BAYESIAN MODEL COMPARISON ASSIGNMENT

Predicting severity of vehicle crash using Bayesian Classification Methods

Prepared by:

Team Data Insighters

Sreekar Bethu

Balacoumarane Vetrivel

Meghna Vinay Amin

Mohammed Ismail Khan

Table of Contents

1. INTRODUCTION

National Highway Traffic Safety Administration (NHTSA) research offices are the Office of Vehicle Safety Research and the Office of Behavioural Safety Research. The Office of Vehicle Safety Research's mission is to strategize, plan, and implement research programs to continually further the Agency's goals in the reduction of crashes, fatalities, and injuries. In addition to its own data input sources NCSA uses data from other governmental agencies, as well as crash files from several states, to support analytical activities. NCSA conducts one-time and ongoing analysis on a wide range of safety issues.

1.1 Objective:

To build Naïve Bayes and Tree Augmented Naive Bayesian models in different tools like GeNIe, R and SPSS and compare results obtained in each tool to determine the better model based on Accuracy and Confusion Matrix.

1.2 Data Source:

The dataset is obtained from the website of 'BayesiaLab' at http://www.bayesia.com/white_papers/data/NASS_data.csv. Different factors that may play a part in determining the intensity of the crash and level of injury or survival rate of the passenger are listed in the dataset. The dataset has a total number of 21 variables and a total number of 20247 observations. The variables are divided into different categories based on the attribute they represent.

The representation of the variable and their attributes are given as:

1. GV (General Vehicle record): 10 variables
2. OA (Occupant Assessment record): 7 variables
3. VE (Exterior Vehicle record): 4 variables

The target variable for this study is '**OA_MAIS**'.

2. Exploratory Data Analysis

On analyzing the dataset, it was found that the dataset has several missing values in each column including the target variable.

2.1 Missing Value Imputation:

As the method of Predictive Mean Matching produces imputed values closer to the real values, this method has been used to perform the imputation of missing values.

Normal regression method for imputation is not performed for this dataset as there are missing values in all the numeric variables, thus the prediction will not be accurate.

The dataset obtained after performing missing value imputation is as such:

```
> summary(completedData)
   GV_CURBWGT      GV_DVLAT             GV_DVLONG          GV_ENERGY        GV_OTVEHWGT
 Min.   : 670   Min.   :-114.0000   Min.   :-145.00   Min.   :   4.0   Min.   : 640
 1st Qu.:1360   1st Qu.:  -5.0000   1st Qu.: -23.00   1st Qu.: 154.0   1st Qu.:1340
 Median :1530   Median :   0.0000   Median : -14.00   Median : 304.0   Median :1550
 Mean   :1618   Mean   :   0.7613   Mean   : -13.84   Mean   : 505.7   Mean   :1630
 3rd Qu.:1830   3rd Qu.:   7.0000   3rd Qu.:  -7.00   3rd Qu.: 598.0   3rd Qu.:1840
 Max.   :4310   Max.   : 118.0000   Max.   :  84.00   Max.   :9852.0   Max.   :4540
   GV_SPLIMIT        OA_AGE         OA_HEIGHT        OA_WEIGHT         VE_PDOF_TR
 Min.   : 0.00   Min.   : 0.00   Min.   : 59.0   Min.   : 28.00   Min.   :  5
 1st Qu.:35.00   1st Qu.:25.00   1st Qu.:163.0   1st Qu.: 64.00   1st Qu.:115
 Median :40.00   Median :37.00   Median :170.0   Median : 77.00   Median :135
 Mean   :40.72   Mean   :40.17   Mean   :170.8   Mean   : 78.69   Mean   :152
 3rd Qu.:45.00   3rd Qu.:52.00   3rd Qu.:178.0   3rd Qu.: 91.00   3rd Qu.:165
 Max.   :75.00   Max.   :97.00   Max.   :216.0   Max.   :150.00   Max.   :355
  GV_FOOTPRINT
 Min.   :2.468
 1st Qu.:3.925
 Median :4.200
 Mean   :4.365
 3rd Qu.:4.560
 Max.   :7.795
```

2.2 Data Discretization:

To improve the classification performance of the model, the numeric variables in the dataset have been binned into different categories.

Absolute values are considered for discretization for the variables that have negative values.

Variables VE_ORIGAVTW & VE_WHEELBAS are not considered in further process as variable GV_FOORPRINT is the product of the two.

The variables and their subsequent categories are as follows:

Original variable	Categorical variable	Category	Range	Count
GV_CURBWGT	GV_CURBWGT_cat	light wt vehicle	0 to 1500	8796
		medium wt vehicle	1500 to 3000	10330
		heavy wt vehicle	3000 to 4500	77
GV_DVLAT	GV_DVLAT_cat	Low speed Lat	1 to 30	18595
		Medium speed Lat	30 to 70	591
		High speed Lat	70 to 120	17
GV_DVLONG	GV_DVLONG_cat	Low speed Lat	1 to 30	16291
		Medium speed Lat	30 to 70	2812
		High speed Lat	70 to 150	100
GV_ENERGY	GV_ENERGY_cat	Low Impact	0 to 500	13302
		Medium Impact	500 to 2000	5352
		High Impact	2000 to 10000	549
GV_OTVEHWGT	GV_OTVEHWGT_cat	light wt vehicle-ot	0 to 1500	8387
		medium wt vehicle-ot	1500 to 3000	10704
		heavy wt vehicle-ot	3000 to 4600	112
GV_SPLIMIT	GV_SPLIMIT_cat	low speed limit	1 to 30	4112
		medium speed limit	30 to 50	11814
		high speed limit	50 to 100	3277
OA_AGE	OA_AGE_cat	age less than 20	0 to 20	2238
		age between 20 & 40	20 to 40	8483
		age between 40 & 60	40 to 60	5764
		age above 60	60 to 100	2718
OA_HEIGHT	OA_HEIGHT_cat	Short	0 to 165	6633
		Medium	166 to 180	9354
		Tall	181 to 250	3216
OA_WEIGHT	OA_WEIGHT_cat	Wt below 60	0 to 60	3413
		Wt between 60 & 90	61 to 90	10830
		Wt above 90	91 to 160	4960
GV_FOOTPRINT	GV_FOOTPRINT_cat	Less area	1 to 4	5721
		Medium area	4 to 5	11105
		Large area	5 to 10	2377
GV_LANES	GV_LANES_cat	Single lane	1	389
		Double lane	2	7350

		Triple lane	3	3520
		Multi lane	>4	7944
GV_MODELYR	GV_MODELYR_cat	Car mftd before 2005	1999 to 2005	14198
		Car mftd between 2005 & 2010	2005 to 2010	4852
		Car mftd after 2010	2010 to 2013	153
VE_PDOF_TR	VE_PDOF_TR_cat	Left	0 to 90	2417
		Front	90 to 180	12699
		Right	180 to 270	2176
		Rear	270 to 500	1911
GV_WGTCDTR	-	Passenger Car	-	11800
		Truck (<=10000 lbs.)		2414
		Truck (<=6000 lbs.)		4989
GV_BAGDEPLY	-	Deployed	-	9017
		Not Deployed		10186
OA_MANUSE	-	0	-	2175
		1		17028
OA_SEX	-	Female	-	9453
		Male		3750
VE_GAD1	-	Left	-	2997
		Front		11947
		Right		2630
		Rear		1629
OA_MAIS	OA_MAIS_cat	Not injured/Minor injury	0 & 1	16133
		Moderate/Serious/Severe injury	2, 3 & 4	2768
		Critical/Maximum	5 & 6	302

2.3 Train and Test Split:

In this analysis, dataset has been split into a 70:30 ratio for training & testing the dataset after imputing missing values into the explanatory variables, removing missing values from target and discretizing.

Justification: Using the trial and error method, the accuracy of the model is measured under three different scenarios in GeNIe starting with a 80:20 ratio split followed by a 70:30 ratio split and lastly a 60:40 split as the train data should be greater than the test data.

The results are mentioned below:

Model	Split	Train	Test
Naïve Bayes	60:40	84.15%	84.03%
	70:30	84.14%	84.42%
	80:20	84.03%	84.48%
TAN	60:40	84.58%	84.79%
	70:30	84.68%	84.98%
	80:20	84.68%	84.78%

Inference drawn from the table is that the Accuracy values of the 70:30 split is higher and there is more similarity between the Train and Test data result. Therefore the 70:30 split is chosen to build the model.

The train and test data for comparing all Bayesian models is constant throughout.

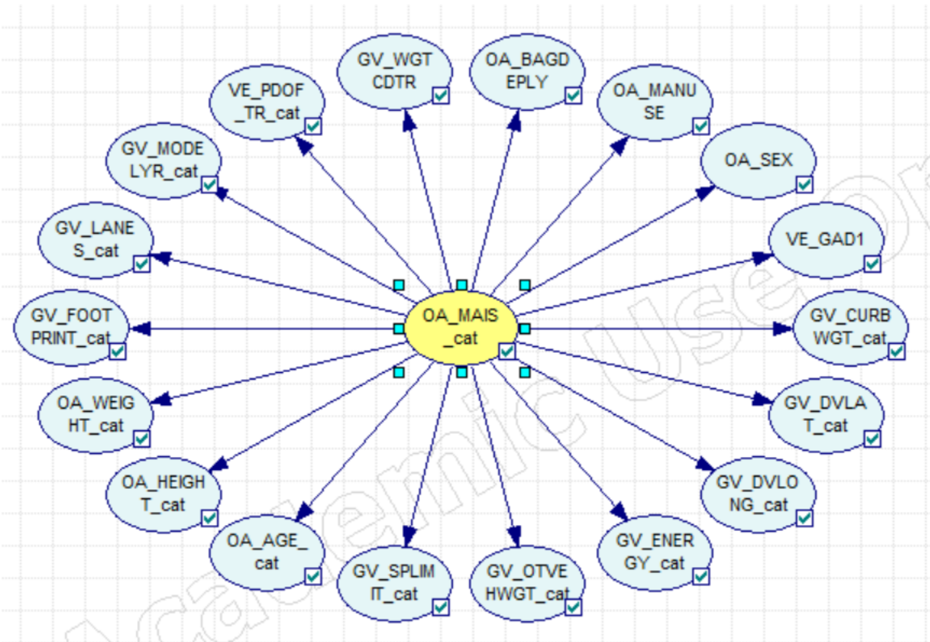
3. BAYESIAN MODELING

Naïve Bayes:

A Naive Bayes model assumes that all the attributes of an instance are independent of each other given the class of that instance. This model does not look for correlation between the explanatory variables and hence there will not be any connections between the explanatory variables.

3.1 Naïve Bayes Model in GeNIe:

3.1.1 Network



The above network shows the dependence of all variables on the target variable '**OA_MAIS_cat**'.

From the above network, it can be inferred that the model does not consider any correlation between the explanatory variables and hence there are no connections among them.

Target variable is connected to all explanatory variables.

3.1.2 Accuracy

The accuracy of the train and test model is given as:

Train Model	Test Model
Accuracy: OA_MAIS_cat = 0.841467 (11311/13442) SCritical_Maximum = 0.127962 (27/211) SModerate_Serious_Severe_injury = 0.207946 (403/1938 SNot_injured_Minor_injury = 0.963517 (10881/11293)	Accuracy: OA_MAIS_cat = 0.844298 (4864/5761) SCritical_Maximum = 0.131868 (12/91) SModerate_Serious_Severe_injury = 0.226506 (188/830) SNot_injured_Minor_injury = 0.963636 (4664/4840)

3.1.3 Confusion Matrix

The confusion matrix of the Train is given as:

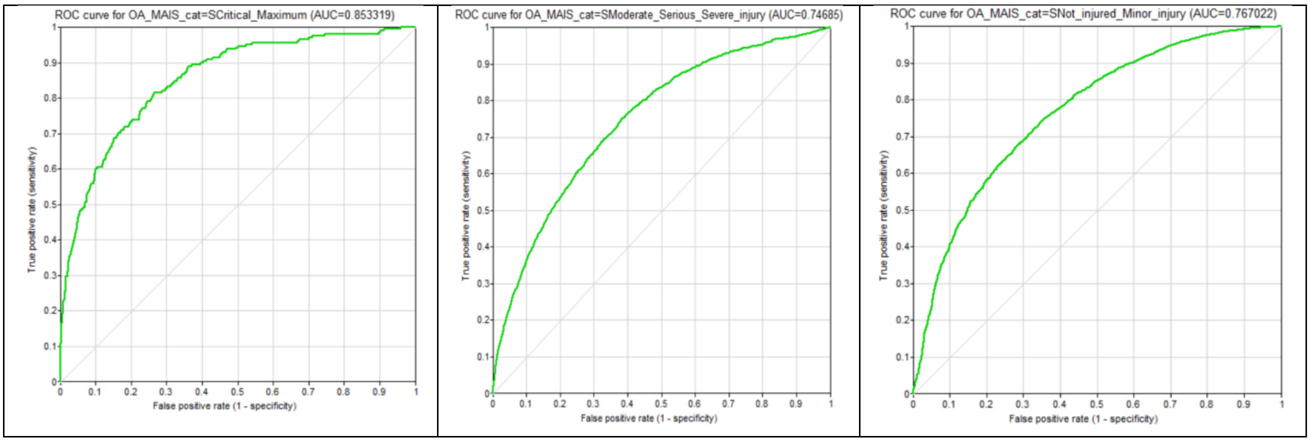
	SCritical_Maximum	SModerate_Serious_Severe_injury	SNot_injured_Minor_injury
SCritical_Maximum	**27**	70	114
SModerate_Serious_Severe_injury	38	**403**	1497
SNot_injured_Minor_injury	14	398	**10881**

The confusion matrix of the Test is given as:

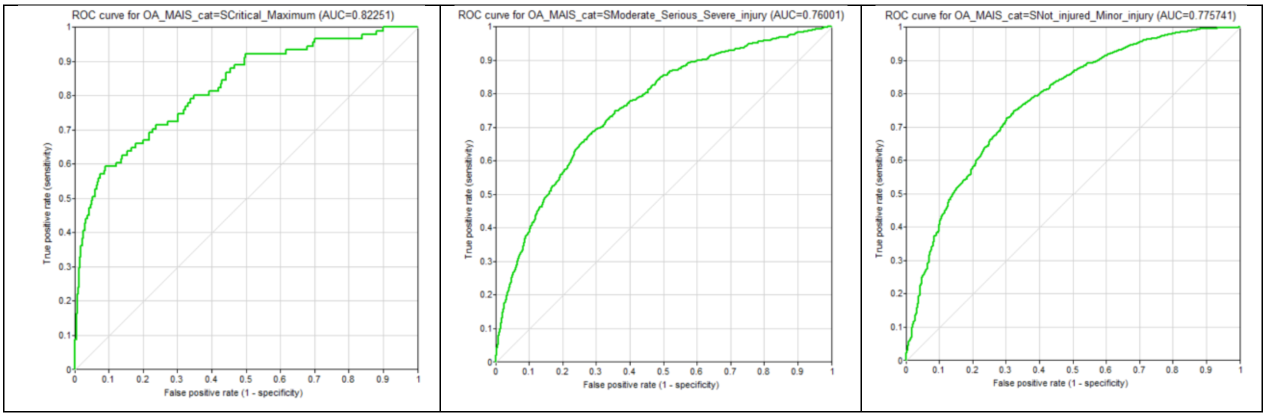
	SCritical_Maximum	SModerate_Serious_Severe_injury	SNot_injured_Minor_injury
SCritical_Maximum	**12**	29	50
SModerate_Serious_Severe_injury	26	**188**	616
SNot_injured_Minor_injury	10	166	**4664**

3.1.4 ROC Curves:

Train Data:



Test Data:



Area under the curve is greater than 70% for all the categories in the target variable.

3.2 Naive Bayes Model in R:

3.2.1 Train Dataset

Naive Bayes model is built in R on Train dataset and below is the confusion matrix of the model:

```
                              true
pred                            Critical/Maximum Moderate/Serious/Severe injury Not injured/Minor injury
  Critical/Maximum                            29                             41                       17
  Moderate/Serious/Severe injury              97                            548                      939
  Not injured/Minor injury                    85                           1349                    10337
```

Model accuracy and Precision accuracy is as below:

```
> Train_accuracy
[1] 0.8119327
> Train_precision
              Critical/Maximum Moderate/Serious/Severe injury       Not injured/Minor injury
                     0.3333333                      0.3459596                      0.8781752
```

Below are the overall details about the model:

```
Overall Statistics

               Accuracy : 0.8119
                 95% CI : (0.8052, 0.8185)
    No Information Rate : 0.8401
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2393
 Mcnemar's Test P-Value : <2e-16

Statistics by Class:

                     Class: Critical/Maximum Class: Moderate/Serious/Severe injury Class: Not injured/Minor injury
Sensitivity                         0.137441                               0.28277                          0.9153
Specificity                         0.995616                               0.90994                          0.3327
Pos Pred Value                      0.333333                               0.34596                          0.8782
Neg Pred Value                      0.986372                               0.88278                          0.4279
Prevalence                          0.015697                               0.14417                          0.8401
Detection Rate                      0.002157                               0.04077                          0.7690
Detection Prevalence                0.006472                               0.11784                          0.8757
Balanced Accuracy                   0.566529                               0.59636                          0.6240
```

3.2.2 Test Data

Below is the overall statistics of test dataset.

```
Confusion Matrix and Statistics

                                Reference
Prediction                        Critical/Maximum Moderate/Serious/Severe injury Not injured/Minor injury
  Critical/Maximum                              13                             28                       11
  Moderate/Serious/Severe injury                35                            218                      380
  Not injured/Minor injury                      43                            584                     4449

Overall Statistics

               Accuracy : 0.8124
                 95% CI : (0.802, 0.8224)
    No Information Rate : 0.8401
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2303
 Mcnemar's Test P-Value : 1.403e-13

Statistics by Class:

                     Class: Critical/Maximum Class: Moderate/Serious/Severe injury Class: Not injured/Minor injury
Sensitivity                         0.142857                               0.26265                          0.9192
Specificity                         0.993122                               0.91584                          0.3192
Pos Pred Value                      0.250000                               0.34439                          0.8765
Neg Pred Value                      0.986337                               0.88066                          0.4292
Prevalence                          0.015796                               0.14407                          0.8401
Detection Rate                      0.002257                               0.03784                          0.7723
Detection Prevalence                0.009026                               0.10988                          0.8811
Balanced Accuracy                   0.567989                               0.58924                          0.6192
```

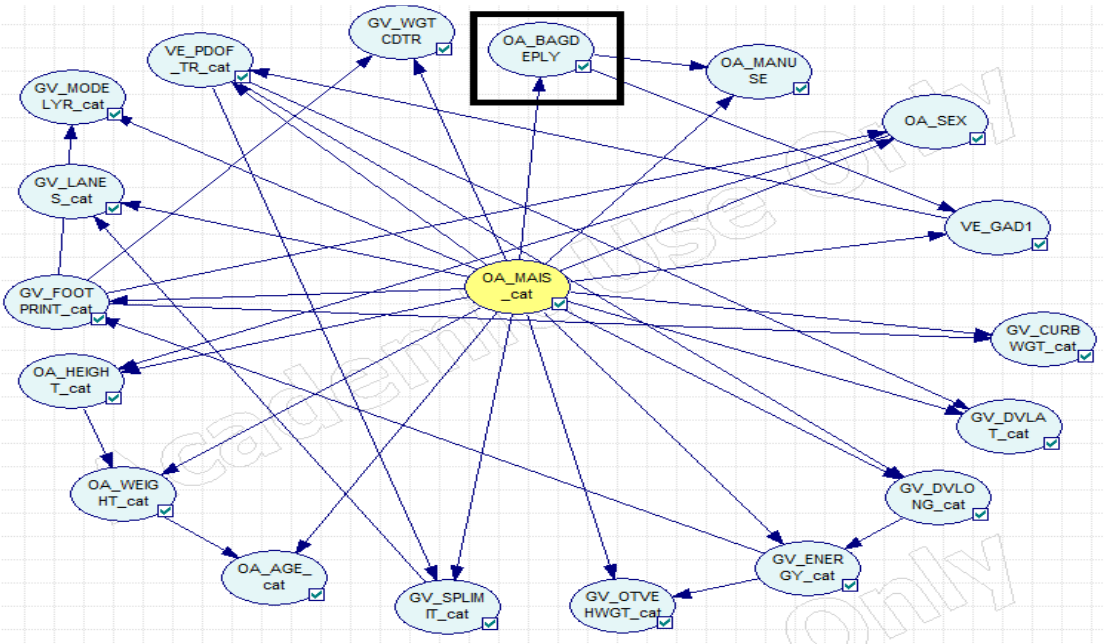
Model built has a train accuracy of 81.19% and a test accuracy of 81.24%. Model precision accuracy is high for 'Not injured/Minor injury' category, that is, for State 0 & State 1.

Tree Augmented Network:

A Tree Augmented Naïve (TAN) Bayesian network maintains the structure of the Naive Bayesian network and augments it by adding edges between the explanatory variables in order to capture the correlations between the attributes. This is where each attribute variable will have one parent among the other attributes (except one). The graph among the attributes will then form a tree (hence the name). TAN has better classification performance than standard Naive Bayes.

3.3 Tree Augmented Naive Bayes Model in GeNIe:

3.3.1. Network:



The above model shows the dependence of the variables on each other and on the target variables.

From the above network, it can be inferred that the model did consider the correlation between the explanatory variables and hence there are connections among them. Root node in the above network built is 'OA_BAGDEPLY'. This variable gives the information whether the air bags in the vehicle were deployed during the crash or not. This variable determines the severity of injury to the passengers. If the air bags are deployed, then severity would be less as compared to when the air bags are not deployed.

3.3.2 Accuracy:

Train Model	Test Model
Accuracy: OA_MAIS_cat = 0.846898 (11384/13442) SCritical_Maximum = 0.113744 (24/211) SModerate_Serious_Severe_injury = 0.200722 (389/1938) SNot_injured_Minor_injury = 0.971487 (10971/11293)	Accuracy: OA_MAIS_cat = 0.849852 (4896/5761) SCritical_Maximum = 0.0989011 (9/91) SModerate_Serious_Severe_injury = 0.206024 (171/830) SNot_injured_Minor_injury = 0.97438 (4716/4840)

3.3.3 Confusion Matrix:

Confusion Matrix for Train data:

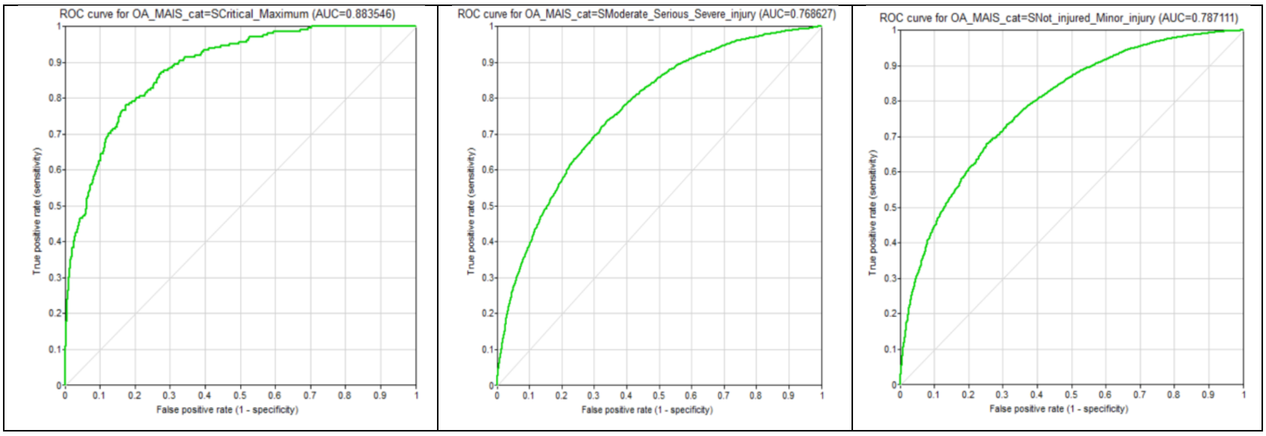
	SCritical_Maximum	SModerate_Serious_Severe_injury	SNot_injured_Minor_injury
SCritical_Maximum	**14**	83	114
SModerate_Serious_Severe_injury	34	**392**	1512
SNot_injured_Minor_injury	24	373	**10896**

Confusion Matrix for Test data:

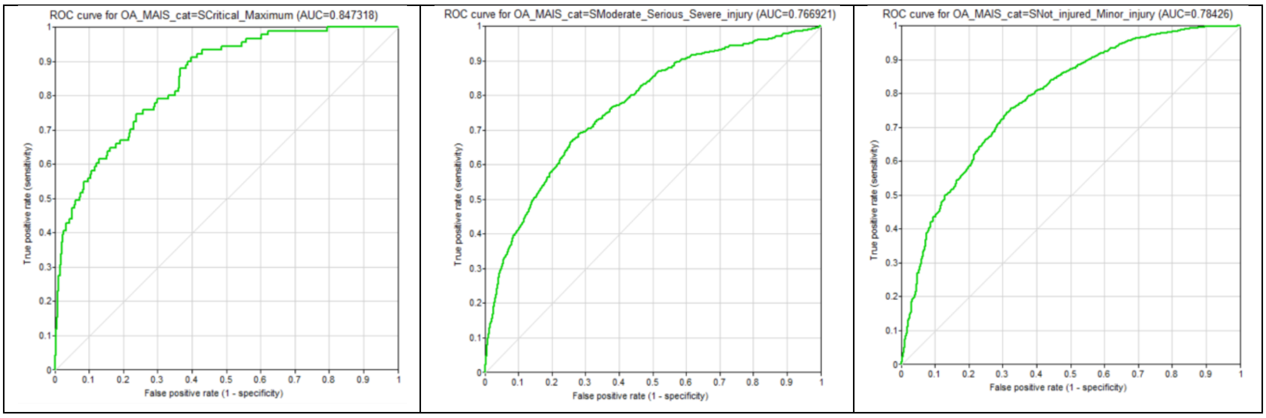
	SCritical_Maximum	SModerate_Serious_Severe_injury	SNot_injured_Minor_injury
SCritical_Maximum	**9**	28	54
SModerate_Serious_Severe_injury	21	**171**	638
SNot_injured_Minor_injury	5	119	**4716**

3.3.4 ROC Curve:

Train Data:



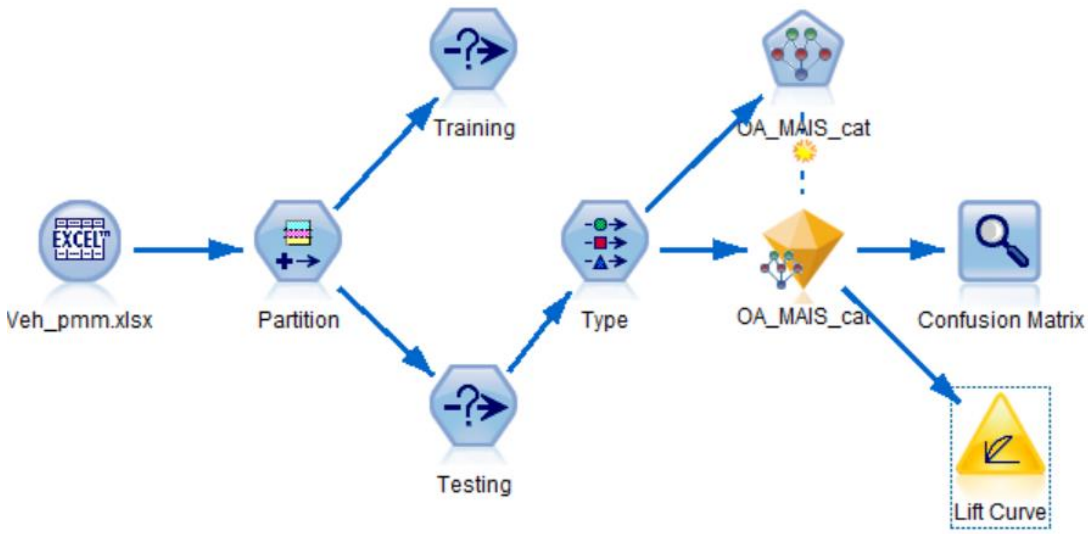
Test Data:



Area under the curve is greater than 70% for all the categories in the target variable.

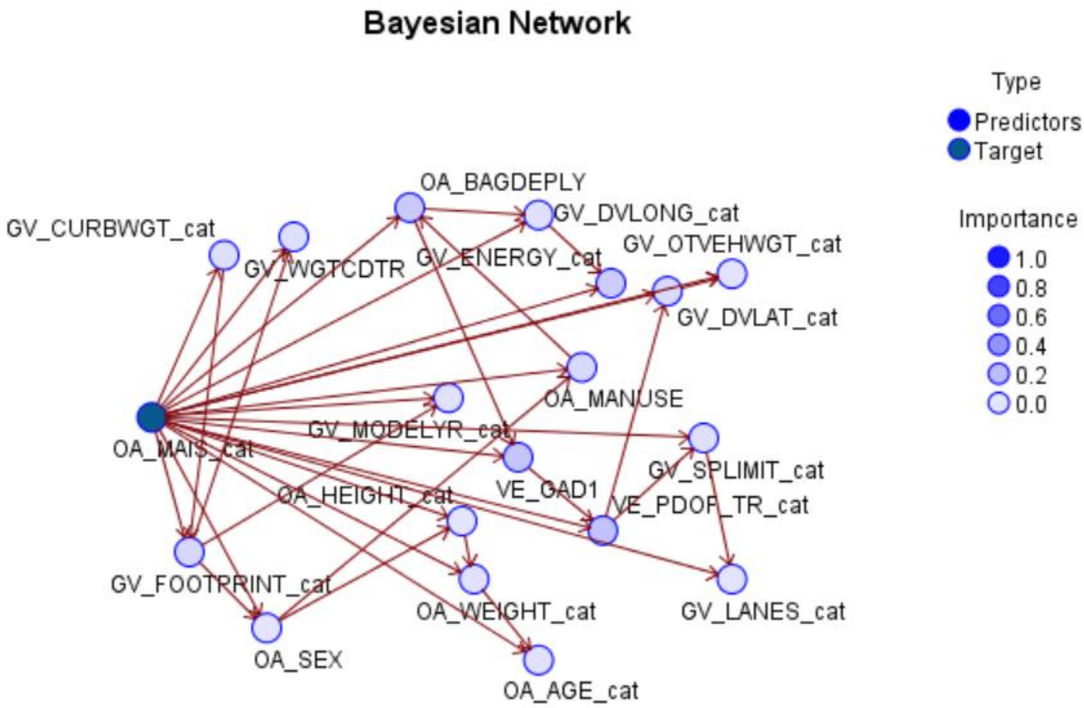
3.4 Tree Augmented Naive Bayes Model in SPSS:

3.4.1 Stream:



The TAN model is built on SPSS with train data and tested with test data.

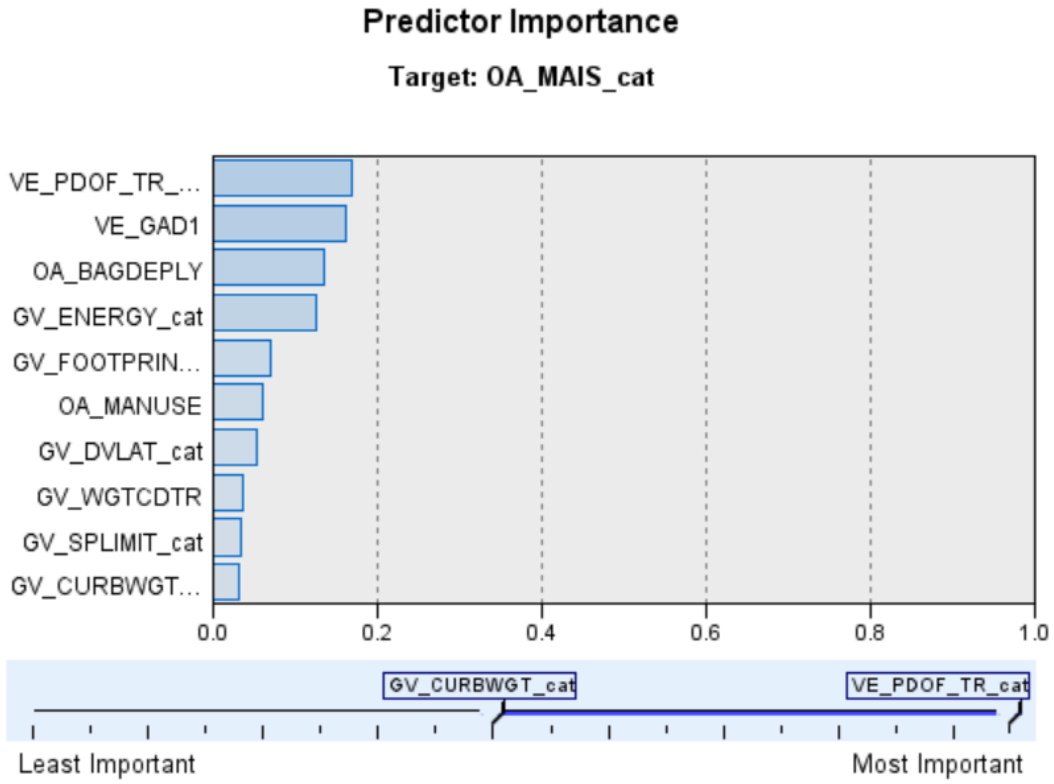
3.4.2 Network:



The dependency between the target variables and the predictors is given in the network above.

3.4.3 Predictor Importance

The Predictor importance of the predictor variables is given as:



3.4.4 Analysis for Train data:

Results for output field OA_MAIS_cat

Comparing $B-OA_MAIS_cat with OA_MAIS_cat

Correct	11,390	84.73%
Wrong	2,052	15.27%
Total	13,442	

Coincidence Matrix for $B-OA_MAIS_cat (rows show actuals)

	Critical/Maximum	Moderate/Serious/Severe injury	Not injured/Minor injury
Critical/Maximum	21	66	124
Moderate/Serious/Severe injury	22	318	1,598
Not injured/Minor injury	7	235	11,051

Performance Evaluation

Critical/Maximum	3.287
Moderate/Serious/Severe injury	1.271
Not injured/Minor injury	0.029

Confidence Values Report for $BP-OA_MAIS_cat

Range	0.351 - 1.0
Mean Correct	0.871
Mean Incorrect	0.734
Always Correct Above	0.995 (1.41% of cases)
Always Incorrect Below	0.384 (0.03% of cases)
90% Accuracy Above	0.732
2.0 Fold Correct Above	0.826 (92.37% of cases)

As per the confusion matrix generated in SPSS, the accuracy of the model is 84.73%

3.4.5 Analysis for Test data:

Results for output field OA_MAIS_cat

Comparing $B-OA_MAIS_cat with OA_MAIS_cat

'Partition'	2_Testing	
Correct	4,916	85.17%
Wrong	856	14.83%
Total	5,772	

Coincidence Matrix for $B-OA_MAIS_cat (rows show actuals)

'Partition' = 2_Testing	Critical/Maximum	Moderate/Serious/Severe injury	Not injured/Minor injury
Critical/Maximum	24	21	43
Moderate/Serious/Severe injury	13	149	680
Not injured/Minor injury	3	96	4,743

Performance Evaluation

'Partition' = 2_Testing	
Critical/Maximum	3.673
Moderate/Serious/Severe injury	1.345
Not injured/Minor injury	0.034

Confidence Values Report for $BP-OA_MAIS_cat

'Partition' = 2_Testing	
Range	0.347 - 1.0
Mean Correct	0.87
Mean Incorrect	0.739
Always Correct Above	0.994 (1.4% of cases)
Always Incorrect Below	0.347 (0% of cases)
90.01% Accuracy Above	0.738
2.0 Fold Correct Above	0.827 (92.59% of cases)

As per the confusion matrix generated in SPSS, the accuracy of the model is 85.21%

4. IMPLEMENTING MODEL IN DIFFERENT SCENARIOS:

The model is implemented in different scenarios and the results are shown as in different cases.

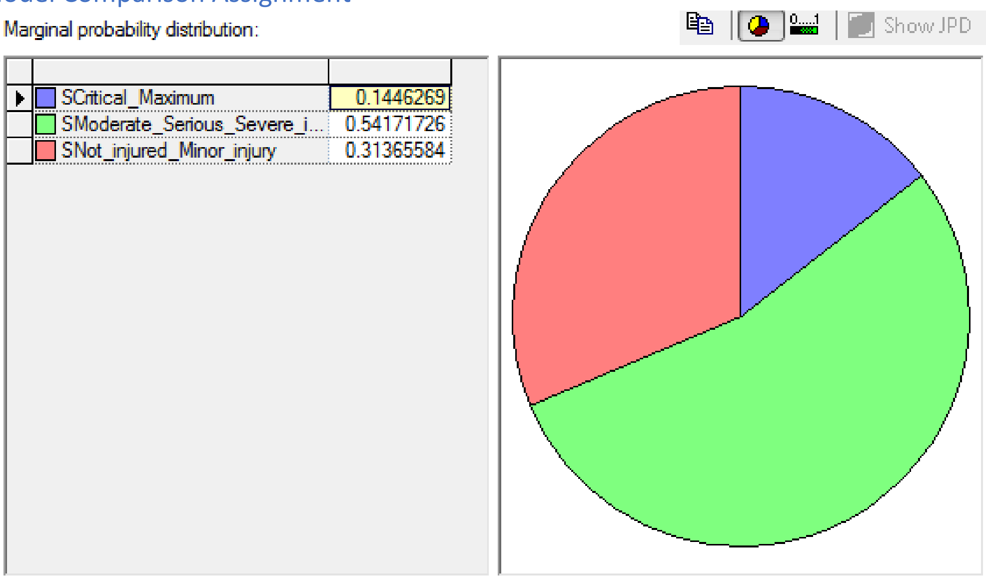
CASE 1:

Consider a 62-year-oldman whose height is 184 CMS with a weight of 96kg driving on a double lane road and has encountered a crash with high impact.

Variables and their categories:

OA_AGE_cat	Age above60
OA_HEIGHT_cat	Tall
OA_WEIGHT_cat	Above 90
OA_SEX	Male
GV_LANES_cat	Double Lane
GV_ENERGY_cat	High Impact

On feeding these input values to the model, the following graph is generated:



From the above graph it can be inferred that the probability of the passenger suffering a minor injury, or a severe injury is less. The passenger maybe moderately, seriously or severely injured if these circumstances would occur. The probability of such an injury is 54.17%.

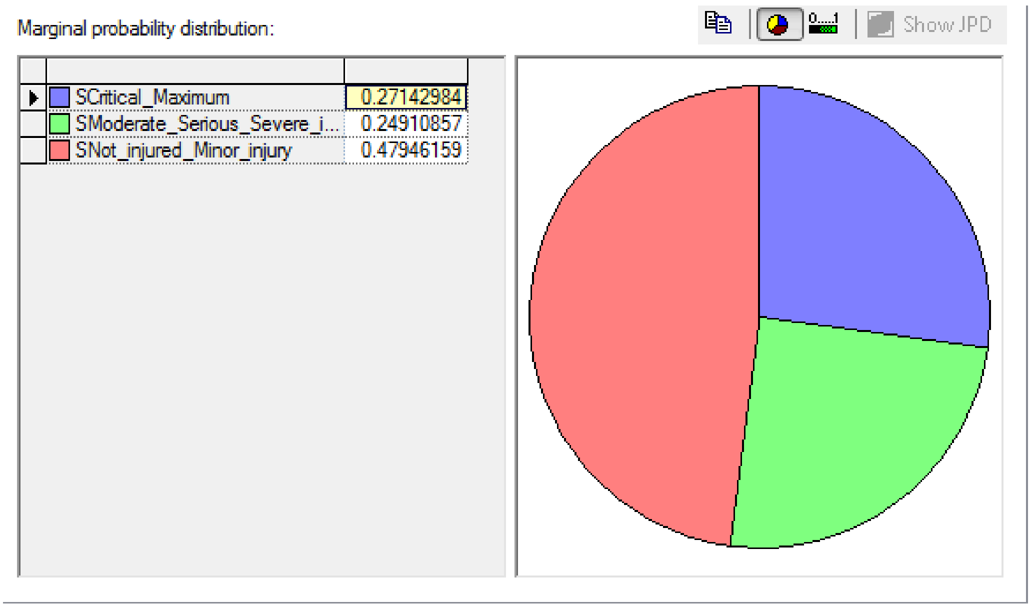
CASE 2:

Consider a woman driving at a high speed in a multi lane road with a low speed limit and has encountered a crash with medium impact.

Variables and their categories:

OA_SEX	**Female**
GV_LANES_cat	Multi Lane
GV_ENERGY_cat	Medium Impact
GV_DVLONG_cat	High speed Long
GV_DVLAT_cat	High speed Lat
GV_SPLIMIT_cat	low speed limit

On feeding these input values to the model, the following graph is generated:



From the above graph it can be inferred that the level of injury of the passenger is minor with the probability being close to 48%.

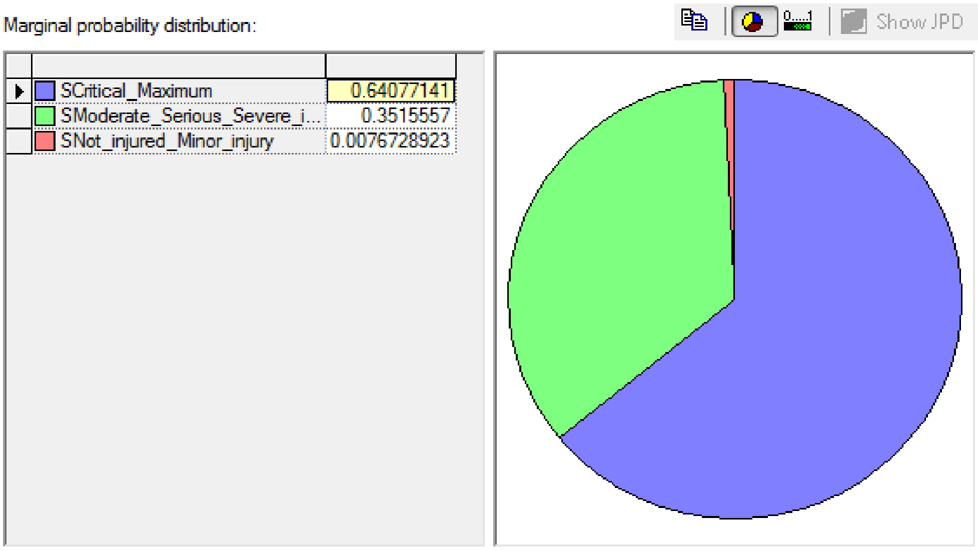
CASE 3:

Consider a person driving a car manufactured before 2005, at a high speed on a road with high speed limit has encountered a crash from the front.

Variables and their categories:

GV_MODELYR_cat	Before 2005
VE_GAD1	Front
GV_SPLIMIT_cat	High speed
GV_DVLONG_cat	High speed Lat
GV_DVLAT_cat	High speed Lat

On feeding these input values to the model, the following graph is generated:



From the above graph it can be inferred that the passenger enduring such a situation will either die or suffer major injuries. The probability for the event is 64%.

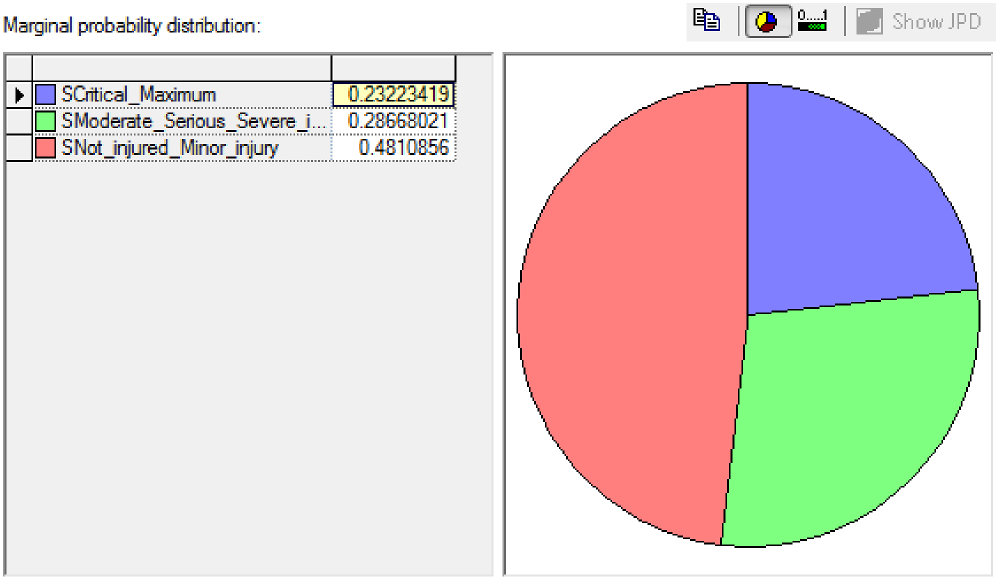
CASE 4:

Consider a person driving a heavy weight car of medium size at a low speed. The vehicle has encountered a low impact crash and has covered less footprint area.

Variables and their categories:

GV_OTVEHWGT_cat	Medium weight vehicle
GV_ENERGY_cat	Low impact
GV_DVLONG_cat	Low speed longitude
GV_DVLAT_cat	low speed Lat
GV_CURBWGT_cat	Heavy weight vehicle
GV_FOOTPRINT_cat	Less area

On feeding these input values to the model, the following graph is generated:



From the above graph it can be inferred that the level of injury to the passenger will be less. The probability of the passenger escaping unhurt or with very minor injuries is 48%.

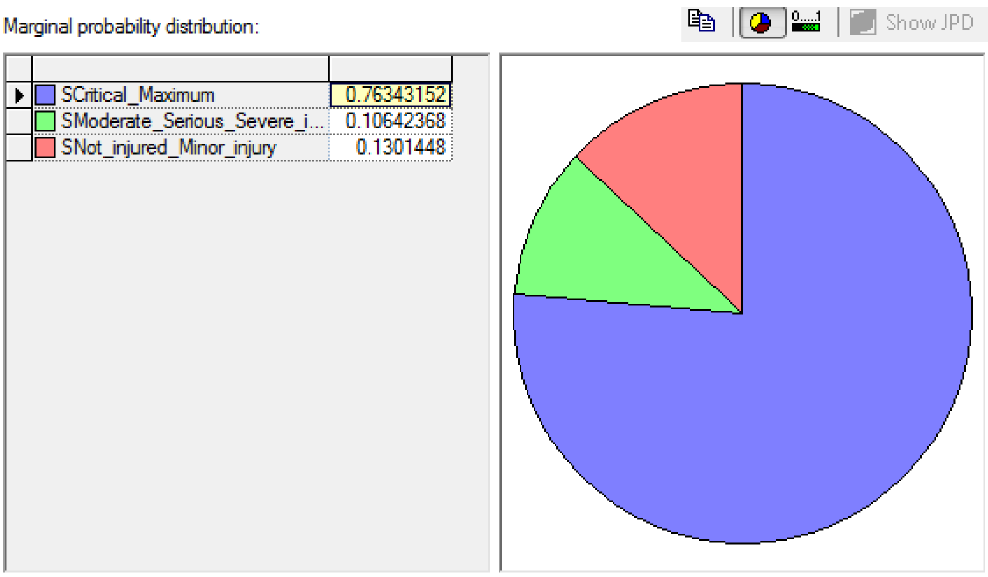
CASE 5:

Consider a person driving a car on a multi-lane is hit by heavy vehicle with high impact on the left side of the car but covered a lesser area post crash.

Variables and their categories:

GV_LANES_cat	Multi lane
GV_OTVEHWGT_cat	Heavy vehicle
GV_ENERGY_cat	High Impact
VE_GAD1	Left side
GV_FOOTPRINT_cat	Less area

On feeding these input values to the model, the following graph is generated:



From the above graph it can be inferred that the level of injury to the passenger will be Critical/Maximum. The probability of the passenger injured with Critical to Maximum injury is 76.3%.

5. COMPARISON OF MODELS:

Given the four models built in GeNIe, R and SPSS a comparative study on each model is given below:

Model	Accuracy of model on Test data set
Naïve bayes in R	81.24
Naïve bayes in GeNIe	84.4
TAN in GeNIe	84.9
TAN in SPSS	85.17

Model	Target Category	Sensitivity %
Naïve Bayes R – Train	Critical/Maximum	13.74
	Moderate/Serious/Severe Injury	28.27
	Not injured_Minor Injury	91.53

Model	Target Category	Sensitivity %
Naïve Bayes R – Test	Critical/Maximum	14.2
	Moderate/Serious/Severe Injury	26.26
	Not injured_Minor Injury	91.92

Model	Target Category	Sensitivity %
Naïve Bayes GeNIe - Train	Critical/Maximum	34.17
	Moderate/Serious/Severe Injury	46.26
	Not injured_Minor Injury	87.1

Model	Target Category	Sensitivity %
Naïve Bayes GeNIe - Test	Critical/Maximum	25
	Moderate/Serious/Severe Injury	49.08
	Not injured_Minor Injury	87.5

Model	Target Category	Sensitivity %
TAN GeNIe – Train	Critical/Maximum	19.4
	Moderate/Serious/Severe Injury	50.06
	Not injured_Minor Injury	87.01

Model	Target Category	Sensitivity %
TAN GeNIe – Test	Critical/Maximum	25.71
	Moderate/Serious/Severe Injury	53.77
	Not injured_Minor Injury	87.2

Model	Target Category	Sensitivity %
TAN SPSS – Train	Critical/Maximum	42
	Moderate/Serious/Severe Injury	51.3
	Not injured_Minor Injury	86.5

Model	Target Category	Sensitivity %
TAN SPSS – Test	Critical/Maximum	43.9
	Moderate/Serious/Severe Injury	55.6
	Not injured_Minor Injury	87.3

Considering the above factors and the results from implementation scenarios performed across tools, it can be concluded that the Tree Augmented Naïve Bayes classification method gives a better result than Naïve Bayes classification. This can also be corroborated with the theory that TAN classifier considers the connection between the predictor variables, whereas Naïve Bayes considers each predictor variables as independent of each other.

6. CONCLUSION

In this study, four Bayesian models were built, namely, Naïve Bayes in GeNIe and R and Tree Augmented Naïve Bayes in GeNIe and SPSS. The models were compared based on their accuracy and sensitivity. A few scenarios have been implemented in Tree Augmented Naïve Bayes model in GeNIe. From the study it can be concluded that Tree Augmented Naïve Bayes model gives better accuracy and sensitivity for categories of target.