DATA MINING APPROACH TO REDUCE ACCIDENTS

MSIS 5633 – Business Intelligence Tools & Techniques

Ankur Bansal
Avinash Kaur
Tanmayee Nettem

CONTENTS

EXECUTIVE SUMMARY

Accidents are events with interaction of various components: driver, vehicle, road, environment and many more. Nevertheless, environment and vehicle are the essential components and are directly related to traffic safety.

This accident rate is an issue that has been increasing due to the traffic rate. According to the National Highway Traffic Safety Administration (NHTSA) and research from UAB, in 2015 alone, a total of more than 32,000 people were killed and 2M people were injured in traffic crash. For this reason, road accidents are considered as a worldwide problem.

NHTSA is National Highway Traffic Safety Administration department striving to reduce the damage of the property and human loss due to motor vehicle crashes. It uses data from sources like National Automotive Sampling System (NASS) and General Estimates Systems (GES). NASS GES obtains the data from a country wide representative probability sample which is selected from more than 5 million police-reported crashes each year. NASS GES data is in form of Statistical Analysis System (SAS) files. The current study includes 4 data files: ACCIDENT, VEHICLE, PERSON, DISTRACT.

The purpose of our analysis is to identify the factors that lead to accidents. With the data involving various components, we wish to be able to predict the causes of major injury. This will help NHTSA in taking necessary steps toward safety. The objective of this research is to use data mining techniques and build prediction models that will aid in analyzing the various factors for injury severity. The data mining algorithms include Decision Tree algorithm, Random Forest Algorithm and Naïve Bayes algorithm. The results of each model will be evaluated to choose a best performing model that serves the purpose.

CRISP-DM

BUSINESS UNDERSTANDING

With road safety being a major concern worldwide wide with accidents costing the nation between one to three percent of annual GDP, it is important to take necessary steps towards address this issue and find a solution.

This study aims to analyze and build different types of prediction models to find whether a crash will cause “Minor" or “Major” injury i.e., INJSEV_IM_binned.

Data from all the files is used to determine factors responsible for Severe injury and recommend significant precautionary measures to avoid injuries.

DATA UNDERSTANDING

The SAS data files for this analysis are:

- ACCIDENT

  The data file has information regarding the characteristics of the crash and prevailing environmental conditions during the time of crash.

  Each crash is reported as one record.

  CASENUM is the unique identifies for each record.

- VEHICLE:

  This data file has information regarding only in-transport motor vehicles and the drivers involved in the crash.

  Each in-transport motor vehicle has one record.

  CASENUM and VEH_NO are the unique identifiers for each record

- PERSON

  This data file consists of information regarding all persons involved in the crash including motors and non-motorists.

  Other information such as sex, age, vehicle occupant restraint use, injury severity is also provided.

  Each person is considered in one record.

  CASENUM, VEH_NO and PER_NO are the unique identifiers for each record.

  The motor vehicle occupants of this file are, PER_TYPE = 1,2,3,9 and non-motor vehicle occupants are PERR_TYUPE = 4,5,6,7,8,10 or 19.

- DISTRACT

  This data file has information about distractions that drivers have before the crash that leads to a crash.

  Each distraction is considered as a separate record.

  CASENUM, VEH_NO and MDRDSTRD are unique identifiers for each record.

A few elements are common to the data files such as

CASENUM: Indicates unique case number assigned to each event.

Common to all the data files.

REGION: Identifies the region of the country where the crash happened.
Common to all data files

VEH_NO: A consecutive number assigned to every vehicle in the crash.
Common to the data files DISTRACT, VEHICLE, PERSON

PER_NO: A consecutive number assigned to each person in the crash (i.., each occupant, non-motorists or pedestrians)
Common to the data files DISTRACT, PERSON

The target element in the analysis is the injury severity:

INJ_SEV:	0	No Apparent Injury
	1	Possible Injury
	2	Suspected Minor Injury
	3	Suspected Serious Injury
	4	Fatal Injury
	5	Injured
	6	Died prior to crash
	9	Unknown.

Since, the target element is focused on if the injury is severe or minor, we would categorize all the above categories into 2 levels: Minor or Major

The variables of interest are

DEFORMED: The extent of damage sustained by the vehicle.

AIR_BAG: Records the availability and deployment of air bag.

ROLLOVER: Indicates if rollover occurred for a vehicle

WEATHER: Records the prevailing weather during the time of crash.

SEAT_IM: Indicates the position of passengers in the car during the crash.

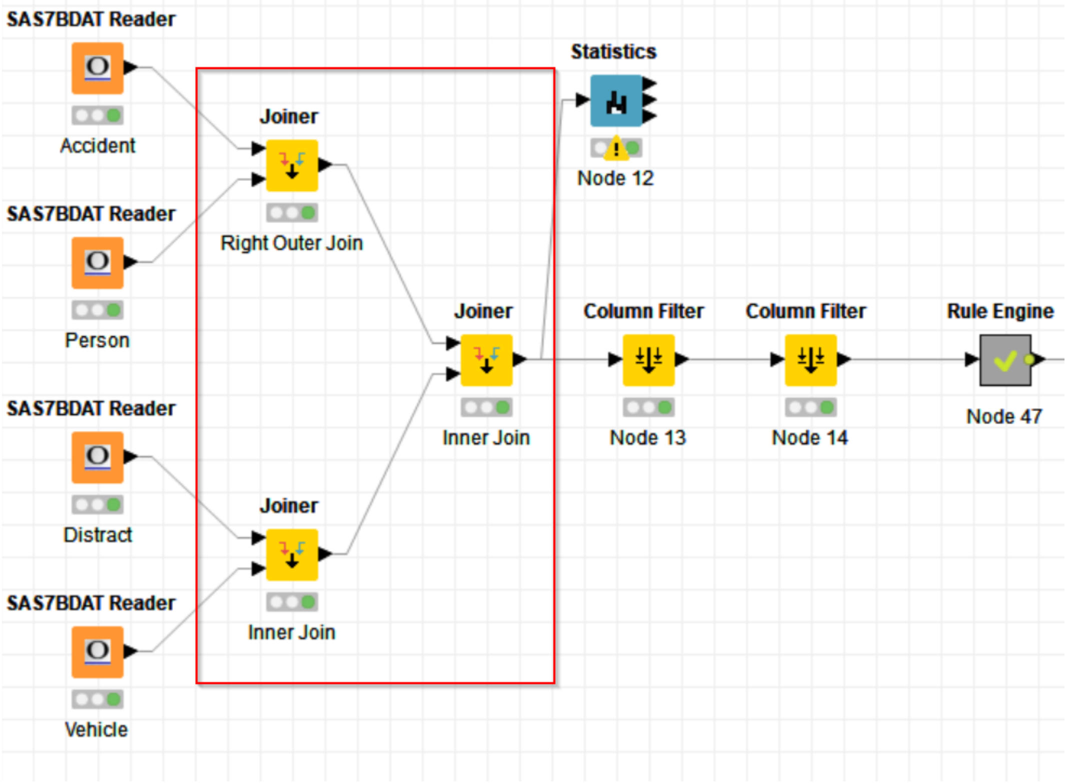
ALC_RES: Records the result of the alcohol test performed on the driver.

VSURCOND: The condition of the surface on which the vehicle was travelling.

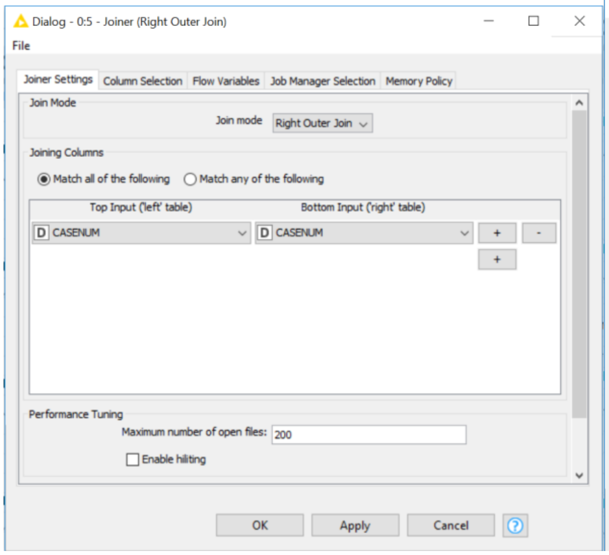
DATA PREPROCESSING

The next phase in CRISP –DM is Data Preprocessing phase. This phase

1. Data integration: - Data Integration is the combination of technical and business processes used to combine data from different multiple sources into meaningful and valuable information to get reliable data source. Here, in our dataset we have combined data from 4 tables namely accident, Person, Distract and vehicle as shown below: -



We have joined Accident and Person data with Right Outer Join using Case number and similarly we have joined Distract and Vehicle data using Inner Join as shown below: -



Join Mode

Join mode: Right Outer Join

Joining Columns

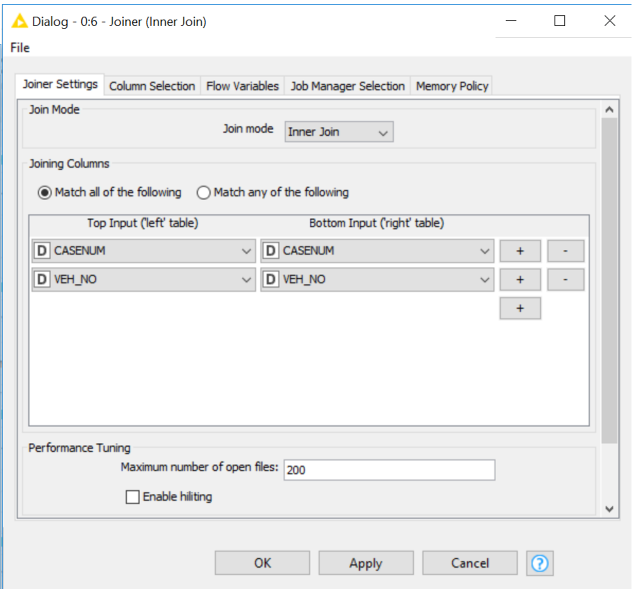
● Match all of the following ○ Match any of the following

Top Input ('left' table)	Bottom Input ('right' table)
D CASENUM	D CASENUM

Performance Tuning

Maximum number of open files: 200

□ Enable hiliting



Join Mode

Join mode: Inner Join

Joining Columns

● Match all of the following ○ Match any of the following

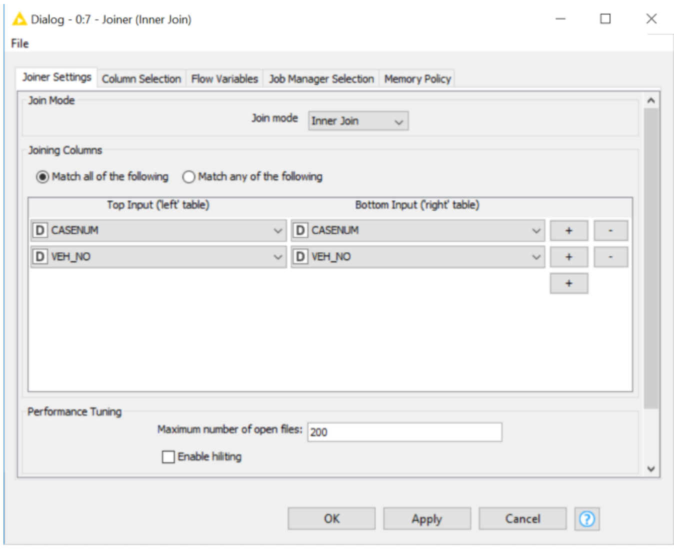
Top Input ('left' table)	Bottom Input ('right' table)
D CASENUM	D CASENUM
D VEH_NO	D VEH_NO

Performance Tuning

Maximum number of open files: 200

□ Enable hiliting

In the end, we will use a final join to join both the resultant tables using inner join as shown below: -



Join Mode

Join mode: Inner Join

Joining Columns

● Match all of the following ○ Match any of the following

Top Input ('left' table)	Bottom Input ('right' table)
D CASENUM	D CASENUM
D VEH_NO	D VEH_NO

Performance Tuning

Maximum number of open files: 200

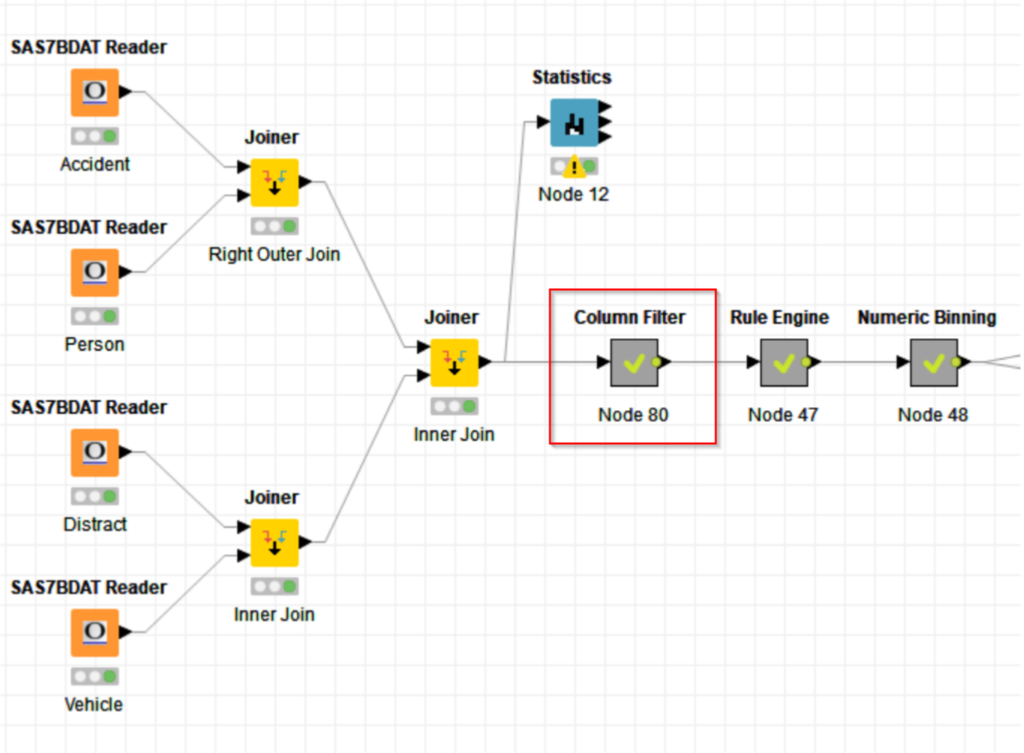
□ Enable hiliting

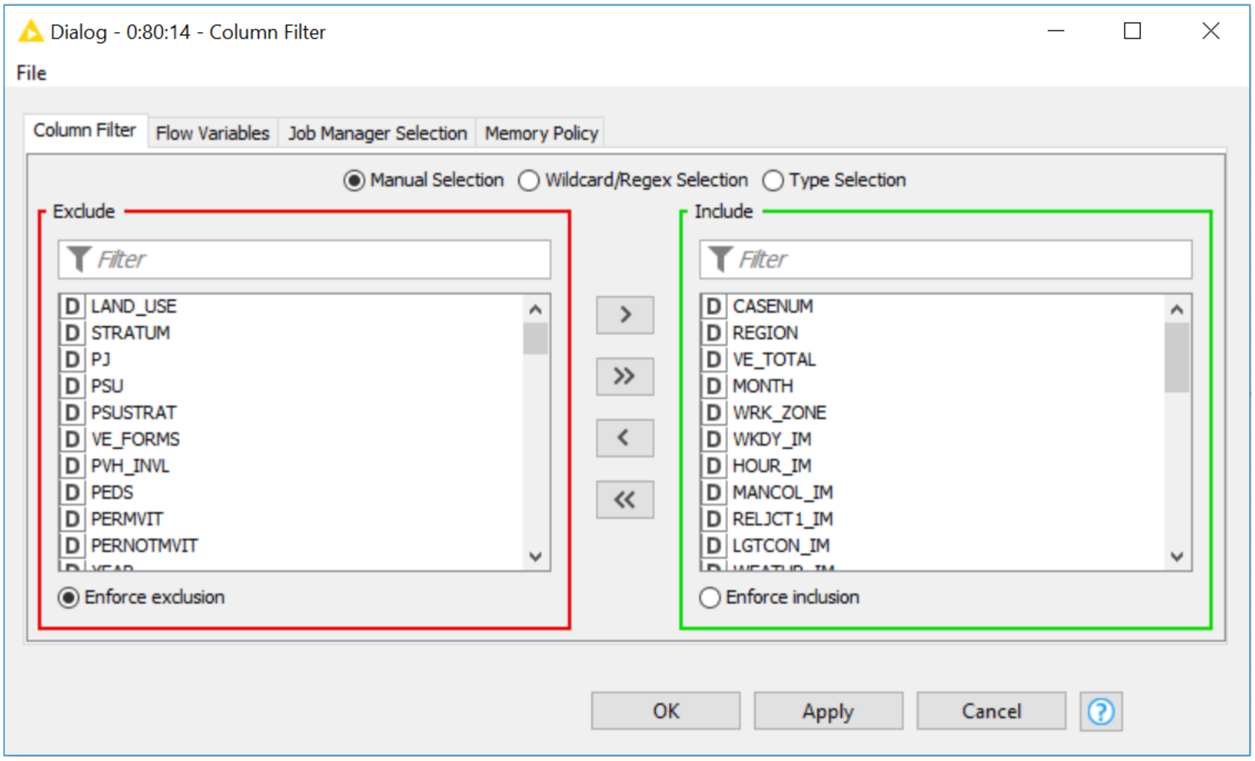
2. Data Cleaning: -

Next step after data integration, is **data cleaning.** We will clean the data and refine our dataset by cleaning the erroneous data or treating the null or the missing values and making sure that we have created columns in such a way that they adhere with the data model formats, so that the models that are formed are efficient, robust and results in optimal performance. Here, we also have the imputed columns for several columns and therefore, which made it easier to further clean the dataset.

3. Dimension Reduction: -

Here, we have reduced the columns by subjective analysis thus reducing the number of initial columns from **195** to **32.** Again, these columns are removed by selecting our target variable. Here, our target variable is Injury severity (INJSEV_IM) and thus the factors that occurred in our data set after the accidents, which makes little sense and no impact on our target variable needs to be excluded from our independent variable list.





Dialog - 0:80:14 - Column Filter

File

Column Filter | Flow Variables | Job Manager Selection | Memory Policy

● Manual Selection ○ Wildcard/Regex Selection ○ Type Selection

Exclude

Filter

D LAND_USE
D STRATUM
D PJ
D PSU
D PSUSTRAT
D VE_FORMS
D PVH_INVL
D PEDS
D PERMVIT
D PERNOTMVIT

● Enforce exclusion

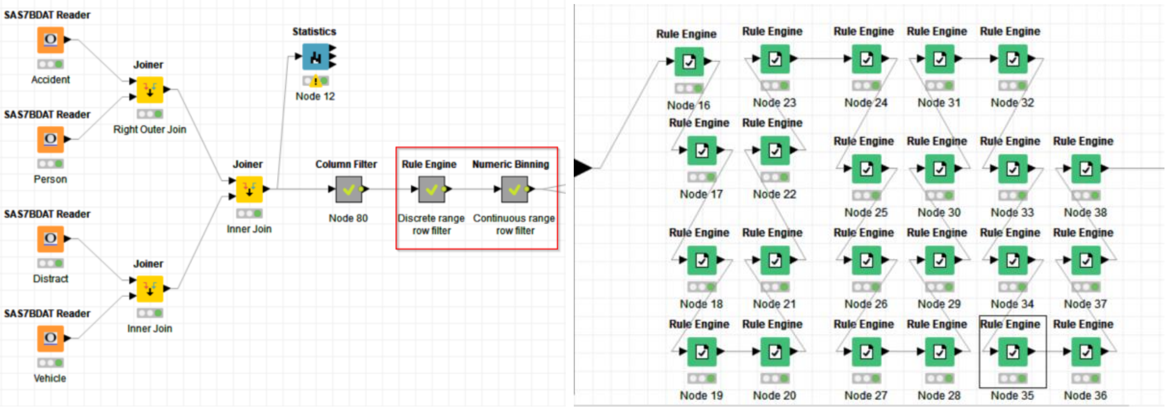
Include

Filter

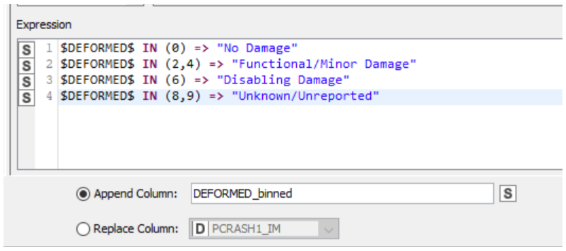
D CASENUM
D REGION
D VE_TOTAL
D MONTH
D WRK_ZONE
D WKDY_IM
D HOUR_IM
D MANCOL_IM
D RELJCT1_IM
D LGTCON_IM

○ Enforce inclusion

After this we will use Rule Engine to categorize several categories for a variable which are similar and, we have created a separate category for a non-reported or unknown value in a variable, thus avoiding any further data loss in our dataset. We have used **22** Rule Engines to categorize the variables and thus category to the required number a shown below: -



We have the following example for the same, Here, we have created rules in Rule Engine for DEFORMED_binned, MONTH_binned, ROLLOVER_binned and VSURCOND_binned into their respective possible minimum categories.

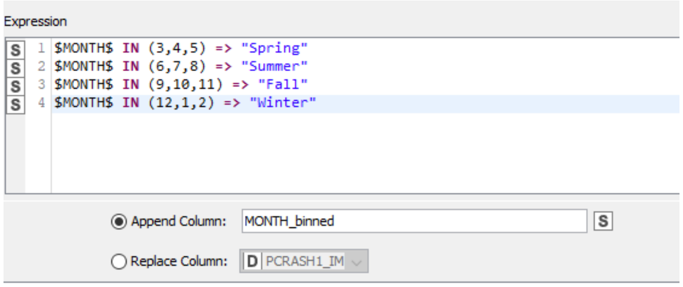


Expression

```
$DEFORMED$ IN (0) => "No Damage"
$DEFORMED$ IN (2,4) => "Functional/Minor Damage"
$DEFORMED$ IN (6) => "Disabling Damage"
$DEFORMED$ IN (8,9) => "Unknown/Unreported"
```

● Append Column: DEFORMED_binned

○ Replace Column: PCRASH1_IM

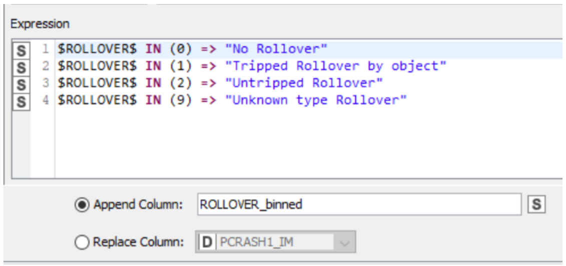


Expression

```
$MONTH$ IN (3,4,5) => "Spring"
$MONTH$ IN (6,7,8) => "Summer"
$MONTH$ IN (9,10,11) => "Fall"
$MONTH$ IN (12,1,2) => "Winter"
```

● Append Column: MONTH_binned

○ Replace Column: PCRASH1_IM

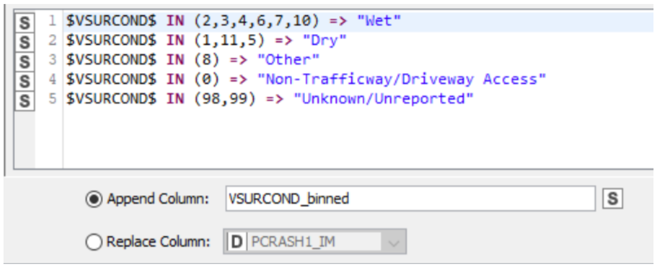


Expression

```
$ROLLOVER$ IN (0) => "No Rollover"
$ROLLOVER$ IN (1) => "Tripped Rollover by object"
$ROLLOVER$ IN (2) => "Untripped Rollover"
$ROLLOVER$ IN (9) => "Unknown type Rollover"
```

● Append Column: ROLLOVER_binned

○ Replace Column: PCRASH1_IM

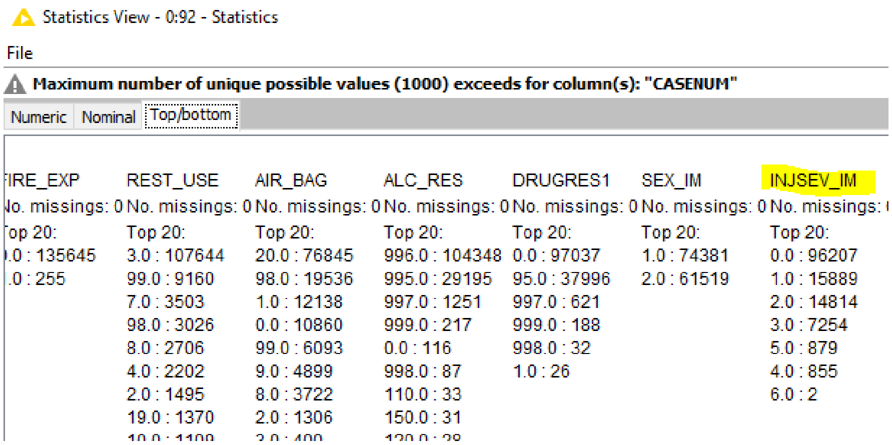


```
$VSURCOND$ IN (2,3,4,6,7,10) => "Wet"
$VSURCOND$ IN (1,11,5) => "Dry"
$VSURCOND$ IN (8) => "Other"
$VSURCOND$ IN (0) => "Non-Trafficway/Driveway Access"
$VSURCOND$ IN (98,99) => "Unknown/Unreported"
```

● Append Column: VSURCOND_binned

○ Replace Column: PCRASH1_IM

Also, apart from this we have used Numeric Binning to categorize and convert our continuous numerical columns to string. Overall, we have used 5 Numeric Binning nodes for our dataset. For example: - We have combined the minimum injury severity and maximum injury severity for 9 different categories of the target variable that was available to us. We have also used statistics analytics to check this before categorizing the variable: -

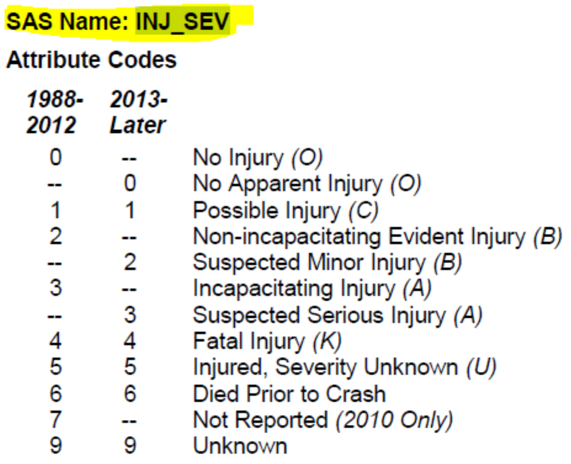


Statistics View - 0:92 - Statistics

File

Maximum number of unique possible values (1000) exceeds for column(s): "CASENUM"

```
IRE_EXP        REST_USE        AIR_BAG         ALC_RES          DRUGRES1        SEX_IM          INJSEV_IM
lo. missings: 0 No. missings: 0 No. missings: 0 No. missings: 0 No. missings: 0 No. missings: 0 No. missings: (
op 20:         Top 20:         Top 20:         Top 20:          Top 20:         Top 20:         Top 20:
.0 : 135645    3.0 : 107644    20.0 : 76845    996.0 : 104348   0.0 : 97037     1.0 : 74381     0.0 : 96207
.0 : 255       99.0 : 9160     98.0 : 19536    995.0 : 29195    95.0 : 37996    2.0 : 61519     1.0 : 15889
               7.0 : 3503      1.0 : 12138     997.0 : 1251     997.0 : 621                     2.0 : 14814
               98.0 : 3026     0.0 : 10860     999.0 : 217      999.0 : 188                     3.0 : 7254
               8.0 : 2706      99.0 : 6093     0.0 : 116        998.0 : 32                      5.0 : 879
               4.0 : 2202      9.0 : 4899      998.0 : 87       1.0 : 26                        4.0 : 855
               2.0 : 1495      8.0 : 3722      110.0 : 33                                       6.0 : 2
               19.0 : 1370     2.0 : 1306      150.0 : 31
               10.0 : 1109     3.0 : 400       120.0 : 28
```

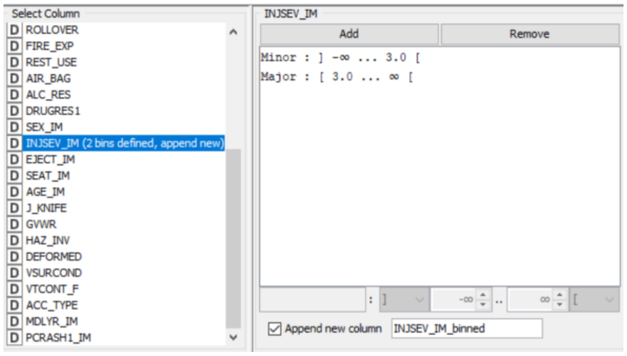


SAS Name: INJ_SEV

Attribute Codes

1988-2012	***2013-Later***	
0	--	No Injury *(O)*
--	0	No Apparent Injury *(O)*
1	1	Possible Injury *(C)*
2	--	Non-incapacitating Evident Injury *(B)*
--	2	Suspected Minor Injury *(B)*
3	--	Incapacitating Injury *(A)*
--	3	Suspected Serious Injury *(A)*
4	4	Fatal Injury *(K)*
5	5	Injured, Severity Unknown *(U)*
6	6	Died Prior to Crash
7	--	Not Reported *(2010 Only)*
9	9	Unknown

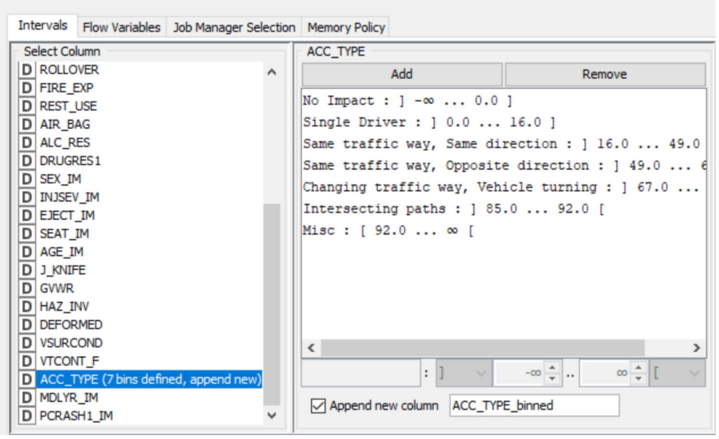
We can see the below Numeric Binner for the same variable:-



```
Minor : ] -∞ ... 3.0 [
Major : [ 3.0 ... ∞ [
```

☑ Append new column: INJSEV_IM_binned

And similarly, we have converted other continuous numeric variables into string. Here, we have converted Accident Type (ACC_TYPE): -

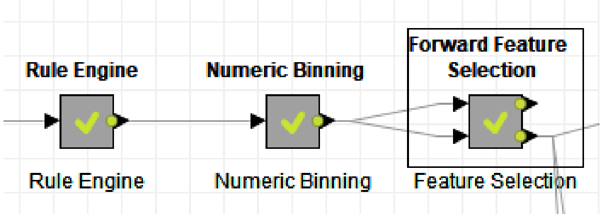


```
No Impact : ] -∞ ... 0.0 ]
Single Driver : ] 0.0 ... 16.0 ]
Same traffic way, Same direction : ] 16.0 ... 49.0
Same traffic way, Opposite direction : ] 49.0 ... 6
Changing traffic way, Vehicle turning : ] 67.0 ...
Intersecting paths : ] 85.0 ... 92.0 [
Misc : [ 92.0 ... ∞ [
```

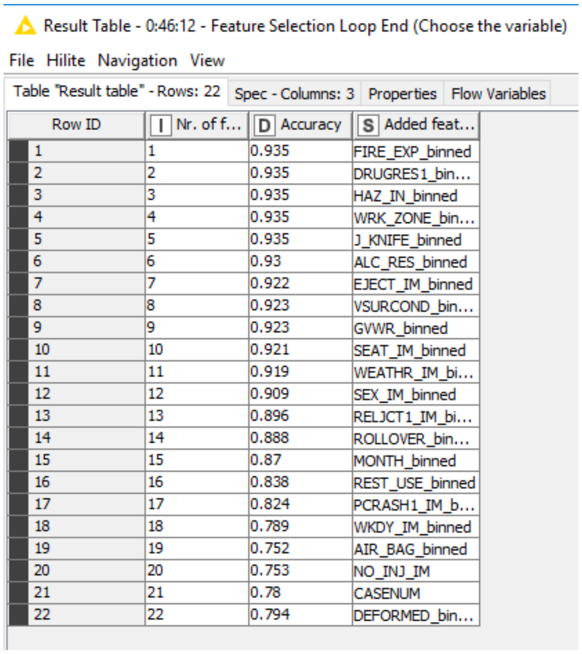
:] -∞ .. ∞ [

☑ Append new column ACC_TYPE_binned

After, performing these operations on our data, we have used **forward feature selection** technique as shown below: -



to further reduce our dimensions from 32 to 22 important features that greatly affect our target variable i.e. injury severity (INJSEV_IM). In this process we under sample the highest category in the population so that the training data doesn't have any biased influence towards the results while predicting the values using train data sets. We use decision tree for this purpose and then decision tree gives us the 22 most influential variables as shown below: -



Result Table - 0:46:12 - Feature Selection Loop End (Choose the variable)

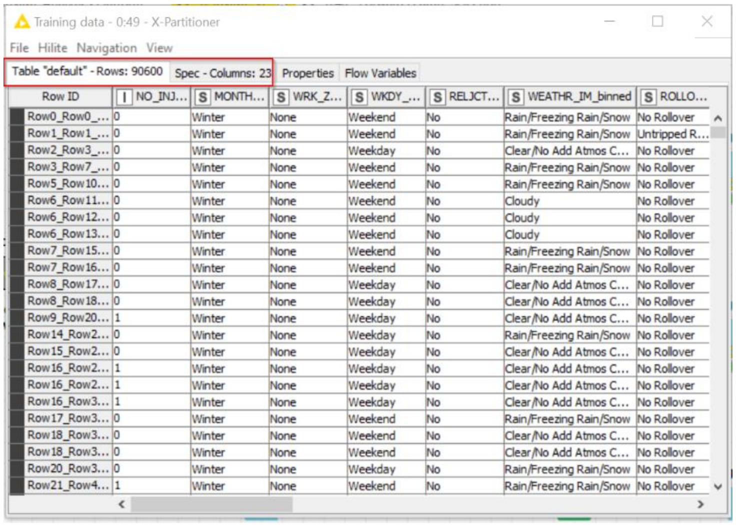
File Hilite Navigation View

Table "Result table" - Rows: 22 Spec - Columns: 3 Properties Flow Variables

Row ID	I Nr. of f...	D Accuracy	S Added feat...
1	1	0.935	FIRE_EXP_binned
2	2	0.935	DRUGRES1_bin...
3	3	0.935	HAZ_IN_binned
4	4	0.935	WRK_ZONE_bin...
5	5	0.935	J_KNIFE_binned
6	6	0.93	ALC_RES_binned
7	7	0.922	EJECT_IM_binned
8	8	0.923	VSURCOND_bin...
9	9	0.923	GVWR_binned
10	10	0.921	SEAT_IM_binned
11	11	0.919	WEATHR_IM_bi...
12	12	0.909	SEX_IM_binned
13	13	0.896	RELJCT1_IM_bi...
14	14	0.888	ROLLOVER_bin...
15	15	0.87	MONTH_binned
16	16	0.838	REST_USE_binned
17	17	0.824	PCRASH1_IM_b...
18	18	0.789	WKDY_IM_binned
19	19	0.752	AIR_BAG_binned
20	20	0.753	NO_INJ_IM
21	21	0.78	CASENUM
22	22	0.794	DEFORMED_bin...

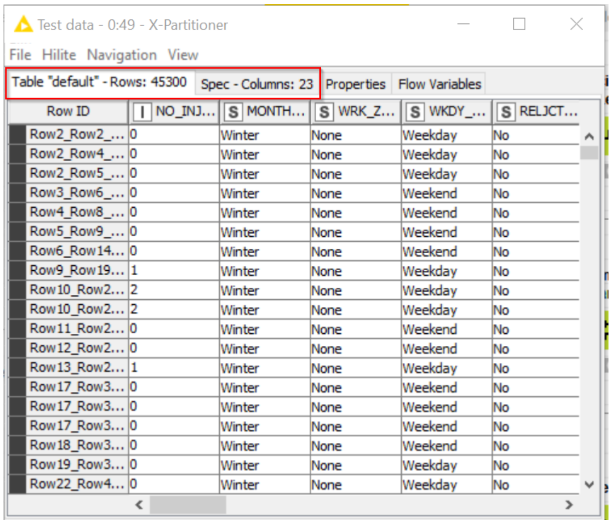
Partitioning of data:

After obtaining the total of 22 variables based on the forward feature selection, it is now required to partition the data into two sets before building the model. Total number of records in the data after feature selection is around 135900 and the ideal partitioning approach with a ratio of 70-30 has been adapted here, with 70% of the data as the training data (90600 records) and 30% of the data as the test data (45300).



Training data - 0:49 - X-Partitioner
File Hilite Navigation View
Table "default" - Rows: 90600
Spec - Columns: 23
Properties
Flow Variables

Row ID	I NO_INJ...	S MONTH...	S WRK_Z...	S WKDY_...	S RELJCT...	S WEATHR_IM_binned	S ROLLO...
Row0_Row0_...	0	Winter	None	Weekend	No	Rain/Freezing Rain/Snow	No Rollover
Row1_Row1_...	0	Winter	None	Weekend	No	Rain/Freezing Rain/Snow	Untripped R...
Row2_Row3_...	0	Winter	None	Weekday	No	Clear/No Add Atmos C...	No Rollover
Row3_Row7_...	0	Winter	None	Weekend	No	Rain/Freezing Rain/Snow	No Rollover
Row5_Row10...	0	Winter	None	Weekend	No	Rain/Freezing Rain/Snow	No Rollover
Row6_Row11...	0	Winter	None	Weekend	No	Cloudy	No Rollover
Row6_Row12...	0	Winter	None	Weekend	No	Cloudy	No Rollover
Row6_Row13...	0	Winter	None	Weekend	No	Cloudy	No Rollover
Row7_Row15...	0	Winter	None	Weekend	No	Rain/Freezing Rain/Snow	No Rollover
Row7_Row16...	0	Winter	None	Weekend	No	Rain/Freezing Rain/Snow	No Rollover
Row8_Row17...	0	Winter	None	Weekday	No	Clear/No Add Atmos C...	No Rollover
Row8_Row18...	0	Winter	None	Weekday	No	Clear/No Add Atmos C...	No Rollover
Row9_Row20...	1	Winter	None	Weekday	No	Clear/No Add Atmos C...	No Rollover
Row14_Row2...	0	Winter	None	Weekday	No	Rain/Freezing Rain/Snow	No Rollover
Row15_Row2...	0	Winter	None	Weekday	No	Clear/No Add Atmos C...	No Rollover
Row16_Row2...	1	Winter	None	Weekday	No	Clear/No Add Atmos C...	No Rollover
Row16_Row2...	1	Winter	None	Weekday	No	Clear/No Add Atmos C...	No Rollover
Row16_Row3...	1	Winter	None	Weekday	No	Clear/No Add Atmos C...	No Rollover
Row17_Row3...	0	Winter	None	Weekend	No	Rain/Freezing Rain/Snow	No Rollover
Row18_Row3...	0	Winter	None	Weekend	No	Clear/No Add Atmos C...	No Rollover
Row18_Row3...	0	Winter	None	Weekend	No	Clear/No Add Atmos C...	No Rollover
Row20_Row3...	0	Winter	None	Weekday	No	Rain/Freezing Rain/Snow	No Rollover
Row21_Row4...	1	Winter	None	Weekend	No	Rain/Freezing Rain/Snow	No Rollover



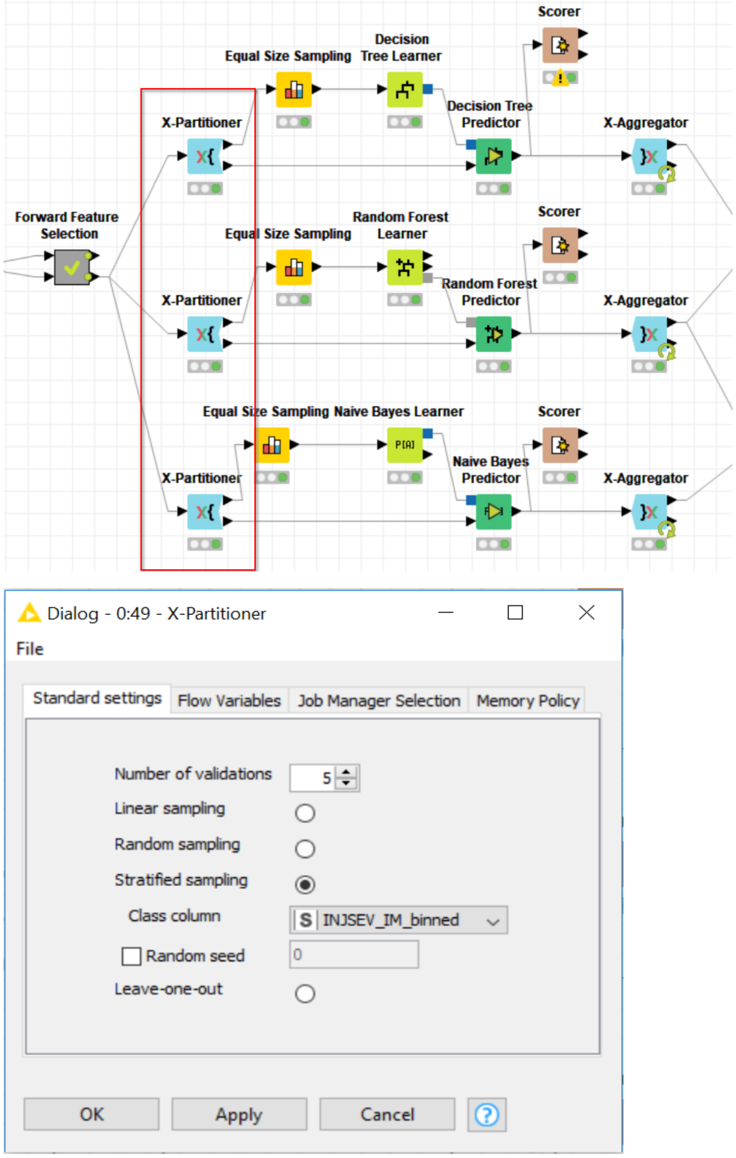
Test data - 0:49 - X-Partitioner
File Hilite Navigation View
Table "default" - Rows: 45300
Spec - Columns: 23
Properties
Flow Variables

Row ID	I NO_INJ...	S MONTH...	S WRK_Z...	S WKDY_...	S RELJCT...
Row2_Row2_...	0	Winter	None	Weekday	No
Row2_Row4_...	0	Winter	None	Weekday	No
Row2_Row5_...	0	Winter	None	Weekday	No
Row3_Row6_...	0	Winter	None	Weekend	No
Row4_Row8_...	0	Winter	None	Weekend	No
Row5_Row9_...	0	Winter	None	Weekend	No
Row6_Row14...	0	Winter	None	Weekend	No
Row9_Row19...	1	Winter	None	Weekday	No
Row10_Row2...	2	Winter	None	Weekday	No
Row10_Row2...	2	Winter	None	Weekday	No
Row11_Row2...	0	Winter	None	Weekend	No
Row12_Row2...	0	Winter	None	Weekend	No
Row13_Row2...	1	Winter	None	Weekday	No
Row17_Row3...	0	Winter	None	Weekend	No
Row17_Row3...	0	Winter	None	Weekend	No
Row17_Row3...	0	Winter	None	Weekend	No
Row18_Row3...	0	Winter	None	Weekend	No
Row19_Row3...	0	Winter	None	Weekday	No
Row22_Row4...	0	Winter	None	Weekday	No

Fig: Training data (on the left) and Test data (on the right)

Approach used for partitioning: X- Partitioning

X-Partitioning is a technique which uses an iterative approach and serves as the start of the cross validation loop where the number of cross validation iterations can be set as required. We have performed it 5 times here with stratified sampling.



Number of validations: 5

Linear sampling ○

Random sampling ○

Stratified sampling ●

Class column: S INJSEV_IM_binned

□ Random seed: 0

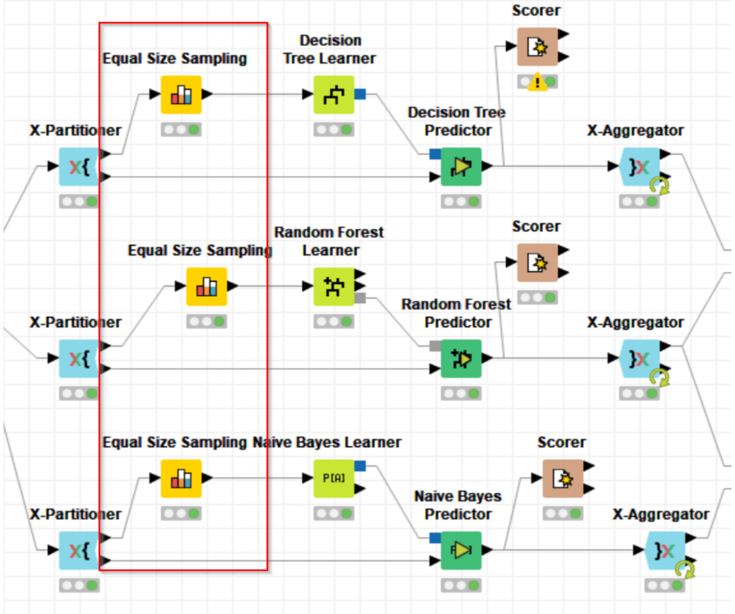
Leave-one-out ○

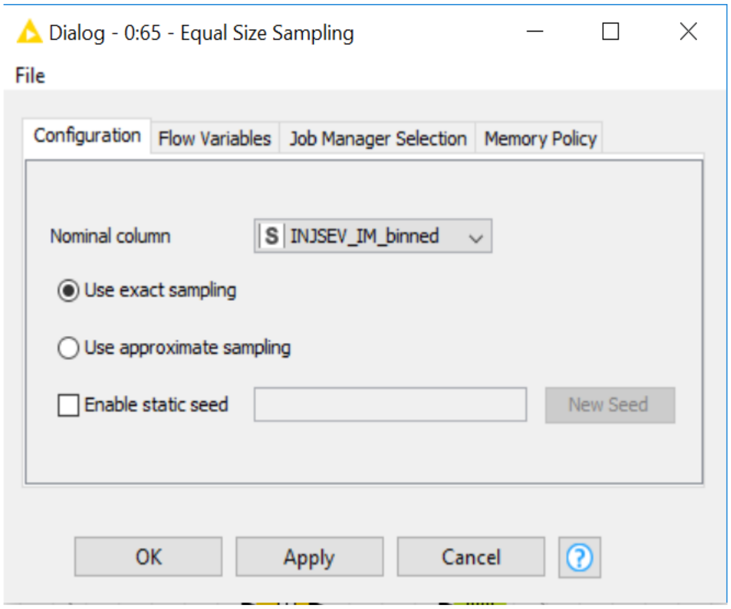
Fig: X- Partitioner node with number of iterations

Training data:

Observation of training data revealed that out of the total records of around 90K, only 6K records represented major injury, leaving around 93% of the data representing minor injury. This made the data highly biased and would eventually result in highly inaccurate model if ignored. Owing to the reason, we decided to perform under-sampling of the minor injury records.

For this purpose, we used Equal size sampling node, which takes into consideration equal number of records representing both the classes.




Dialog - 0:65 - Equal Size Sampling

Nominal column: S INJSEV_IM_binned

● Use exact sampling

○ Use approximate sampling

□ Enable static seed

MODEL BUILDING:

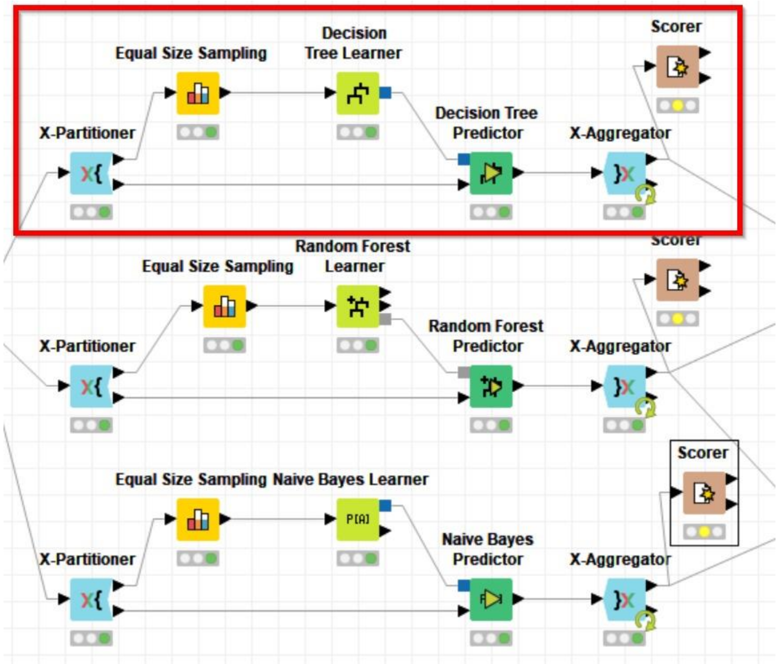
Out of several different machine learning models available, selection of model which fits the needs of the data in hand and which more accurately answers the business question is a very vital task. Not all the models perform best in every situation and owing to this reason we decided to build three different models and then evaluate to see which is performing better.

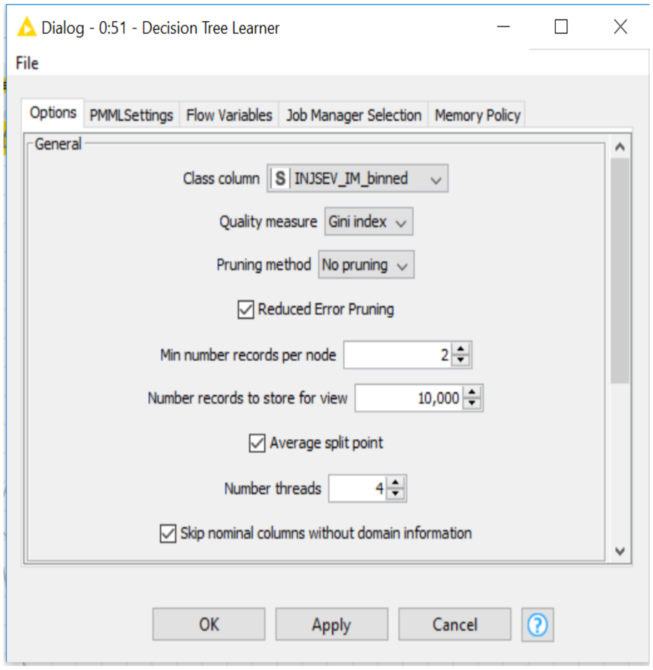
Following are the three models that we built:

1. Decision tree
2. Random Forest
3. Naïve Bayes

DECISION TREE MODEL:

- Since the target variable here (injury severity) is dichotomous in nature and we have around 22 variables, a combination of categorical and continuous variables, it is of our interest to identify that which variable is influencing the target variable the most or is most significant in determining the severity of injury, we decided to build a decision tree.
- The training data is given as the input to the decision tree learner to train the model. And based on that (output goes to the decision tree predictor), test data provided to the predictor node to predict the severity of the injury.
- Since we have used X-Partitioner node to start the loop, we also must use X-Aggregator node to finish that cross-validation iteration and to aggregate the output coming out from the predictor node.
- Also, once the predictions are made and the probability is generated, it is necessary to evaluate the model based on performance measures like accuracy, sensitivity and specificity, suing a Knime node known as scorer. We have explained it in the next step i.e. evaluation.




Dialog - 0:51 - Decision Tree Learner

File

General

Class column: S INJSEV_IM_binned

Quality measure: Gini index

Pruning method: No pruning

☑ Reduced Error Pruning

Min number records per node: 2

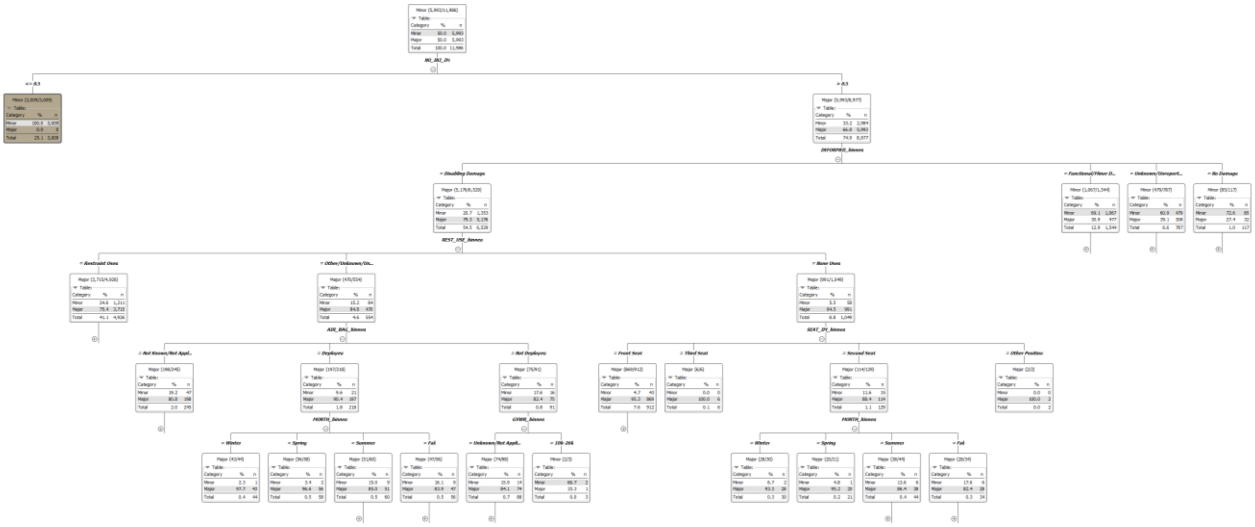
Number records to store for view: 10,000

☑ Average split point

Number threads: 4

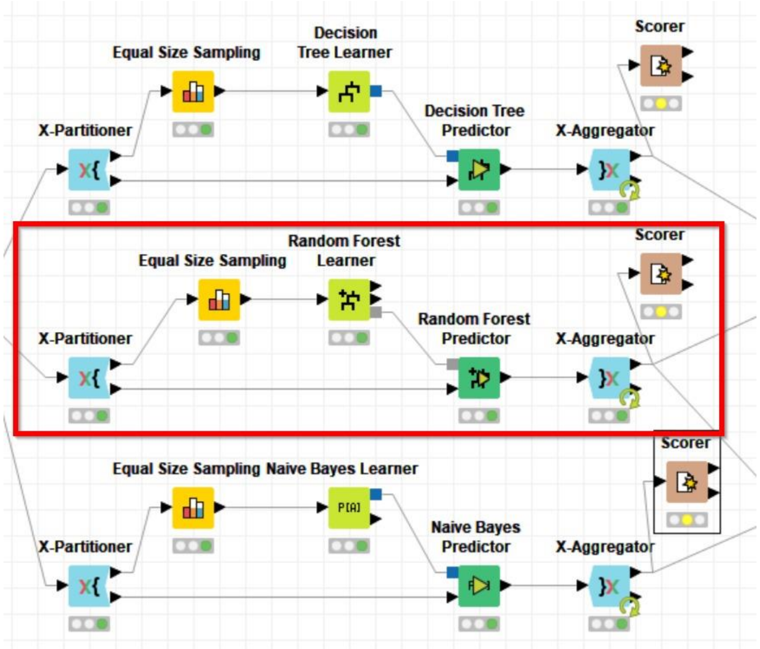
☑ Skip nominal columns without domain information

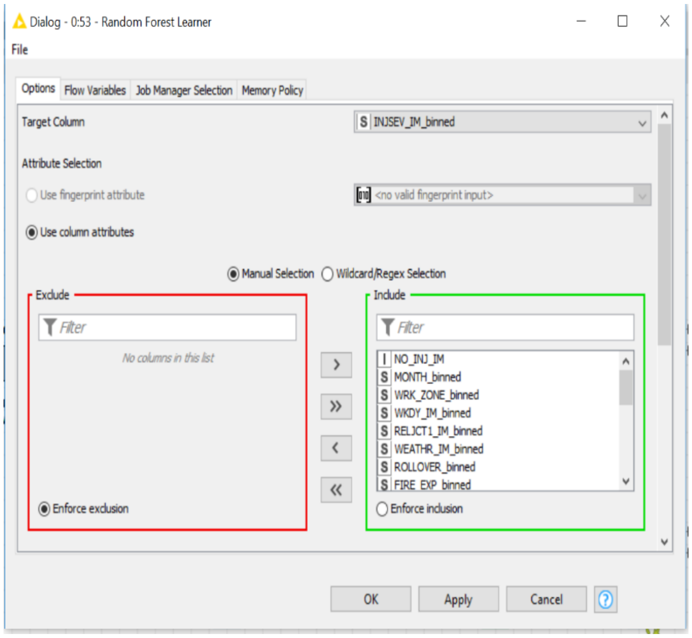
Here, from the below decision tree we can see that our variables like MONTH_binned, REST_USE_binned, DEFORMED_binned and SEAT_IM_binned are important factors in predicting the severity of the injury. Further on exploring the tree we can find VSURCON_binned, GVWR_binned are also very important factors which influence the severity of the injury.



RANDOM- FOREST MODEL:

- Yet another model that we built is Random forest which works on the principal of building several different decision trees for classification purpose and then coming up with a final class. This model is usually more stable than a decision tree and hence made us build it.
- Like decision tree model, we first used the X- Partitioner node and then the equal sized training and test data were used for making predictions.




Dialog - 0:53 - Random Forest Learner

Target Column: S INJSEV_IM_binned

Attribute Selection

○ Use fingerprint attribute <no valid fingerprint input>

● Use column attributes

● Manual Selection ○ Wildcard/Regex Selection

Exclude

Filter

No columns in this list

● Enforce exclusion

Include

Filter

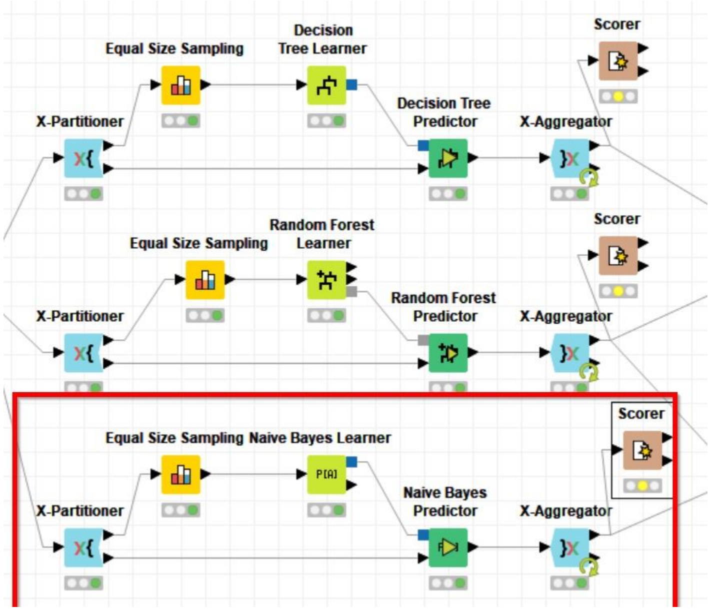
I NO_INJ_IM
S MONTH_binned
S WRK_ZONE_binned
S WKDY_IM_binned
S RELJCT1_IM_binned
S WEATHR_IM_binned
S ROLLOVER_binned
S FIRE_EXP_binned

○ Enforce inclusion

NAÏVE- BAYES MODEL:

- Naïve- Bayes classification works on Bayes theorem of probability, with a strong assumption that all the predictor variables are completely independent of one another, which is usually not possible in the real-life scenarios, however, we still intended to explore and see how this model performs in our problem.
- Like the two models we have built above, once again the training data obtained after X-Partitioning and is made equal sized for majority and minority class and used for training

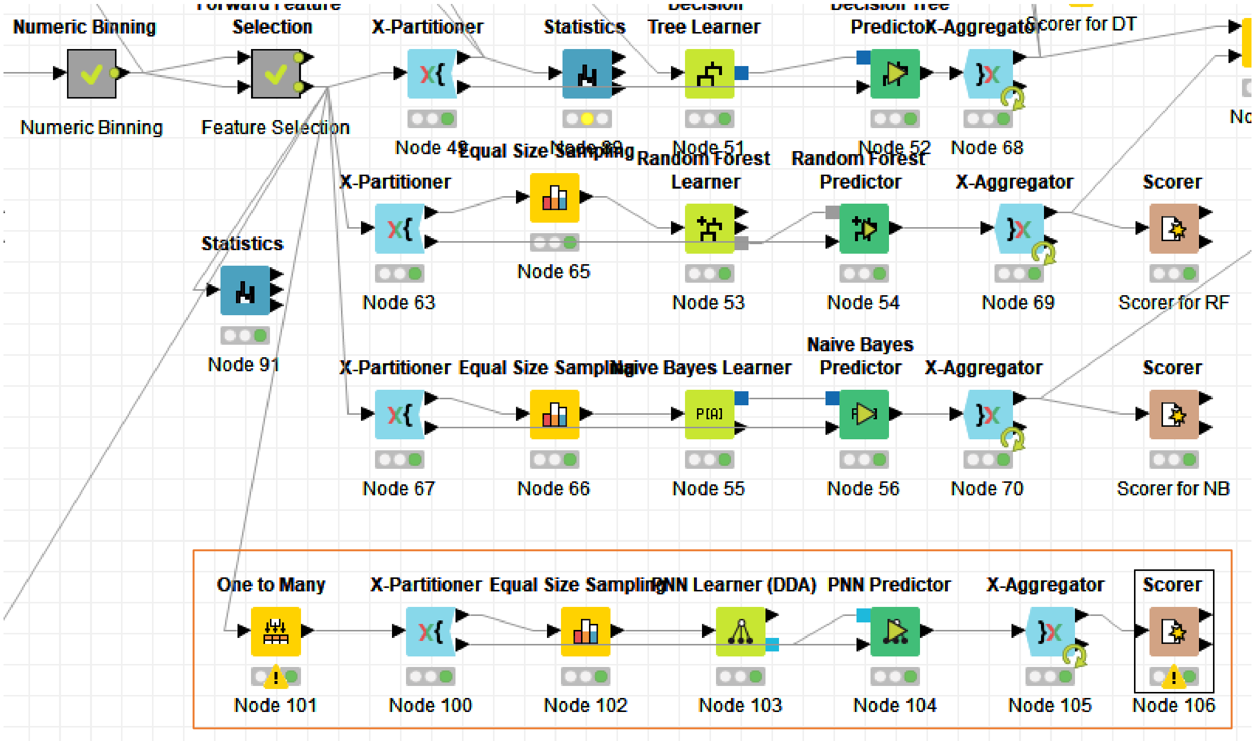
the model and the tested against a test data by making predictions for the severity of injury and then aggregated.



PNN MODEL:

Note: Included as per sir's recommendation during presentation

We have used neural network as it can be seen below. The neural network performs well with the numerical independent variables, so we had to create dummy variables for categories of more than 2.


Numeric Binning
Forward Feature Selection
X-Partitioner
Statistics
Decision Tree Learner
Decision Tree Predictor
X-Aggregator
Scorer for DT
Numeric Binning
Feature Selection
Node 49
Node 89
Node 51
Node 52
Node 68
Equal Size Sampling
Random Forest Learner
Random Forest Predictor
X-Partitioner
X-Aggregator
Scorer
Statistics
Node 65
Node 63
Node 53
Node 54
Node 69
Scorer for RF
Node 91
Naive Bayes Predictor
X-Partitioner
Equal Size Sampling
Naive Bayes Learner
X-Aggregator
Scorer
Node 67
Node 66
Node 55
Node 56
Node 70
Scorer for NB
One to Many
X-Partitioner
Equal Size Sampling
PNN Learner (DDA)
PNN Predictor
X-Aggregator
Scorer
Node 101
Node 100
Node 102
Node 103
Node 104
Node 105
Node 106

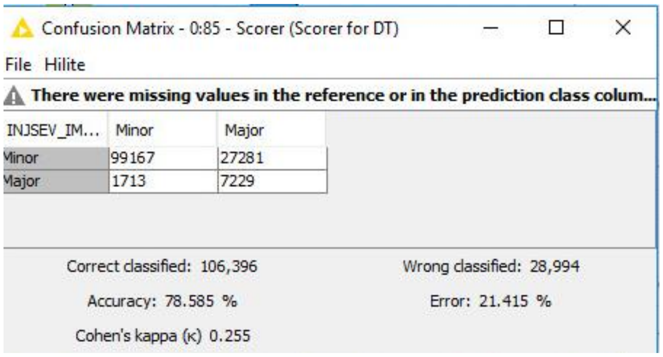
EVALUATION

ACCURACY

Once we obtained the predictions and probabilities from all the models, it is now required to evaluate all the models based on some measurable factors like accuracy, sensitivity, specificity, and ROC curves.

Following is an elaborated comparison of model performances:

1. Decision Tree Model:



Confusion Matrix - 0:85 - Scorer (Scorer for DT)

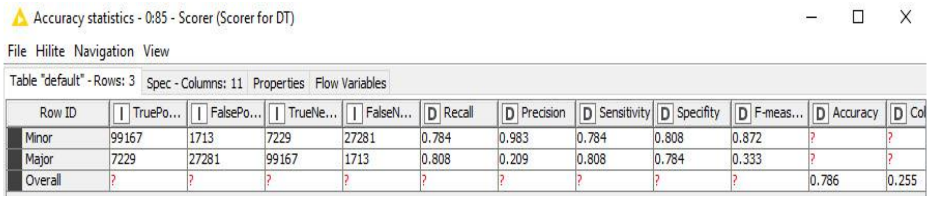
There were missing values in the reference or in the prediction class colum...

INJSEV_IM...	Minor	Major
Minor	99167	27281
Major	1713	7229

Correct classified: 106,396 Wrong classified: 28,994

Accuracy: 78.585 % Error: 21.415 %

Cohen's kappa (κ) 0.255

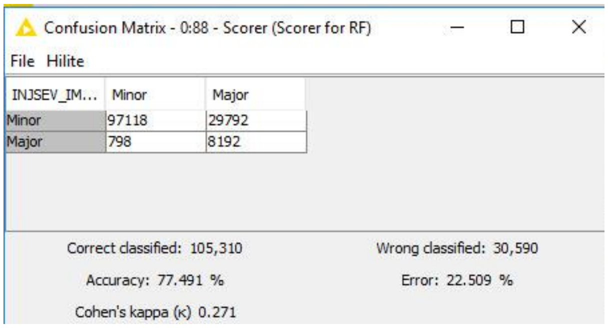


Accuracy statistics - 0:85 - Scorer (Scorer for DT)

Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specifty	D F-meas...	D Accuracy	D Col
Minor	99167	1713	7229	27281	0.784	0.983	0.784	0.808	0.872	?	?
Major	7229	27281	99167	1713	0.808	0.209	0.808	0.784	0.333	?	?
Overall	?	?	?	?	?	?	?	?	?	0.786	0.255

- An accuracy of 78.585% is obtained for this model. The sensitivity (for major injury), i.e. when the major injury is predicted as a major one is 80.8%.

2. Random Forest Model:



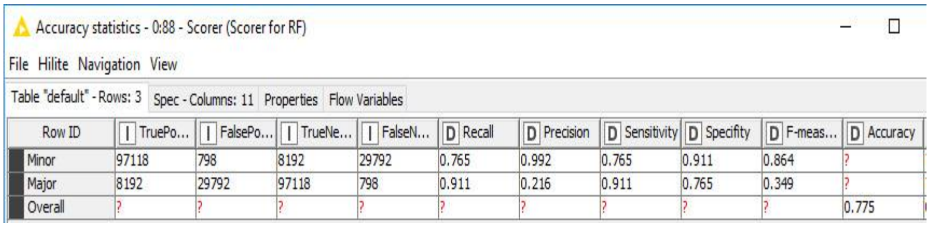
Confusion Matrix - 0:88 - Scorer (Scorer for RF)

INJSEV_IM...	Minor	Major
Minor	97118	29792
Major	798	8192

Correct classified: 105,310 Wrong classified: 30,590

Accuracy: 77.491 % Error: 22.509 %

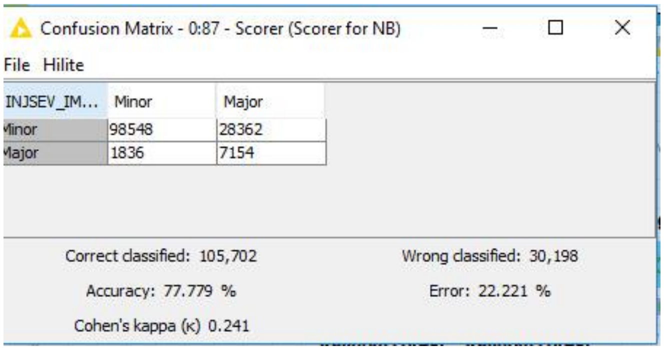
Cohen's kappa (κ) 0.271


Accuracy statistics - 0:88 - Scorer (Scorer for RF)
File Hilite Navigation View
Table "default" - Rows: 3
Spec - Columns: 11
Properties
Flow Variables

Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specifity	D F-meas...	D Accuracy
Minor	97118	798	8192	29792	0.765	0.992	0.765	0.911	0.864	?
Major	8192	29792	97118	798	0.911	0.216	0.911	0.765	0.349	?
Overall	?	?	?	?	?	?	?	?	?	0.775

- The accuracy of this model is obtained to be 77.491% with a sensitivity of 91.1%, which is quite significant.

3. Naïve- Bayes Model:

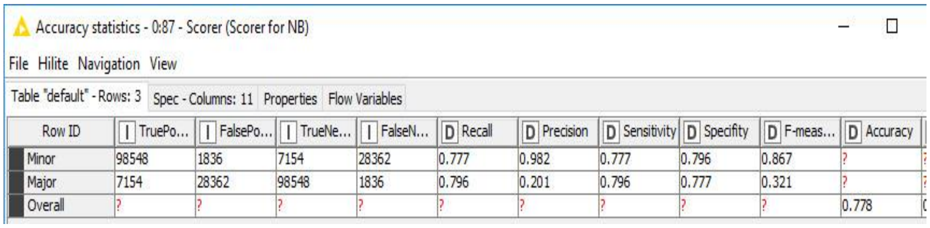


INJSEV_IM...	Minor	Major
Minor	98548	28362
Major	1836	7154

Correct classified: 105,702 Wrong classified: 30,198

Accuracy: 77.779 % Error: 22.221 %

Cohen's kappa (κ) 0.241

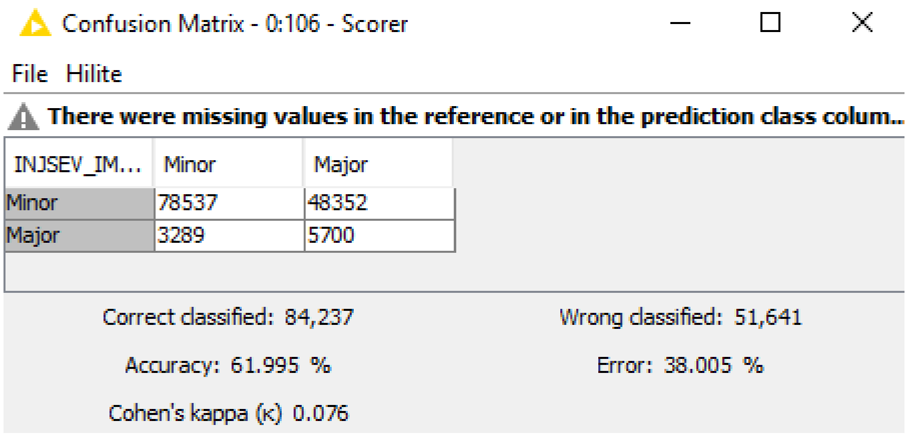

Accuracy statistics - 0:87 - Scorer (Scorer for NB)
File Hilite Navigation View
Table "default" - Rows: 3
Spec - Columns: 11
Properties
Flow Variables

Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specifity	D F-meas...	D Accuracy
Minor	98548	1836	7154	28362	0.777	0.982	0.777	0.796	0.867	?
Major	7154	28362	98548	1836	0.796	0.201	0.796	0.777	0.321	?
Overall	?	?	?	?	?	?	?	?	?	0.778

- Accuracy of 77.779 and a sensitivity of 79.6% is observed for Naïve- Bayes model.

4.PNN Model:

We can see below that neural network created has minimum accuracy of approximately 62% which is very less comparatively to other models we have created earlier like decision tree, random forest and naïve Bayes.



Confusion Matrix - 0:106 - Scorer

There were missing values in the reference or in the prediction class colum...

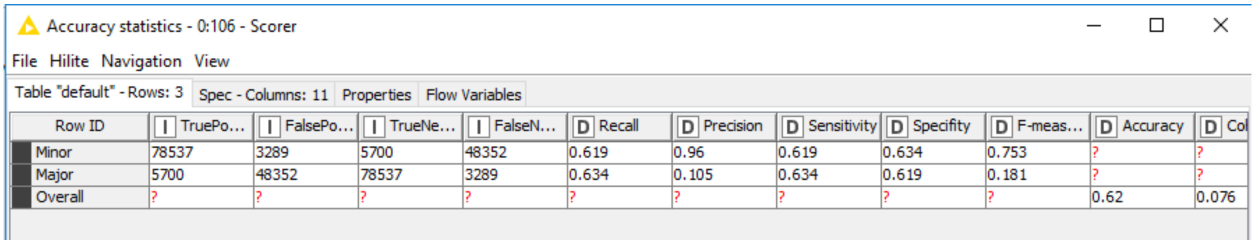
INJSEV_IM...	Minor	Major
Minor	78537	48352
Major	3289	5700

Correct classified: 84,237 Wrong classified: 51,641

Accuracy: 61.995 % Error: 38.005 %

Cohen's kappa (κ) 0.076

Also, the specificity (63.4%) and the sensitivity (61.9%) for neural network is very less as seen below: -

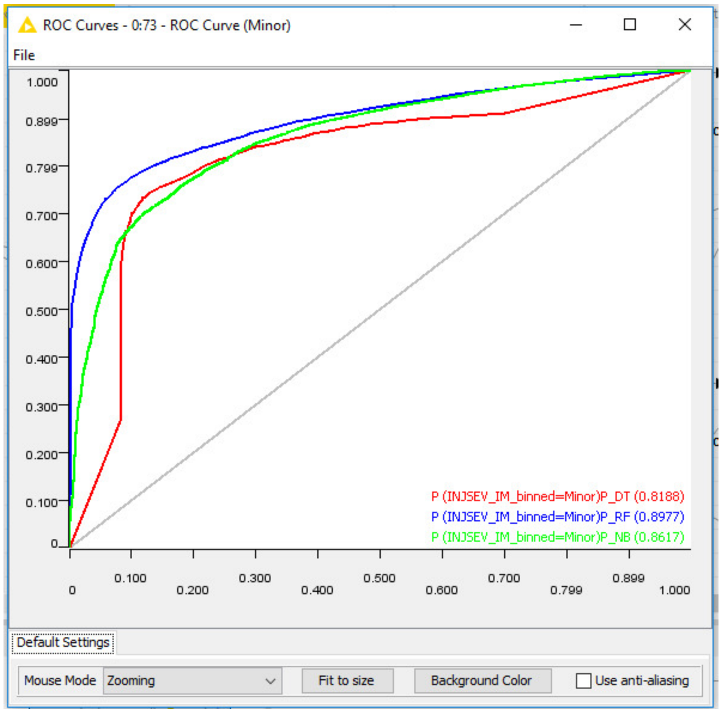


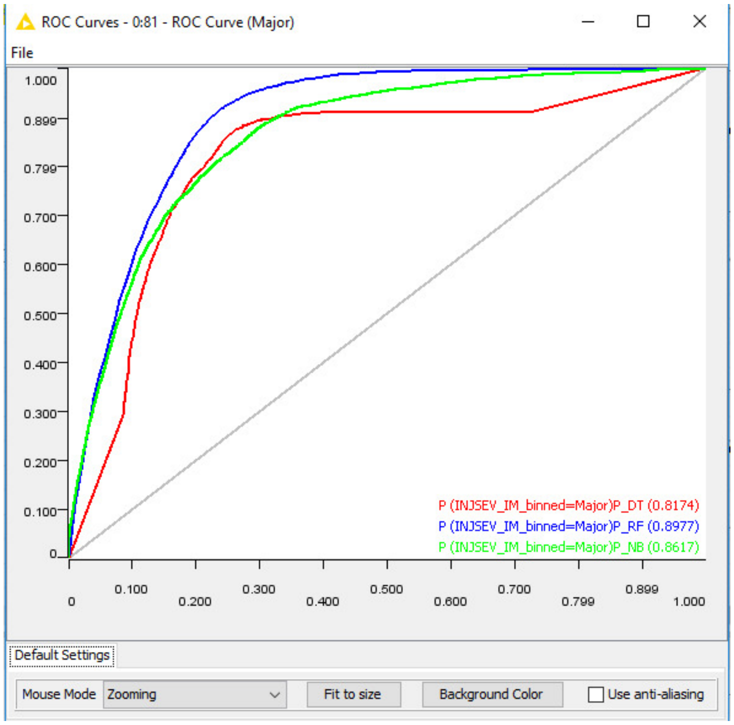
Accuracy statistics - 0:106 - Scorer

Row ID	I TruePo...	I FalsePo...	I TrueNe...	I FalseN...	D Recall	D Precision	D Sensitivity	D Specifity	D F-meas...	D Accuracy	D Col
Minor	78537	3289	5700	48352	0.619	0.96	0.619	0.634	0.753	?	?
Major	5700	48352	78537	3289	0.634	0.105	0.634	0.619	0.181	?	?
Overall	?	?	?	?	?	?	?	?	?	0.62	0.076

ROC CURVES

Since it is not an ideal practice to base the fitness of a model, entirely on the accuracy. We have plotted two ROC curves with the probability values for Major injury as well as for the Minor injury.





- For prediction of Minor injury, we see that the area under the curve is maximum for Random Forest with 0.8977, followed by Naïve- Bayes and then decision tree.
- Similarly, for the prediction of Major injury as well, the area under the curve is maximum for the Random Forest model, with 0.8977.

CONCLUSION

The objective of the study was to build predictive data mining models that could determine various factors impacting injury severity. In this research, we have proposed a framework for the analysis of accident patterns during different events.

To summarize the important factors that affect Injury severity, we can look at the predictive model we have built. Decision Tree and Random forest have both shown that DEFORMED, MONTH and SEAT_IM are all important variables. The common ground between both these models was that vehicles that were more deformed and during Fall were more likely to result in fatal injury.

According to a report by NHTSA, during Fall, due to dusk and dawn timings change, drivers get confused with the lighting and most accidents happen at the start of Fall. The models predict that despite the restraint used, people get ejected from their seat during the crash. SEAT_IM indicating the position of the seat is an important factor as it obviously relates to injury directly. When the crash occurs, people in the front seat are more likely to get injured as the front part of the car is directly exposed to crash.

The classification accuracy on the test results reveals that Decision tree performs better than Random Forest, Naïve Bayes and Neural Network with 78.6%. From the confusion matrix, Decision tree has less amount of error rate compared to the other two models. On the contrary, when an ROC curve is plotted, Random Forest has more area under the curve (89.7%) for both the categories INJSEV_IM_binned = Minor and Major.

With all the accuracy metrics taken into consideration, we have concluded that Random Forest models would yield better predictions when all the cases are taken into consideration.

RECOMMENDATIONS

The main recommendations from our study would depend on the following factors:

DEFORMED: This element describing the extent to which the vehicle was damaged during the crash directly impacts the severity of injury.

Hence, it is highly recommended that car manufacturers find out innovative ways of producing stronger metals that would make the outer part of the car. This is the most important factor that would aid the passengers during a crash.

EJECT_IM_binned: The element describes the angle at which the passenger ejects from the seat. Irrespective of the angle, it is highly important for the restraint system like seat belt to function properly. With advancement in technology, it would be advisable to have safer and more functional restrain systems used inside the vehicle.

MONTH_binned: Fall and Summer are observed to have high accidents events. Since, the light conditions change during the season transition from Fall to Summer, government needs to take precautionary steps like having more street lights used at high accident-prone areas.

SEAT_IM: Irrespective of the seat position, it is necessary to have stronger materials used for outer part of the car.

Citations:

1. National Pedestrian Crash Report by National Highway Traffic Safety Administration (NHTSA): https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/810968