:2734
VSURCOND	Road surface condition	Nominal	4	Dry: 22384 Unknown:189
V_ALCH_IM	Driver drinking in vehicle	Binary	2	No: 25313 Yes: 1496
INJ_SEV	Injury severity	Binary	2	Low Injury: 21008 High Injury:5801
AIR_BAG	Air bag deployed	Nominal	3	Not Deployed: 14607 Unknown: 2094
SEX_IM	Gender of the person involved in crash	Binary	2	Male: 14097 Female: 12712
IMPACT1_IM	Area of vehicle that produced first instant of injury	Nominal	4	Front impact:13244 Other impact:107

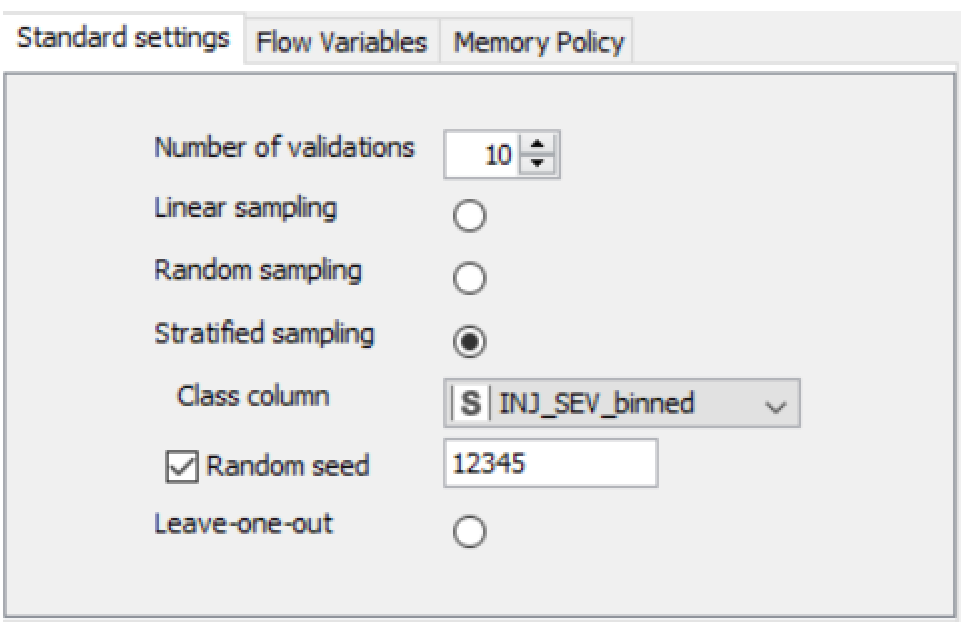
4. Modeling

4.1. Data validation methods

We will use k-fold cross validation method for training and testing our dataset. The complete data set is split in to k mutually exclusive subsets with approximately equal size. The model will be trained using k-1 subsets and validated using the remaining subset. This process repeats k times, which mean every observation are used in both training dataset and validating dataset, and each observation is used for validation exactly once which may reduce the bias of model result.



Studies shows that k = 10 seems to be an optimal value for the number of folds to use. X-Partitioner node in KNIME will be used to perform 10-fold cross validation. We will choose Stratified sampling on target variable INJ_SEV:



4.2. Model selection

Our predictors have 3 numerical variables, and others are categorical. The target variable is injury severity which is binary with two possible outcomes: low injury severity and high injury severity. Based on our knowledge of data mining and machine learning technique for solving this binary classification problem, we will choose the following statistical methods which are used frequently: Decision Tree, Random Forest, Naïve Bayes, Logistic Regression, Artificial Neural Network and Support Vector Machines.

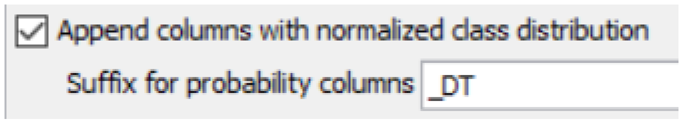
For all models, we choose Learner and Predictor node for training and validating the model; X-Aggregator node for aggregating the results and Scorer node to assess the confusion matrix and model accuracy statistics. However, with Support Vector Machines and Artificial Neural Network model, the input data should be transformed before putting in building model. That's why we use 'One to Many' node to convert categorical variables to dummy variables, and Normalization node to normalize the numerical variables. The transformation method we used is Min-Max normalization with range from 0 to 1.

I 3_REGION_Binned	I 2_REGION_Binned	I 0_REGION_Binned	I 1_REGION_Binned
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
0	1	0	0
0	1	0	0
0	1	0	0
0	1	0	0
0	1	0	0
0	1	0	0
0	1	0	0
0	1	0	0
0	1	0	0
0	1	0	0
0	1	0	0
0	1	0	0
0	1	0	0
0	1	0	0
0	1	0	0

D NUMINJ_IM	D AGE_IM	D VEH_AGE
0	0.276	0.056
0	0.486	0.27
0	0.238	0.011
0.08	0.267	0.281
0	0.248	0.112
0	0.076	0.045
0	0.105	0.135
0	0.105	0.022
0	0.457	0.045
0	0.162	0.112
0.24	0.343	0.112
0	0.21	0.135
0	0.181	0.067
0	0.562	0.011
0	0.381	0.101
0	0.371	0.045
0	0.143	0.124
0	0.362	0.011
0.04	0.105	0.112

The data is unbalance, so first we try to use SMOTE for balancing the data, and the result we've had shows good accuracy, but the sensitivity of the models is low (around 0.3). We try to switch to Equal Size Sampling node in KNIME for training data to make the class attributes occur equally often. Model accuracy reduced a little bit, but the sensitivity is nearly double (around 0.6). Since we prefer the sensitivity which indicate the true positive rate – the rate in predicting high severity injuries, we accept the trade-off to switch to Equal Size Sampling.

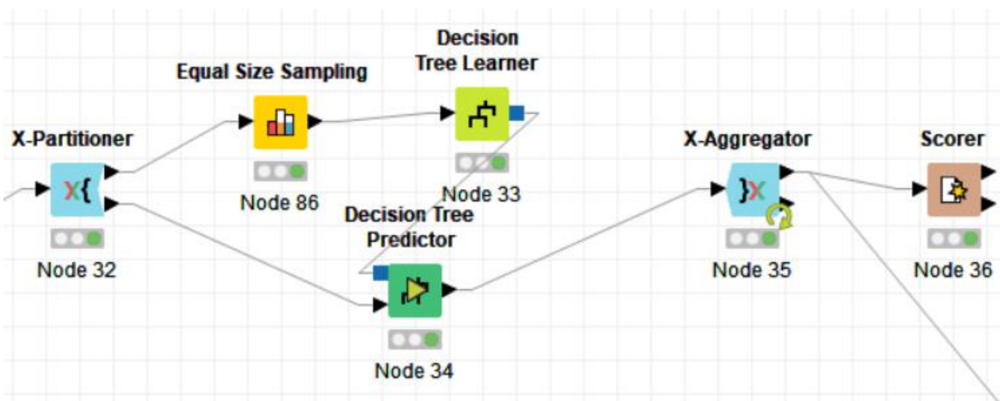
In every Predictor node, we will add a suffix for probability columns, related to model's name. This will make us easier to add these columns for ROC curves assessment between models.



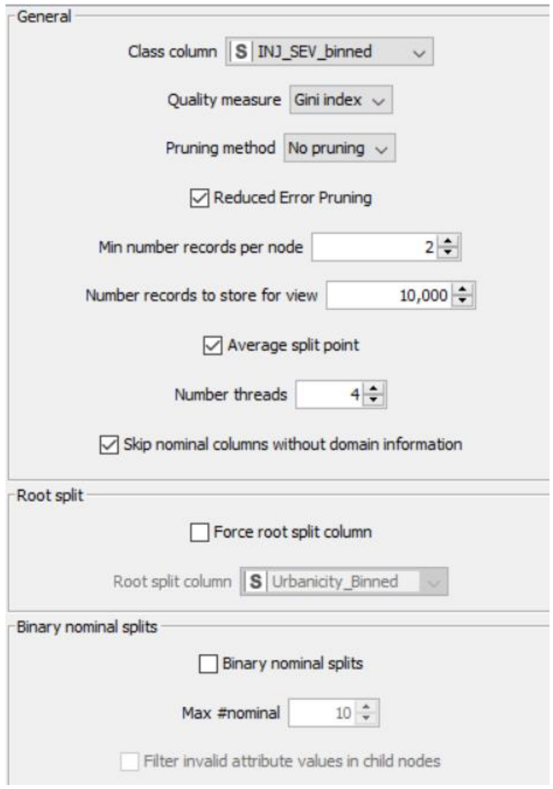
4.2.1 Decision Tree

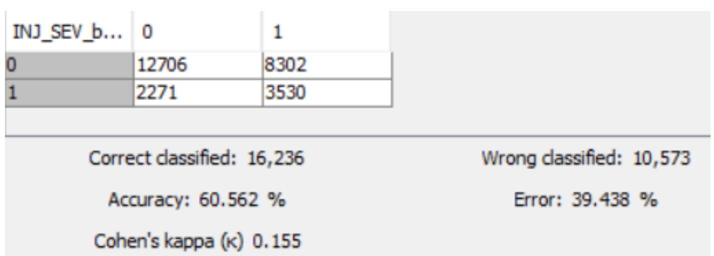
Decision Tree model is a collection of rules that specify how a dataset to be broken up into smaller groups based on the target variable. This model works fine with both categorical and numerical variables.

Below figures indicate the workflow we use to train and validate Decision Tree model in KNIME.



Decision Tree model settings and confusion matrix of the model. Model accuracy is fairly good - 60.562%





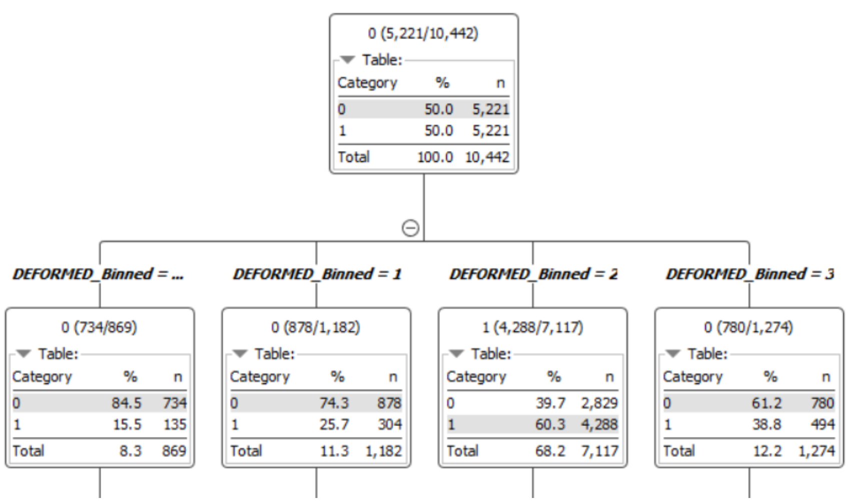
INJ_SEV_b...	0	1
0	12706	8302
1	2271	3530

Correct classified: 16,236 Wrong classified: 10,573

Accuracy: 60.562 % Error: 39.438 %

Cohen's kappa (κ) 0.155

Our model first split using DEFORMED variable which indicate it is the most important variable for our model:

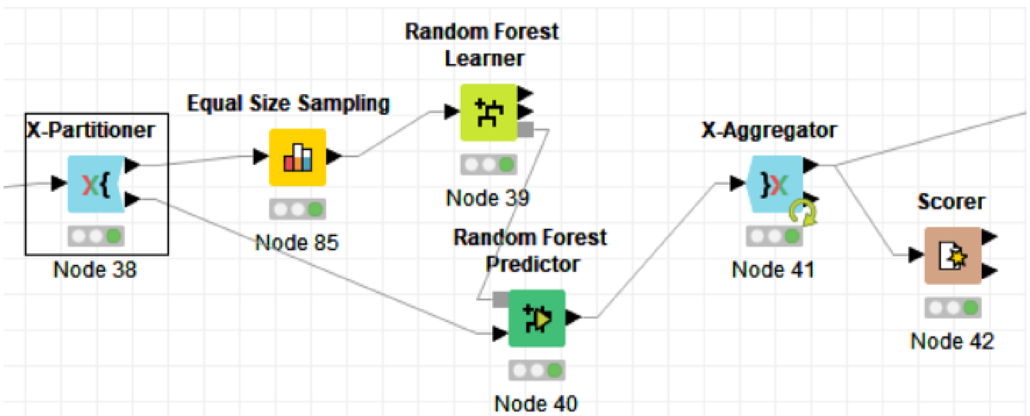


Below table shows the Decision Tree model results like Precision, Sensitivity, Specificity. This result will be used to compare between model in the evaluation part.

Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specifity
0	12706	2271	3530	8302	0.605	0.848	0.605	0.609
1	3530	8302	12706	2271	0.609	0.298	0.609	0.605
Overall	?	?	?	?	?	?	?	?

4.2.2 Random Forest

Random Forest is a machine learning technique which is less prone to overfitting which will work better with minor change in the data. Below figures show the workflow in KNIME, confusion matrix of the model and model performance statistics. Accuracy is higher than Decision Tree model – 68.958%. We use the model default settings in Learner node.



INJ_SEV_b...	0	1
0	14827	6181
1	2141	3660

Correct classified: 18,487 Wrong classified: 8,322

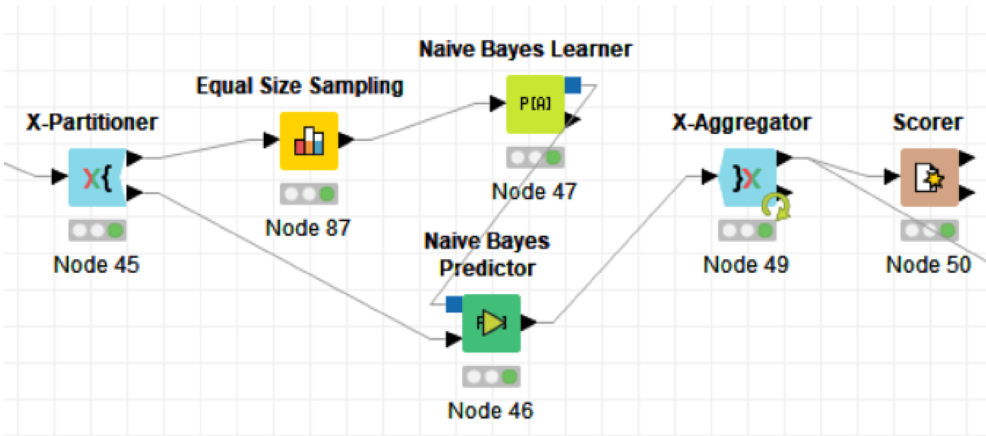
Accuracy: 68.958 % Error: 31.042 %

Cohen's kappa (κ) 0.269

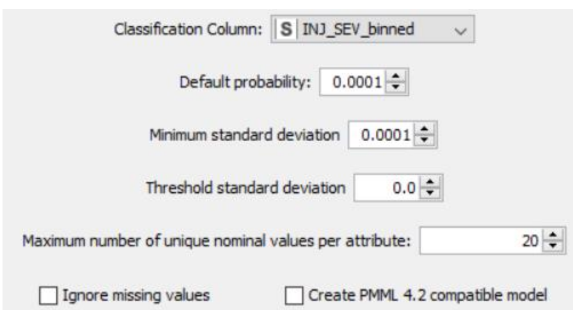
Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specifity
0	14827	2141	3660	6181	0.706	0.874	0.706	0.631
1	3660	6181	14827	2141	0.631	0.372	0.631	0.706
Overall	?	?	?	?	?	?	?	?

4.2.3 Naïve Bayes

This is one of the most simple and popular machine learning classification algorithms – Naïve Bayes algorithm. The workflow in KNIME for implementing Naïve Bayes is in below figure:



We use the default settings for model training. Also look at the tables below, we can see the model confusion matrix, and some statistics of model performance. The accuracy is 67.5%



INJ_SEV_b...	0	1
0	14447	6561
1	2152	3649

Correct classified: 18,096 Wrong classified: 8,713

Accuracy: 67.5 % Error: 32.5 %

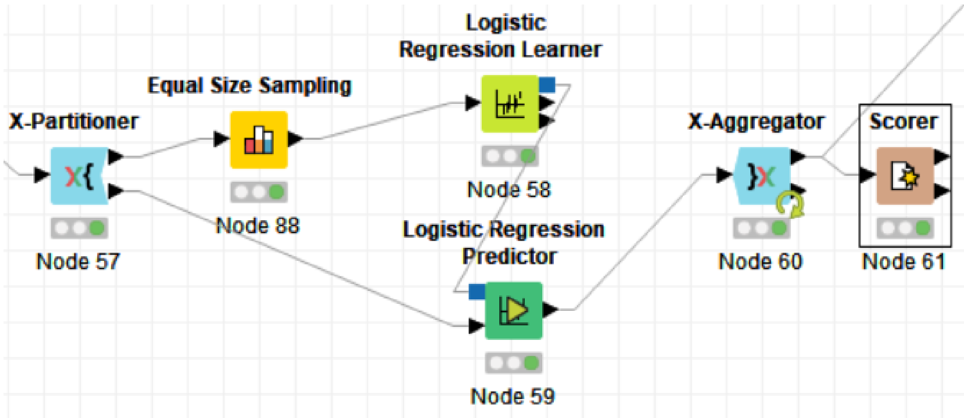
Cohen's kappa (κ) 0.248

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specifity
0	14447	2152	3649	6561	0.688	0.87	0.688	0.629
1	3649	6561	14447	2152	0.629	0.357	0.629	0.688
Overall	?	?	?	?	?	?	?	?

4.2.4 Logistic Regression

Logistic Regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. The dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.), and the model will predict the probability for success or failure of the event.

Logistic Regression is widely used in data mining project. We setup the workflow in KNIME is the same as previous model, using default settings for Learner node:



The confusion matrix show 65.788% in accuracy. Model's sensitivity and specificity will be used to compare with other models:

INJ_SEV_b...	0	1
0	13527	7481
1	1691	4110

Correct classified: 17,637 Wrong classified: 9,172

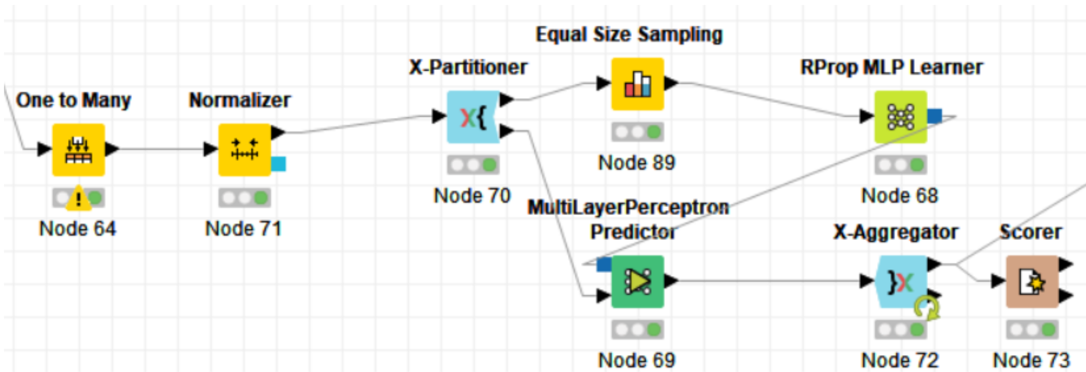
Accuracy: 65.788 % Error: 34.212 %

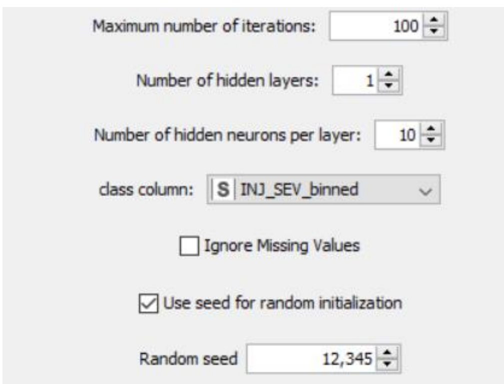
Cohen's kappa (κ) 0.259

Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specifity
0	13527	1691	4110	7481	0.644	0.889	0.644	0.708
1	4110	7481	13527	1691	0.708	0.355	0.708	0.644
Overall	?	?	?	?	?	?	?	?

4.2.5 Artificial Neural Network

ANN or Multi-Layer Perceptron model is developed with the intention to resemble how the human brain works with its ability to learn from experience. For using this model, the data needs to be transformed before putting in X-Partitioner, so the workflow in KNIME is different. We choose one hidden layer with random seed = 12345 in ANN model settings. The accuracy shows 65.124% for this model.





INJ_SEV_b...	0	1
0	13387	7621
1	1729	4072

Correct classified: 17,459 Wrong classified: 9,350

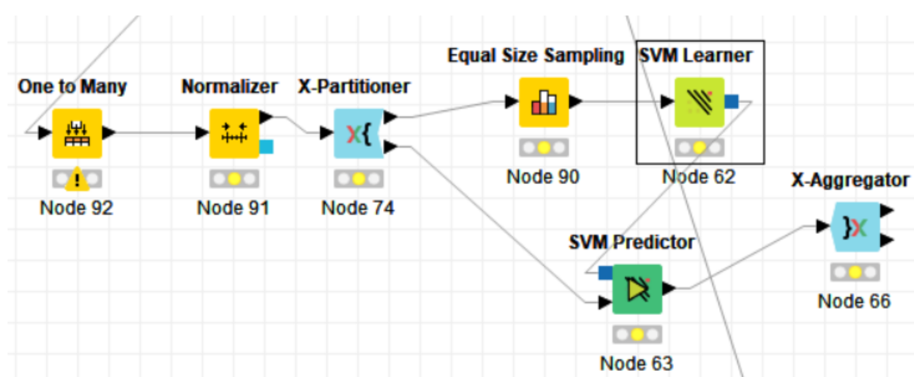
Accuracy: 65.124 % Error: 34.876 %

Cohen's kappa (κ) 0.248

Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specifity
0	13387	1729	4072	7621	0.637	0.886	0.637	0.702
1	4072	7621	13387	1729	0.702	0.348	0.702	0.637
Overall	?	?	?	?	?	?	?	?

4.2.6 Support Vector Machine:

It is a "number-loving" model, that why we use the same workflow as ANN. However we cannot get the results after 6 hours of running the model.

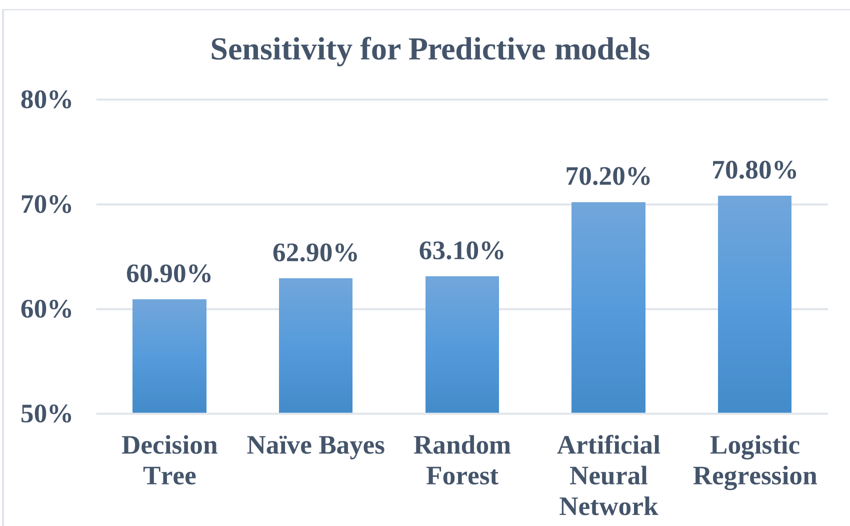


We tried to switch between SVM kernels (Polynomial, HyperTangent, RBF), however the results still didn't show up. Maybe we have to try several SVM optimizations on our future projects to get the results.

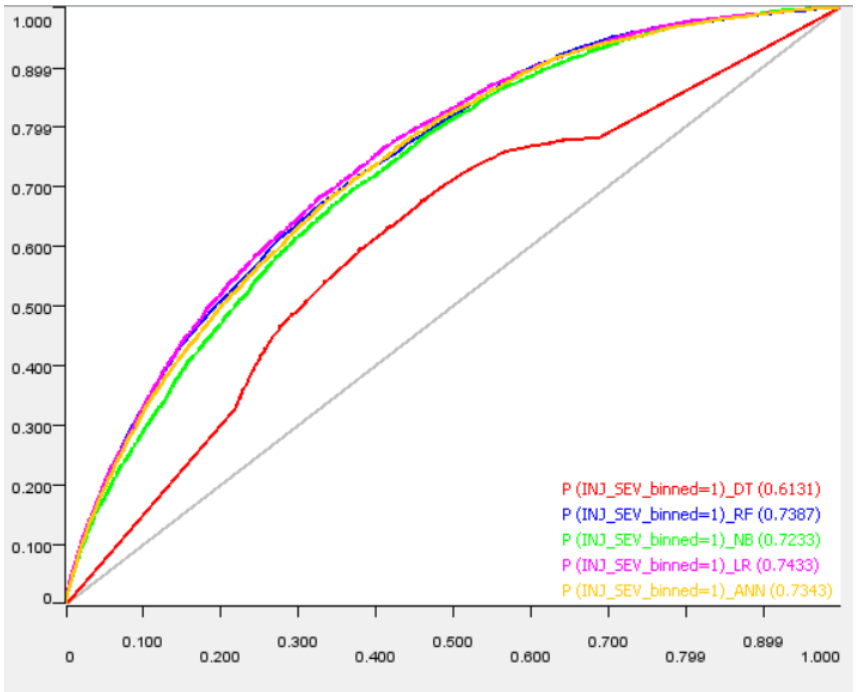
5. Evaluation

We have built 5 models using common machine learning technique, and the next part is to see which the best model is to determine people with high injury severity. Table below shows the comparison between models' accuracy, sensitivity and specificity. Since we are focusing on predicting high severity injuries, so the sensitivity of the model is the criterion we choose.

Model Name	**Accuracy**	**Sensitivity**	**Specificity**
Decision Tree	60.562%	60.9%	60.5%
Random Forest	68.958%	63.1%	70.6%
Naïve Bayes	67.5%	62.9%	68.8%
Logistic Regression	65.788%	70.8%	64.4%
Artificial Neural Network	65.124%	70.2%	63.7%
Support Vector Machine	N/A	N/A	N/A



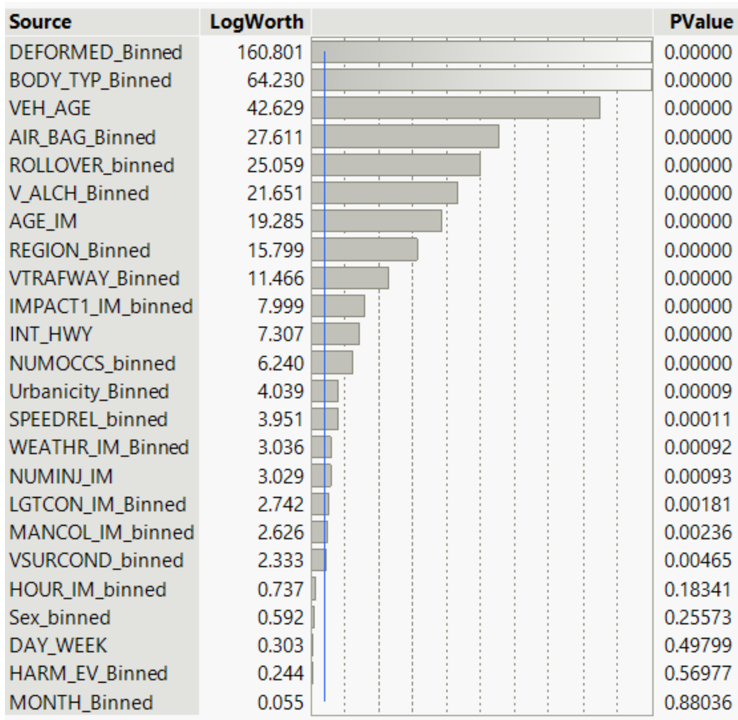
Logistic Regression has the highest sensitivity among 5 models. We will use other determining criterion which are ROC curves of the models. We've configured the ROC plot to show the curves between the actual high injury probabilities versus predicted high injury probabilities. Below figure indicate that Logistic Regression has the best area under the curve (P=0.7433).



Based on these evaluations, we will conclude that Logistic Regression as our best model to predict injury severity for our dataset.

6. Deployment

After evaluating model results, we will use JMP to look at the importance of the variables in our final Logistic Regression model. DEFORMED – amount of damage sustained by the vehicle, BODY_TYP – car type such as sedan or SUV and Vehicle Age is the most important variables, or they affect the most to the high injury severity probabilities of drivers.



Source	LogWorth	PValue
DEFORMED_Binned	160.801	0.00000
BODY_TYP_Binned	64.230	0.00000
VEH_AGE	42.629	0.00000
AIR_BAG_Binned	27.611	0.00000
ROLLOVER_binned	25.059	0.00000
V_ALCH_Binned	21.651	0.00000
AGE_IM	19.285	0.00000
REGION_Binned	15.799	0.00000
VTRAFWAY_Binned	11.466	0.00000
IMPACT1_IM_binned	7.999	0.00000
INT_HWY	7.307	0.00000
NUMOCCS_binned	6.240	0.00000
Urbanicity_Binned	4.039	0.00009
SPEEDREL_binned	3.951	0.00011
WEATHR_IM_Binned	3.036	0.00092
NUMINJ_IM	3.029	0.00093
LGTCON_IM_Binned	2.742	0.00181
MANCOL_IM_binned	2.626	0.00236
VSURCOND_binned	2.333	0.00465
HOUR_IM_binned	0.737	0.18341
Sex_binned	0.592	0.25573
DAY_WEEK	0.303	0.49799
HARM_EV_Binned	0.244	0.56977
MONTH_Binned	0.055	0.88036

The surprise thing is that environment factors like weather, surface condition seem not affect injury severity. Vehicle factors like car age, body type, air bag seem to cause the most effects the risk of high severity. For cars which have risk factors may cause high injury severity, government should propagate the information for citizen about safe driving, and law enforcement should be more strictly implemented.

The critical nature and importance of the subject matter necessitates further analysis before deployment. Government and other stakeholders must refine and meticulously tune models before deployment/scaling.

Conclusion

This project studied the influential factors in the prediction of the severity of injury to drivers in traffic incidents. Six varying machine learning techniques (Decision Tree, Artificial Neural Network, Support Vector Machine, Naïve Bayes, Logistic Regression and Random Forest) with equal size sampling methods are deployed

Logistic Regression is the model we choose to predict the injury severity of people in car crashes. This conclusion is based on ROC curves and accuracy statistics of the models. Below is some insight gained from the models result, which can be used as a springboard for further studies and traffic safety implementations:

- The odds of getting high injury severity will be 5.4 times more if the car has disabling damage.
- Motor cycles will have likely 3 times more the odds of getting high injury severity, compared to Sedan and SUV.
- One unit change in vehicle age will increase the odds of high injury severity by 3.31%.