

**Data mining approach to reduce accidents**

MSIS 5633 – Business Intelligence Tools & Techniques



Ankur Bansal

Avinash Kaur

Tanmayee Nettem

Contents

[EXECUTIVE SUMMARY 1](#_Toc532699254)

[CRISP-DM 2](#_Toc532699255)

[BUSINESS UNDERSTANDING 2](#_Toc532699256)

[DATA UNDERSTANDING 3](#_Toc532699257)

[DATA PREPROCESSING 5](#_Toc532699258)

[MODEL BUILDING: 16](#_Toc532699259)

[Decision Tree Model: 16](#_Toc532699260)

[Random- Forest Model: 18](#_Toc532699261)

[Naïve- Bayes Model: 19](#_Toc532699262)

[PNN Model: 20](#_Toc532699263)

[EVALUATION 22](#_Toc532699264)

[Accuracy 22](#_Toc532699265)

[ROC Curves 24](#_Toc532699266)

[CONCLUSION 26](#_Toc532699267)

[RECOMMENDATIONS 27](#_Toc532699268)

# 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.

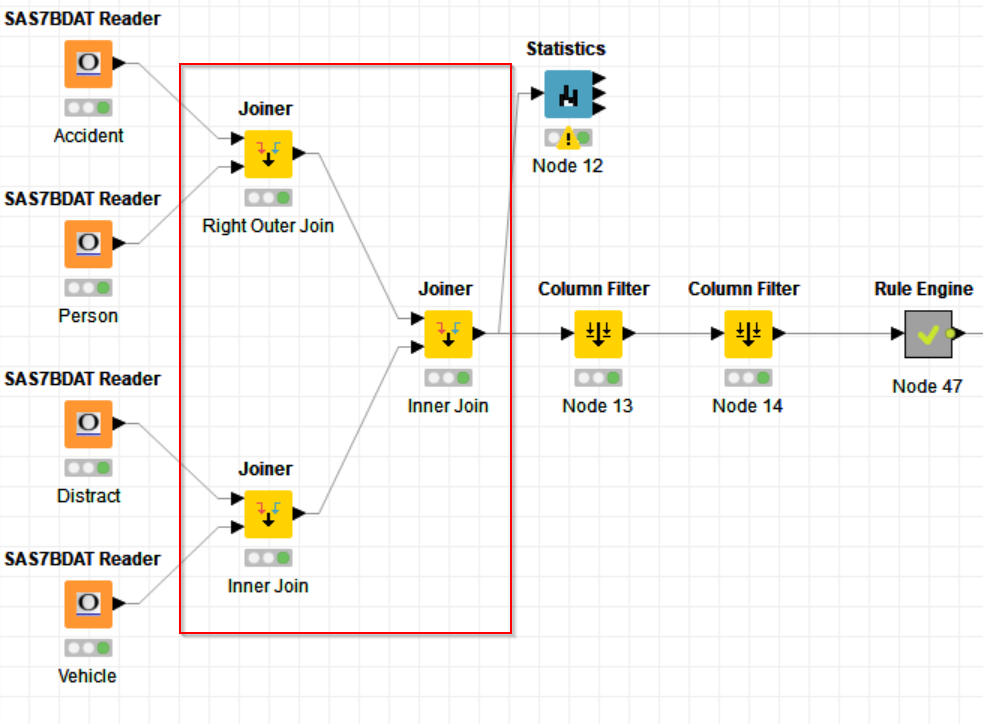
ALC\_RES: Records the result of the alcohol test performed on the driver.

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

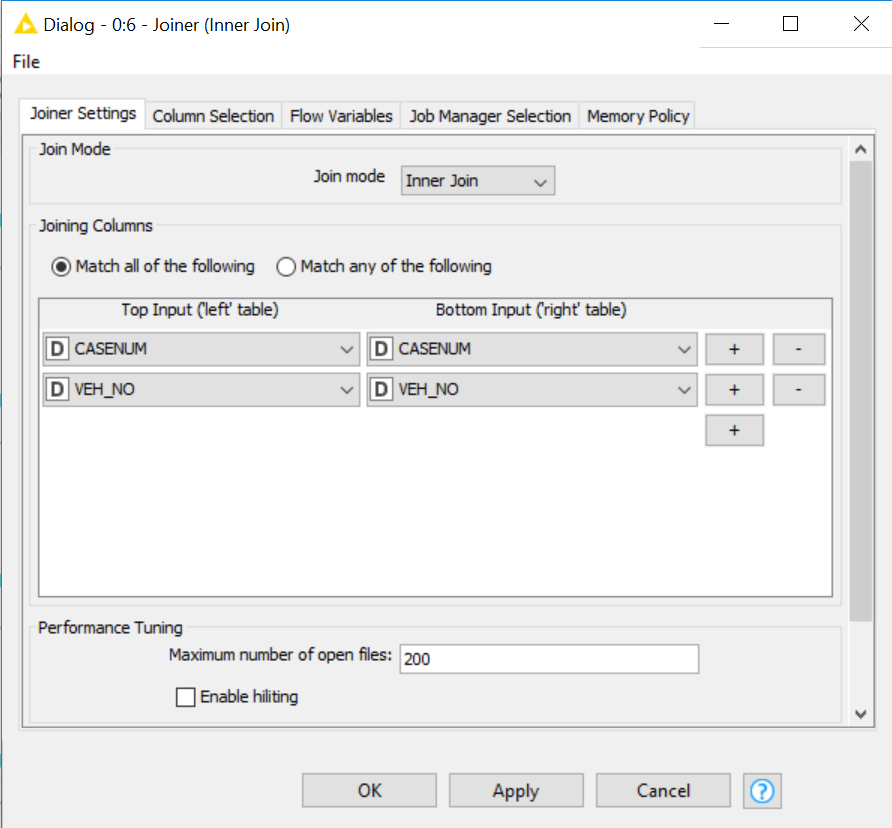
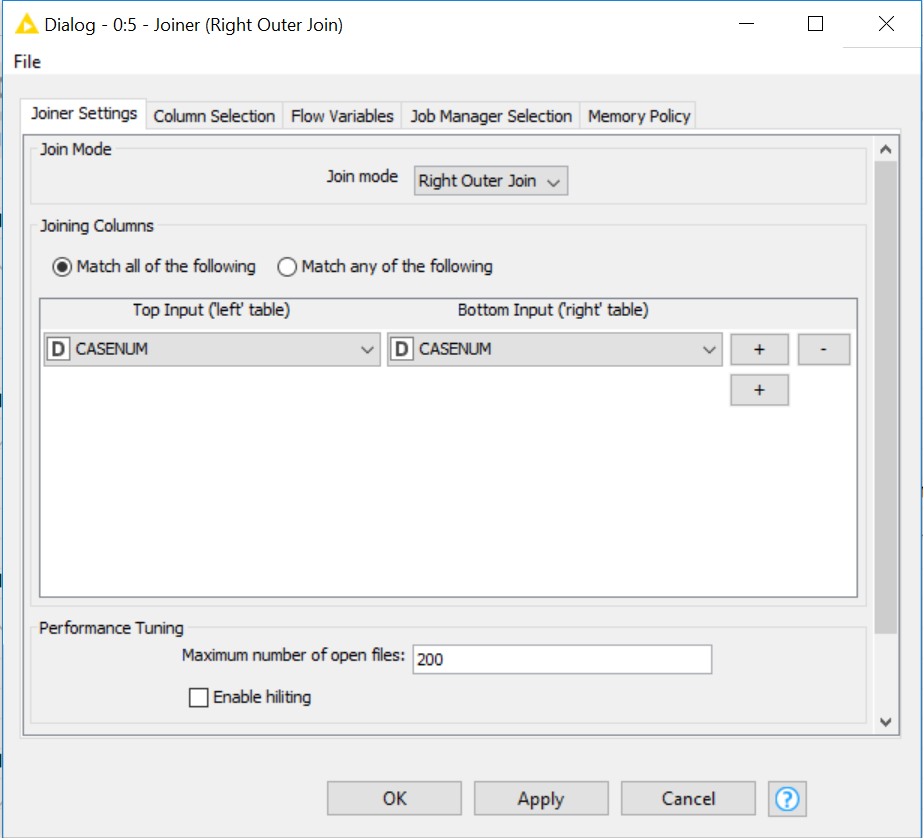
## DATA PREPROCESSING

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

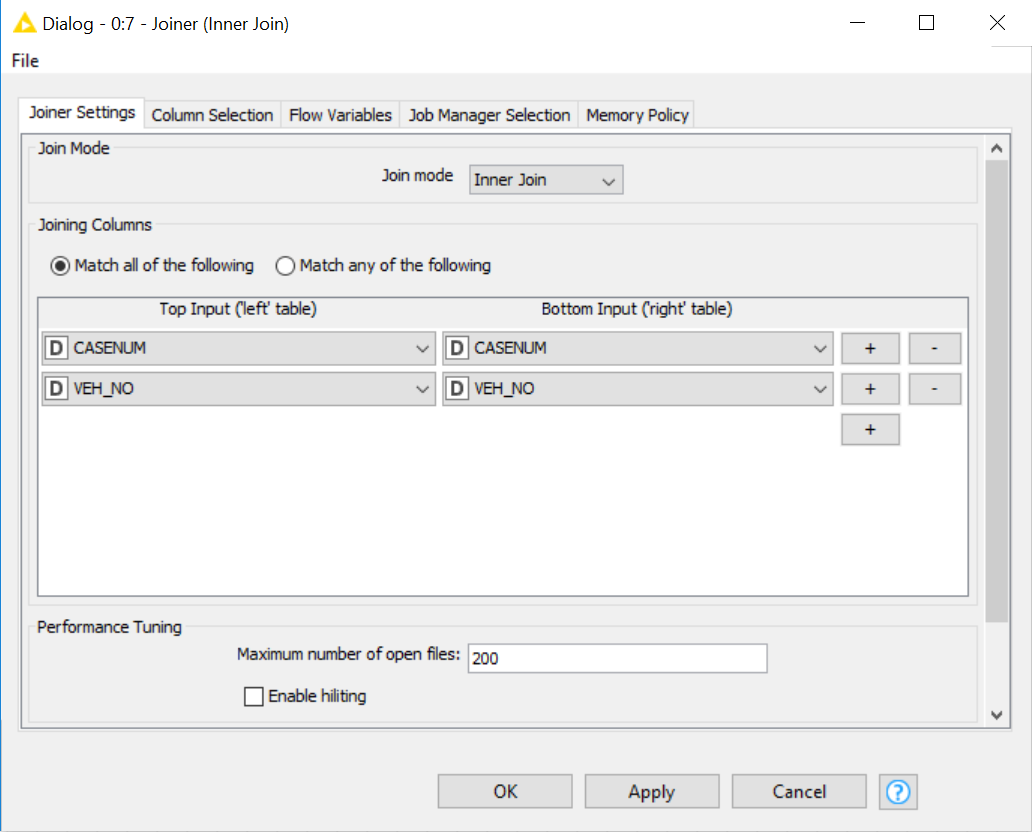
1. Data integration: - Data Integration is the combination of technical and business processes used to combine data from different multiple sources into meaningful and valuable information to get reliable data source. Here, in our dataset we have combined data from 4 tables namely accident, Person, Distract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

****

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

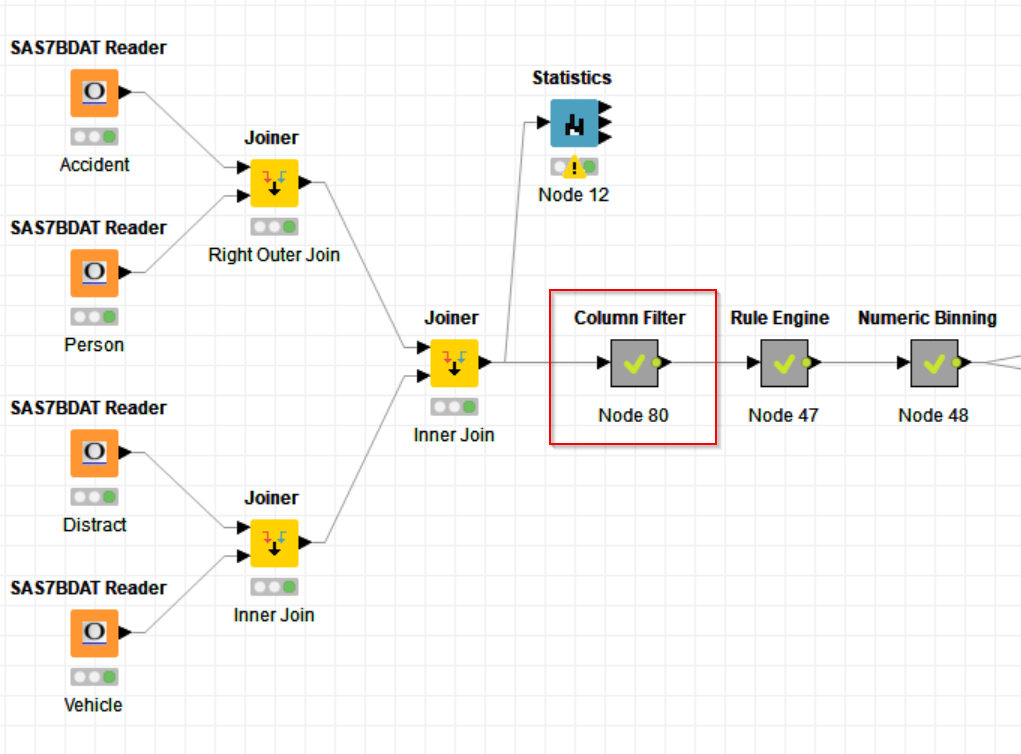
****

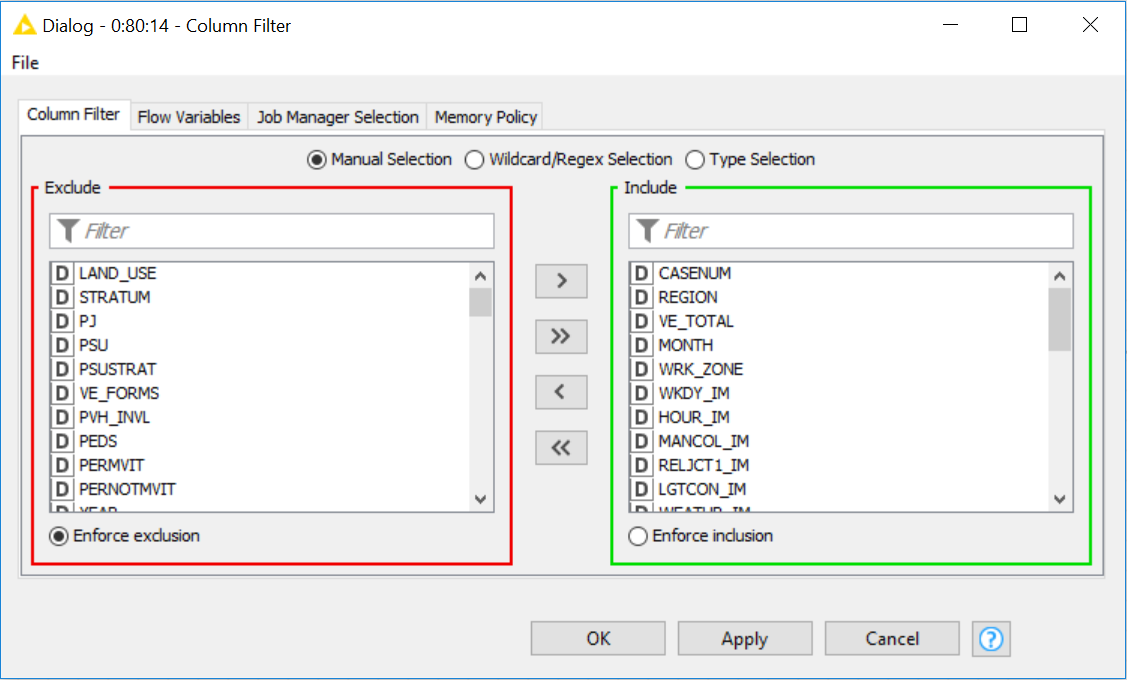
1. **Data Cleaning: -**

**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.

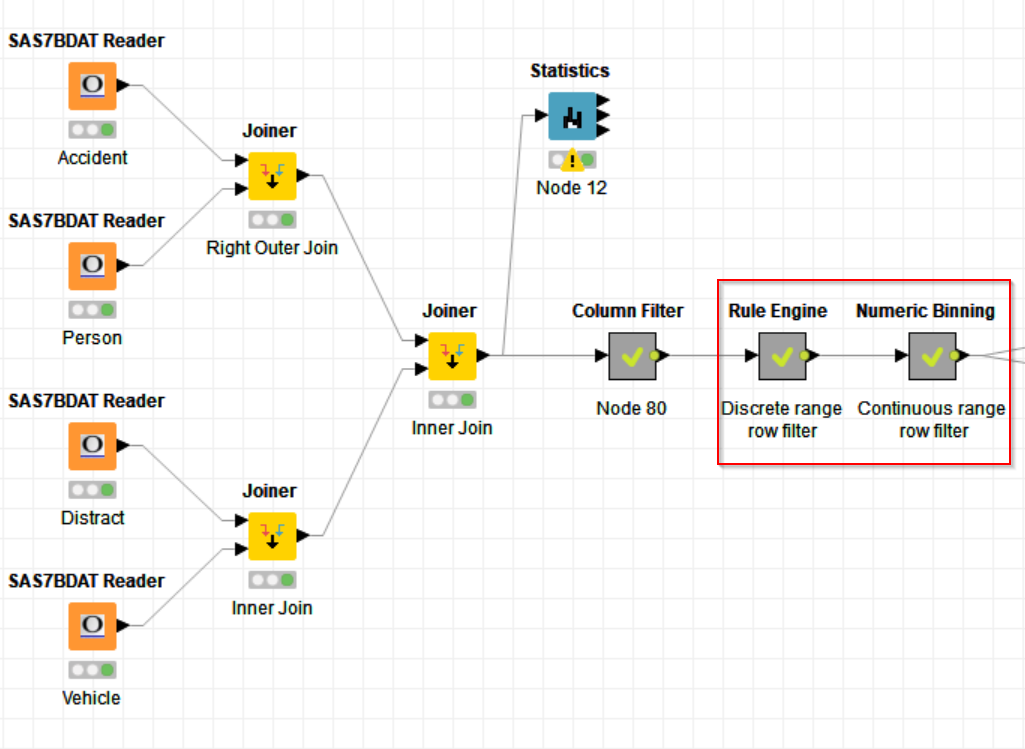
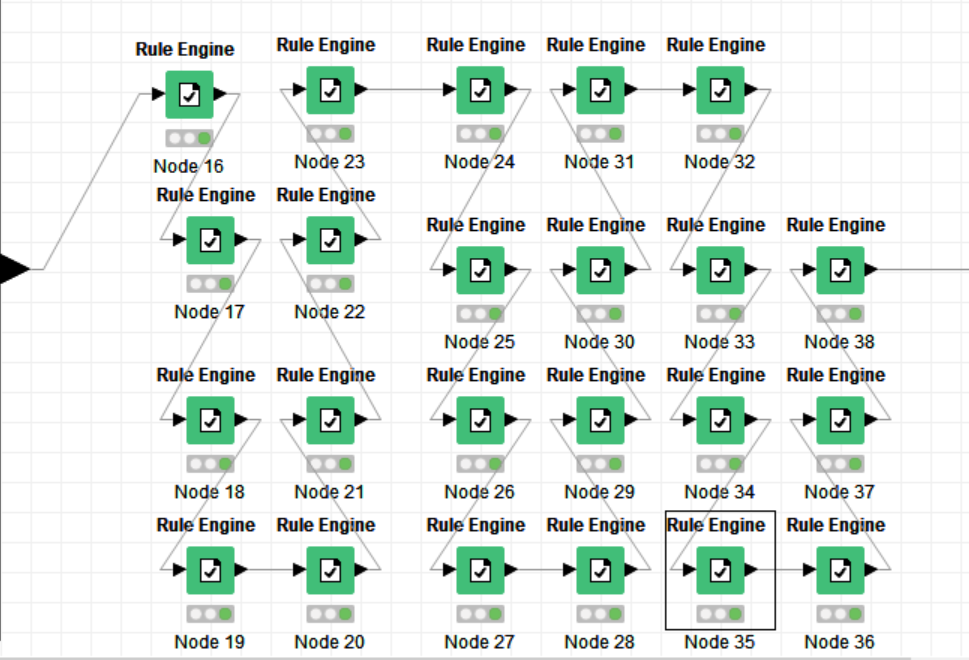
1. **Dimension Reduction: -**

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.

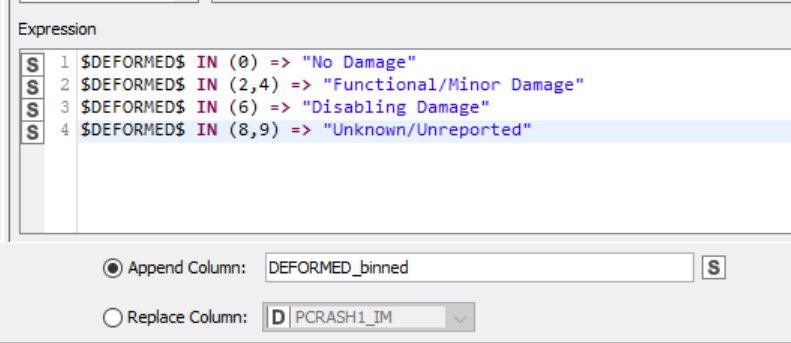


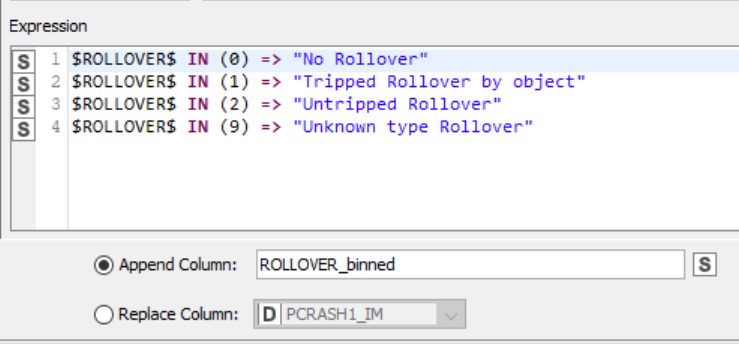
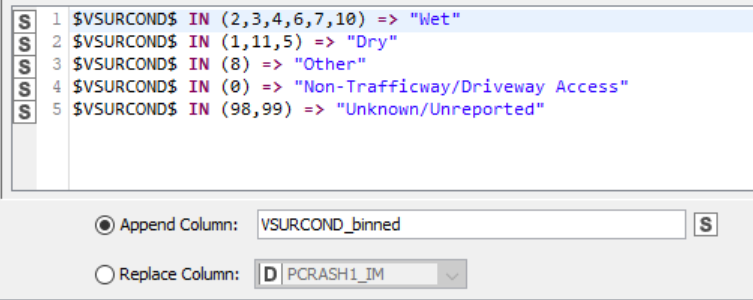


After this we will use Rule Engine to categorize several categories for a variable which are similar and, we have created a separate category for a non-reported or unknown value in a variable, thus avoiding any further data loss in our dataset. We have used **22** Rule Engines to categorize the variables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

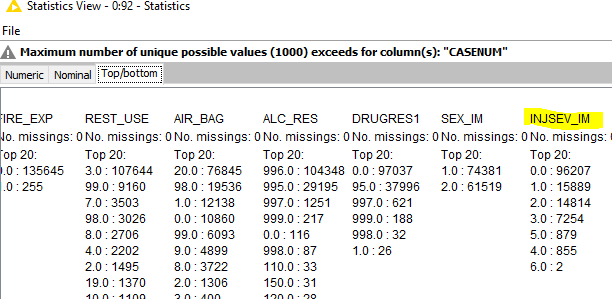
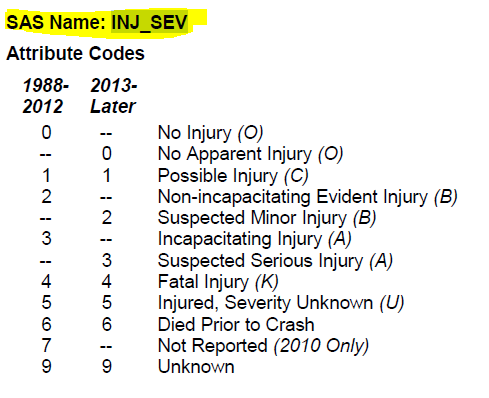
 

We have the following example for the same, Here, we have created rules in Rule Engine for DEFORMED\_binned, MONTH\_binned, ROLLOVER\_binned and VSURCOND\_binned into their respective possible minimum categories.

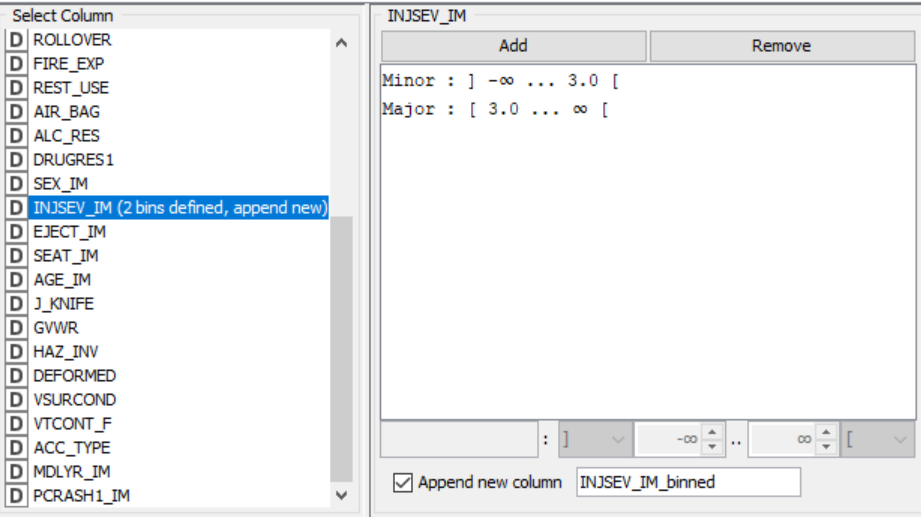
** **

** **

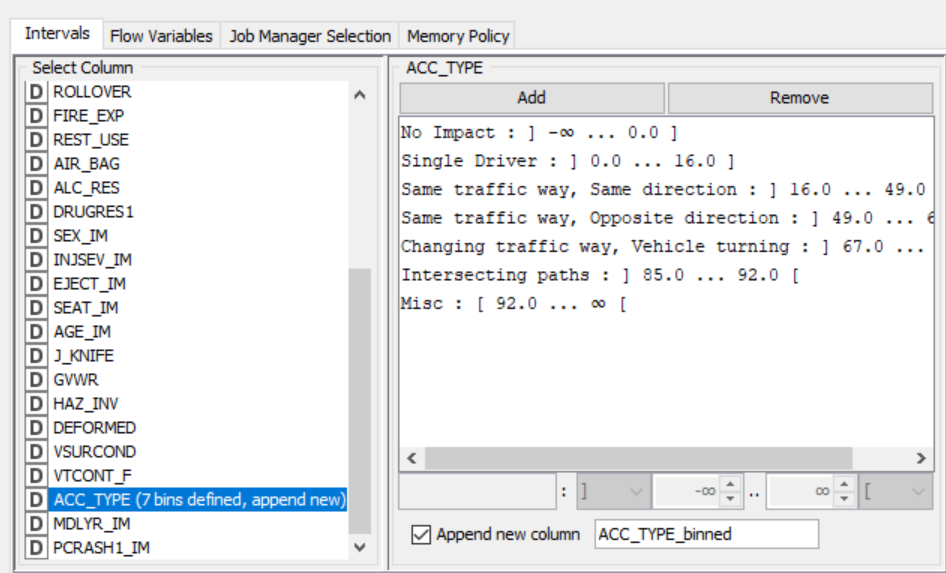
Also, apart from this we have used Numeric Binning to categorize and convert our continuous numerical columns to string. Overall, we have used 5 Numeric Binning nodes for our dataset. For example: - We have combined the minimum injury severity and maximum injury severity for 9 different categories of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

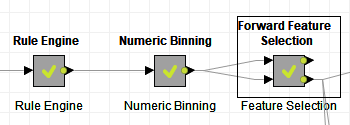
We can see the below Numeric Binner for the same variable:-



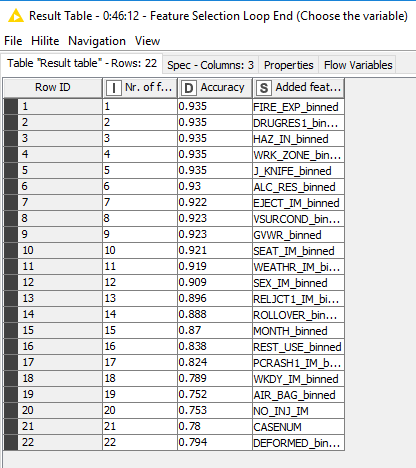
And similarly, we have converted other continuous numeric variables into string. Here, we have converted Accident Type (ACC\_TYPE): -



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: -



**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).

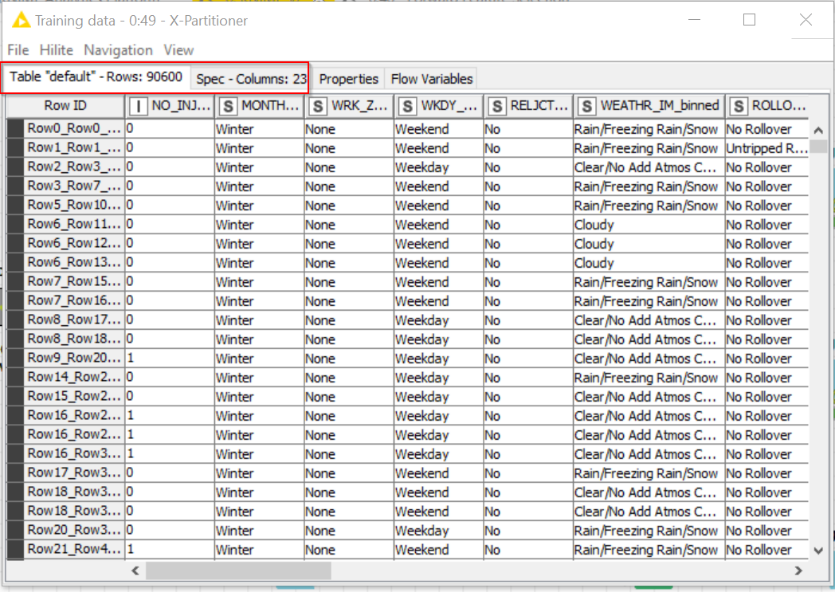


Fig: Training data (on the left) and Test data (on the right)

**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.

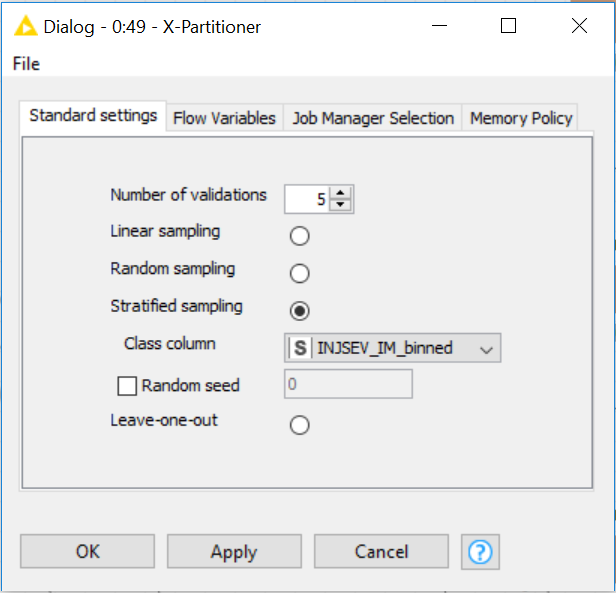
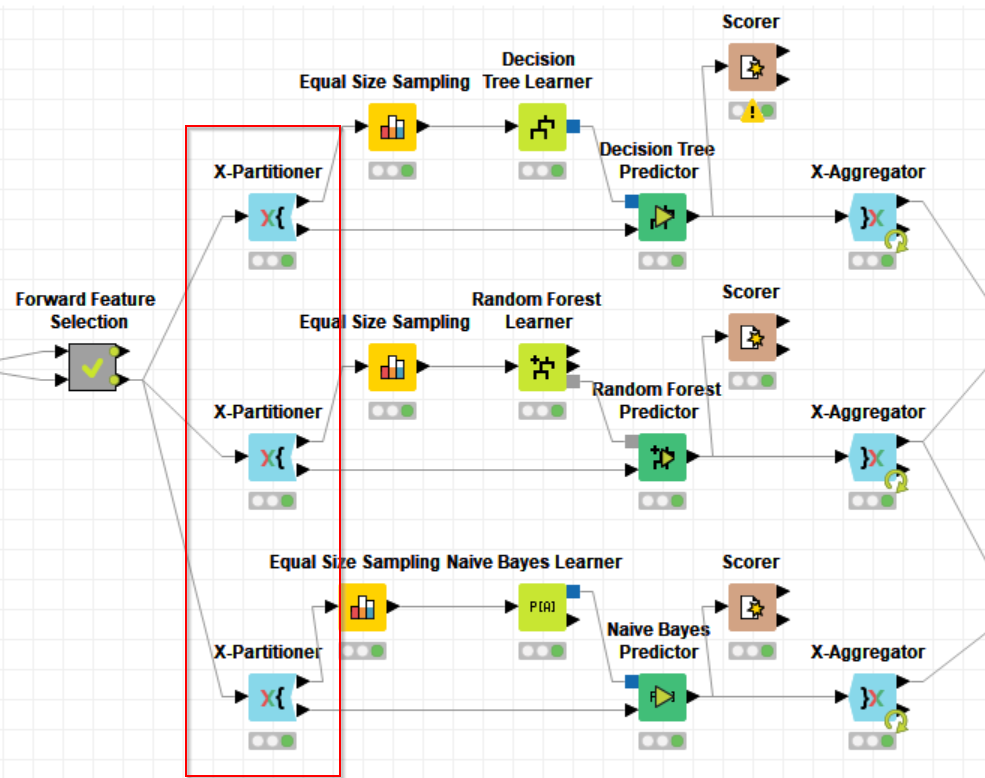
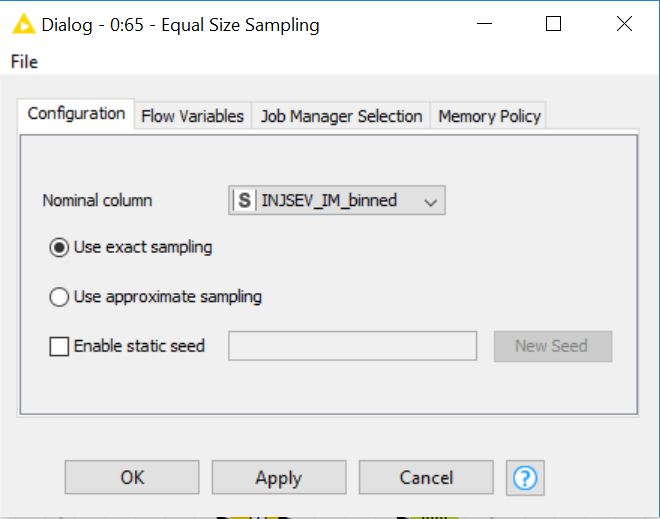
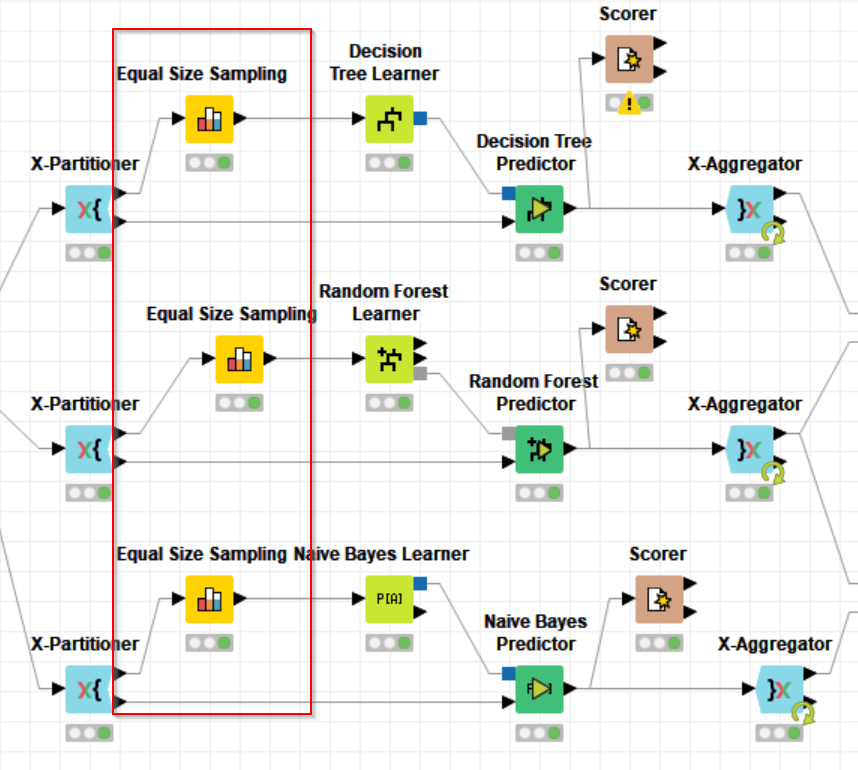


Fig: X- Partitioner node with number of iterations

**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.



## MODEL BUILDING:

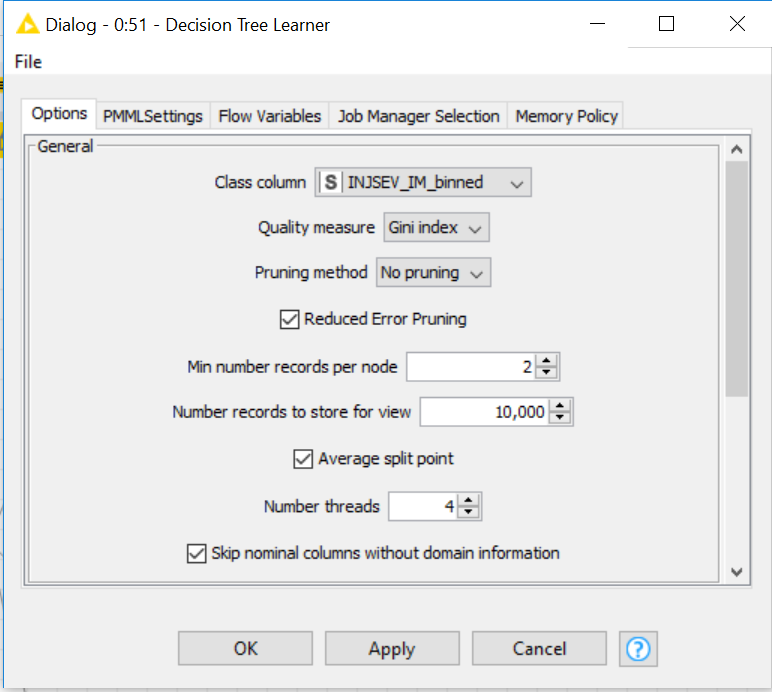
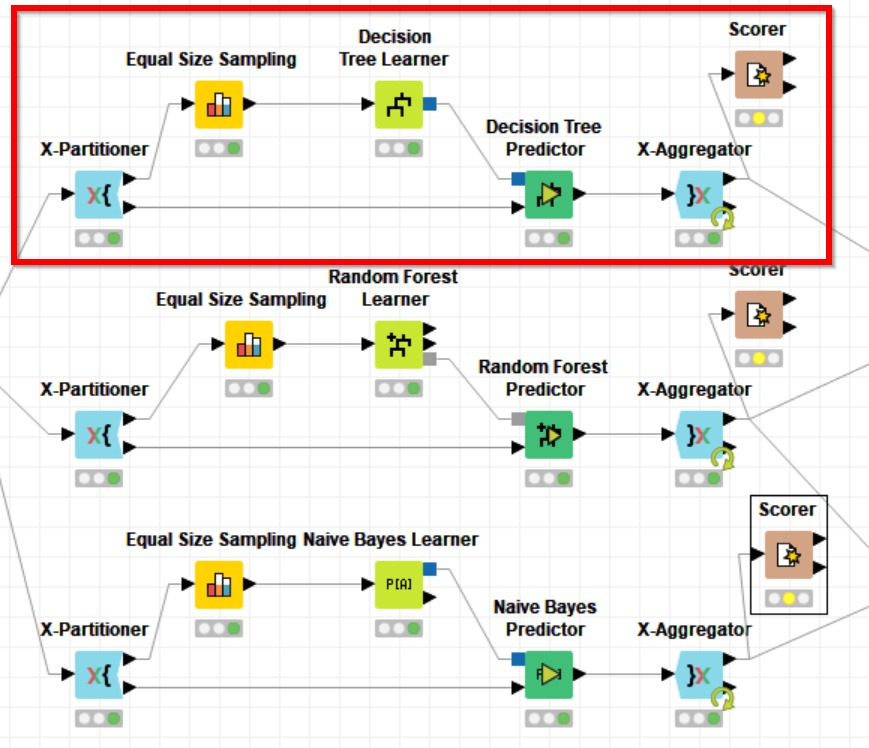
Out of several different machine learning models available, selection of model which fits the needs of the data in hand and which more accurately answers the business question is a very vital task. Not all the models perform best in every situation and owing to this reason we decided to build three different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

Following are the three models that we built:

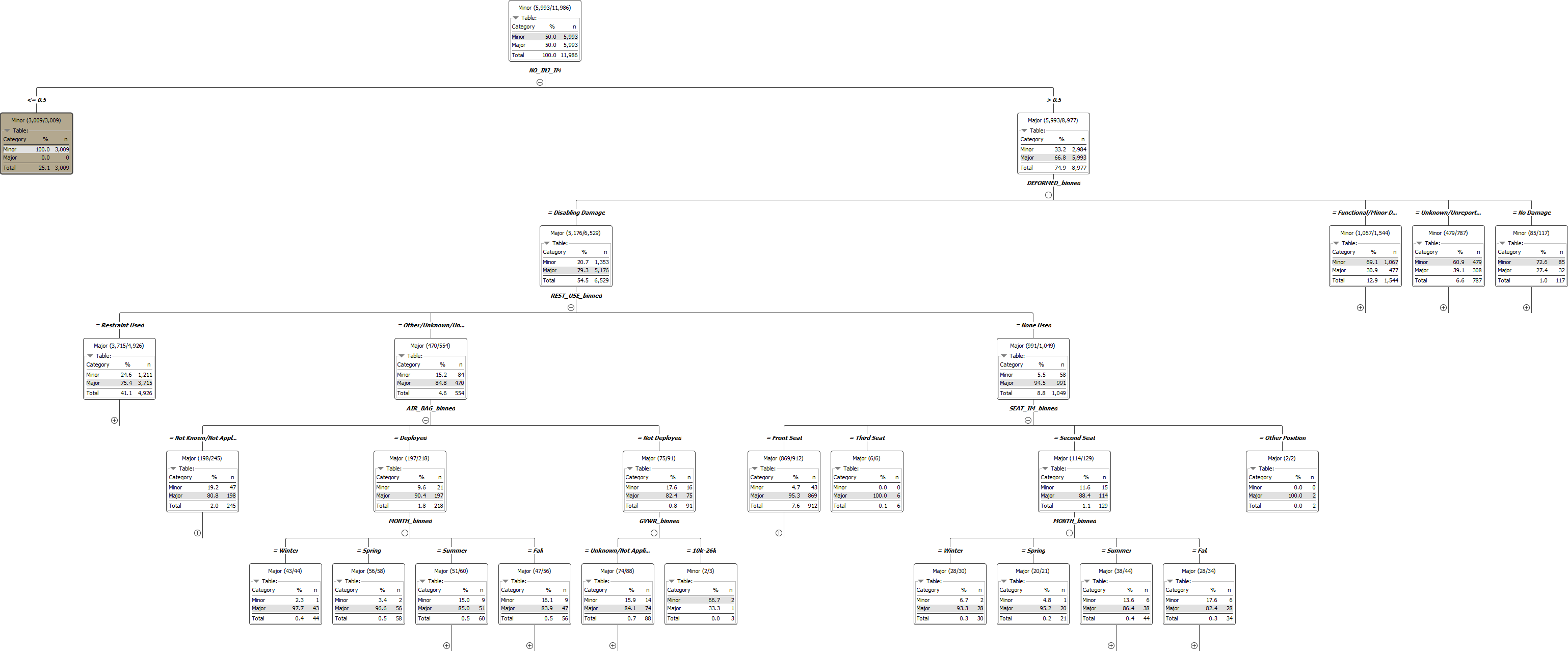
1. Decision tree
2. Random Forest
3. Naïve Bayes

### 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.

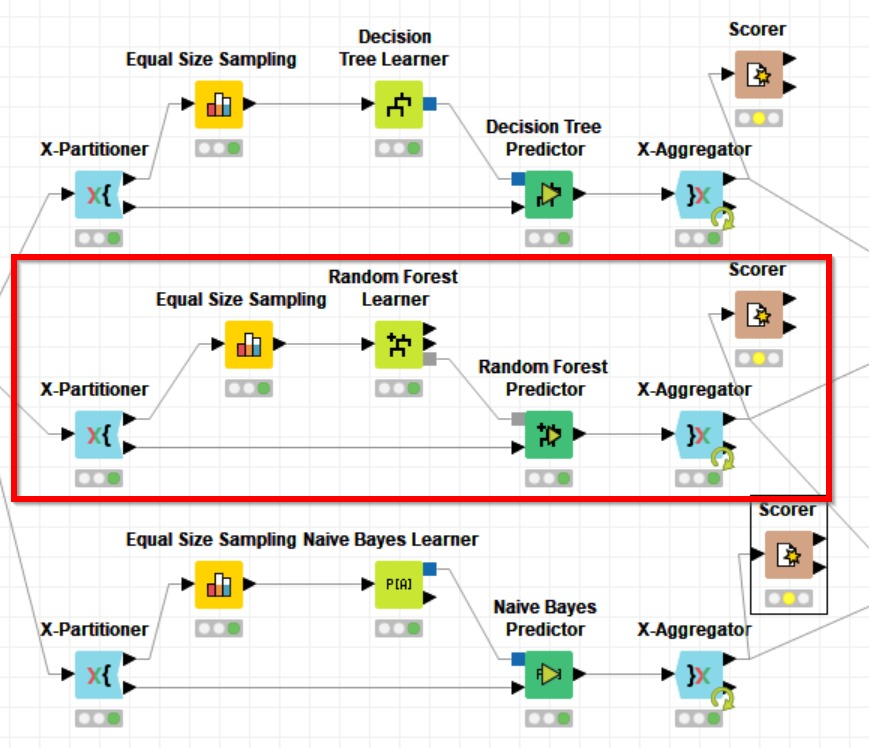


Here, from the below decision tree we can see that our variables like MONTH\_binned, REST\_USE\_binned, DEFORMED\_binned and SEAT\_IM\_binned are important factors in predicting the severity of the injury. Further on exploring the tree we can find VSURCON\_binned, GVWR\_binned are also very important factors which influence the severity of the injury.



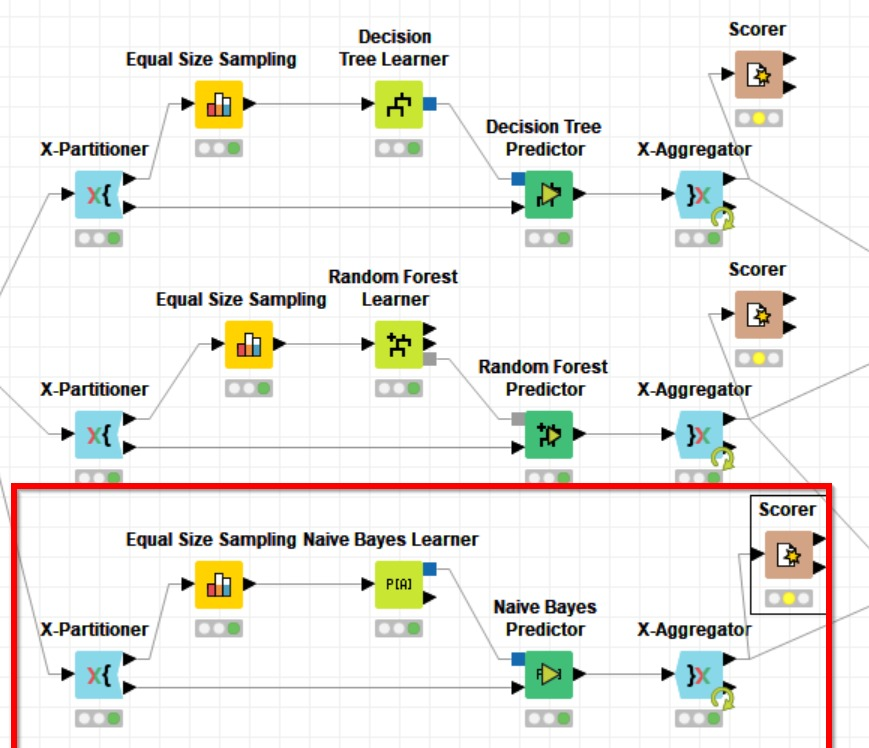
### Random- Forest Model:

* 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.



### Naïve- Bayes Model:

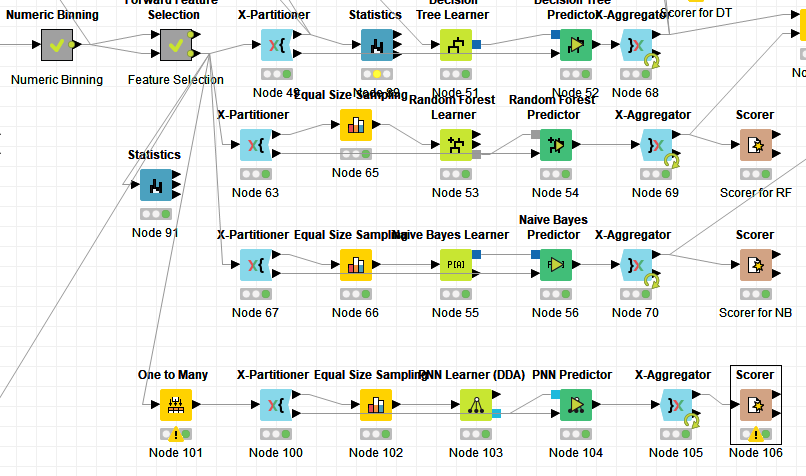
* Naïve- Bayes classification works on Bayes theorem of probability, with a strong assumption that all the predictor variables are completely independent of one another, which is usually not possible in the real-life scenarios, however, we still intended to explore and see how this model performs in 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.



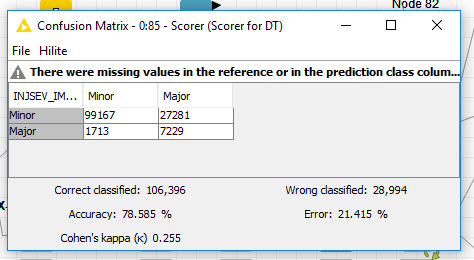
## EVALUATION

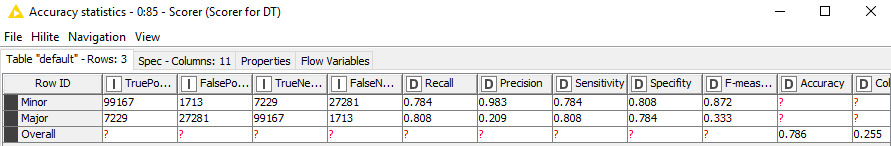
### Accuracy

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:

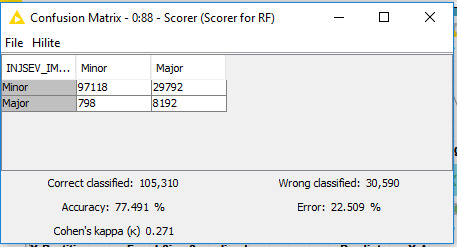
1. Decision Tree Model:

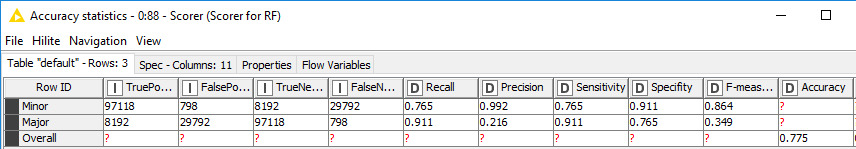




* An accuracy of 78.585% is obtained for this model. The sensitivity (for major injury), i.e. when the major injury is predicted as a major one is 80.8%.

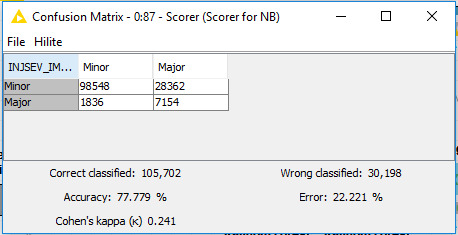
2. Random Forest Model:

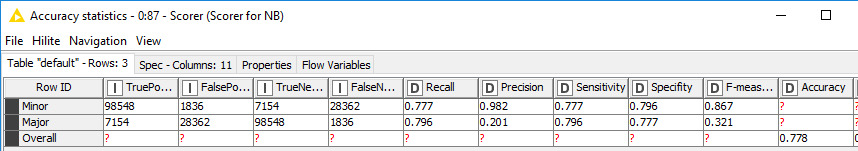




* The accuracy of this model is obtained to be 77.491% with a sensitivity of 91.1%, which is quite significant.

3. Naïve- Bayes Model:

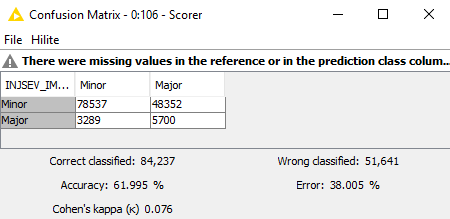




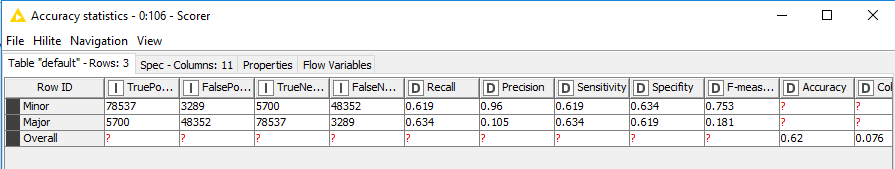
* Accuracy of 77.779 and a sensitivity of 79.6% is observed for Naïve- Bayes model.

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.

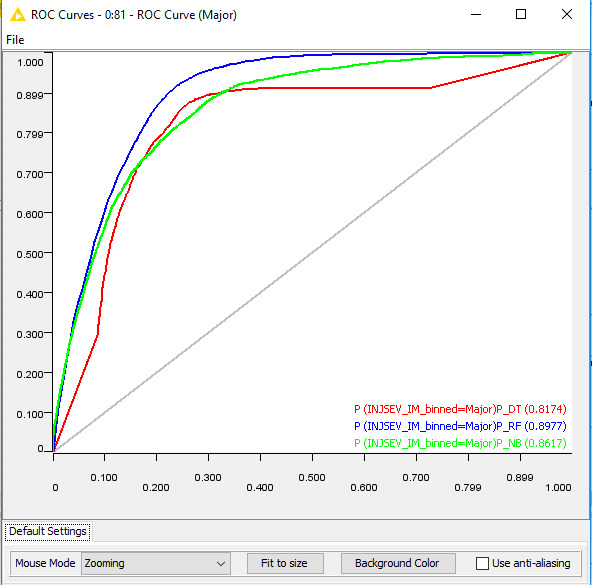
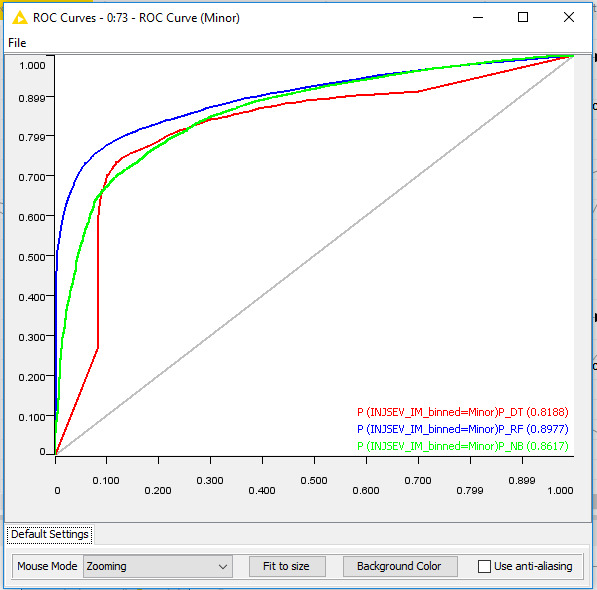


Also, the specificity (63.4%) and the sensitivity (61.9%) for neural network is very less as seen below: -



### 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>