Bayesian Modelling Workshop

ASSIGNMENT – 2

EB-5103 ADVANCED BUSINESS ANALYTICS

Submitted By: -

1. Praman Shukla(A0163239A)
2. Prashant Jain(A0163380J)
3. Praveen Tiwari(A0163322R)
4. Li Meiyao (A0163379U)
5. Pooja Gupta (A0163281J)

Table of Contents

TABLE OF CONTENTS

List of Figures: -

List of tables: -

➢ BAYESIAN NETWORK

A **Bayesian network**, Bayes network, belief network, Bayes (IAN) model or probabilistic directed acyclic graphical model is a **probabilistic graphical model** (a type of statistical model) that represents a set of random variables and their conditional dependencies via a directed acyclic graph (**DAG**).

Naive Bayes (NB) and **Tree Augmented Naive Bayes** (**TAN**) are probabilistic graphical models used for modelling huge datasets involving lots of uncertainties among its various interdependent feature sets. The **instances** are described using a set of variables called **attributes**. A **Naive Bayes** model assumes that all the **attributes** of an instance are **independent** of each other given the class of that instance.

Whereas in the **TAN** model, every **attribute** is **dependent** on its **class** and one other **attribute** from the **feature** set. Since this model incorporates the **dependencies** among the attributes, it is **more realistic** than a **Naive Bayes** model. This project analyses the performance of these two models on Vehicle safety dataset [1].

➢ DATASET DESCRIPTION & PROCEDURAL FLOWCHART

The dataset was collected from Bayesia website, containing **21 variables** and more than **20,000 observations**. Out of these 21 variables, two of them **(WHEELBAS & ORIGAVTW)** were converted into a calculated field **GV_FOOTPRINT** and amongst the rest of the attributes, **13** are **numerical variables** and **6** are **categorical variables**.

Table 1. Data Set Description

S.No.	Name	Description	Variable type	Data Type
1.	GV_CURBWGT	Vehicle curb weight	Explanatory	Numerical
2.	GV_DVLAT	Lateral component of Delta V	Explanatory	Numerical
3.	GV_DVLONG	Longitudinal component of Delta V	Explanatory	Numerical
4.	GV_ENERGY	Energy absorption	Explanatory	Numerical
5.	GV_LANES	Number of Lanes	Explanatory	Numerical
6.	GV_MODELYR	Vehicle model year	Explanatory	Numerical
7.	GV_OTVEHWGT	Weight of the other vehicle	Explanatory	Numerical
8.	GV_SPLIMIT	Speed limit	Explanatory	Numerical
9.	GV_WGTCDTR	Truck weight code	Explanatory	Categorical
10.	OA_AGE	Age of Occupant	Explanatory	Numerical
11.	OA_BAGDEPLY	Air Bag System Deployed	Explanatory	Categorical
12.	OA_HEIGHT	Height of Occupant	Explanatory	Numerical
13.	**OA_MAIS**	**Maximum known Occupant AIS**	**Response**	**Categorical**
14.	OA_MANUSE	Manual belt system use	Explanatory	Categorical
15.	OA_SEX	Occupant's Sex	Explanatory	Categorical
16.	OA_WEIGHT	Occupant's Weight	Explanatory	Numerical
17.	VE_GAD1	Deformation Location	Explanatory	Categorical
18.	VE_PDOF_TR	Clock Direction for Principal Direction of Force	Explanatory	Numerical
19.	GV_FOOTPRINT	Vehicle Footprint	Explanatory	Numerical
20.	GV_CURBWGT	Vehicle curb weight	Explanatory	Numerical
21.	**OA_BMI**	**Body Mass Index (Weight/Height^2)**	**Calculated Explanatory**	**Numerical**

The Overview of **Procedural Flowchart** that has been followed in this project is shown below: -

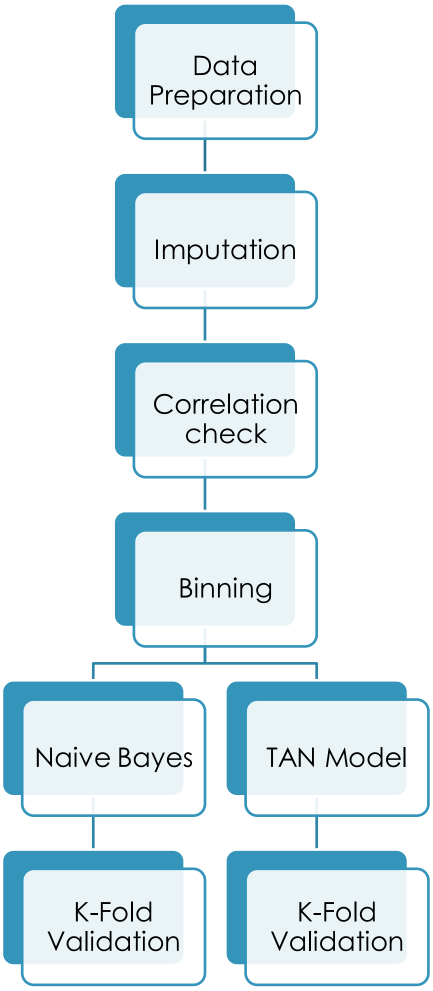


Figure 1: Procedural Flow Chart

➢ DATA PREPARATION

1. **Dealing with Missing values**

   a. **Filling up missing Values in Numerical variables with R-package "mice"**

   To proceed further, all the missing values were replaced with imputed values using **'mice' package in R**. MICE stands for **Multivariate Imputation by Chained Equations**. It generates multiple imputations for incomplete multivariate data by Gibbs sampling. A small piece of code has been written in R to implement imputation on numerical data.

   Table 3 shown below summarizes the **key statistical indicators** of the attributes that have been imputed. The **Original dataset** and **Imputed dataset** have the almost the **same Mean, Median** and **standard deviation** after imputation.

Table 2: Comparison of Statistical indicators before and after Imputation

		Original Dataset				Imputed Dataset			
S. No	Variables	Rows	Mean	Standard Deviation	Median	Rows	Mean	Standard Deviation	Median
1	GV_CURBWGT	20204	1617.26	393.57	1530	20247	1618.31	394.58	1530
2	GV_DVLAT	14049	0.04	13.02	0	20247	0.71	12.58	0
3	GV_DVLONG	14049	-14.76	17.66	-15	20247	-13.88	18.66	-14

4	GV_ENERGY	14049	505.24	645.74	306	20247	503.85	663.98	298
5	GV_LANES	20244	3.28	1.36	3	20247	3.28	1.36	3
6	GV_MODELYR	20247	2003.62	2.77	2003	20247	2003.62	2.77	2003
7	GV_OTVEHWGT	18147	1630.16	411.35	1550	20247	1633.02	413.75	1550
8	GV_SPLIMIT	20016	40.73	11.24	40	20247	40.73	11.25	40
9	OA_AGE	20190	40.17	17.37	37	20247	40.17	17.37	37
10	OA_HEIGHT	17508	170.84	10.75	170	20247	170.82	10.69	170
11	OA_MAIS	19203	0.91	1.04	1	20247	0.91	1.04	1
12	OA_MANUSE	19774	0.88	0.32	1	19774	0.88	0.32	1
13	OA_WEIGHT	17599	78.72	19.64	77	20247	78.71	19.74	77
14	VE_ORIGAVTW	20014	154.75	7.66	154	20247	154.72	7.76	154
15	VE_WHEELBAS	20238	281	28.72	272	20247	280.99	28.72	272
16	VE_PDOF_TR	18298	152.62	67.51	135	20247	152.34	67.03	135
17	GV_FOOTPRINT	20010	4.36	0.64	4.19	20247	4.36	0.64	4.2

After this imputation, dataset contains **20,247 observations** without any missing value in Numerical variables. This final dataset that has been used for further analytical operations.

b. Missing Values in Categorical variables

There are some small number of missing values in three of the Categorical variables as shown below in Table 2 which have been removed from analysis.

Table 3: Missing values in Categorical variables

Variable	Number of Missing values	Action
VE_GAD1	789	Removed in TAN
OA_SEX	234	Removed in TAN
OA_MAIS (class)	1044	Removed from dataset

After removing the missing values from class variables OA_MAIS, the dataset is **left with 19,203 observations** to finally work on. This implies working with almost **95% of the original dataset**.

2. Correlation check of Numerical variables

The second step is to analyse the **multi-collinearity effects** in between the numerical variables and **eliminate the highly-correlated** variables from the analysis of **Naïve Bayes**. Data analysis tool in Excel has been used to compute the correlation matrix using **Pearson's coefficient** (Table 4) and for better interpretation, the results have been compiled in a correlation chart as depicted in Figure 2. It was found that the following pairs of variables are strongly correlated:

Table 4: Correlation between numerical variables (1-7)

	GV_CURBWGT	*GV_DVLAT*	*GV_DVLONG*	*GV_ENERGY*	*GV_LANES*	*GV_OTVEHWGT*	*GV_SPLIMIT*
GV_CURBWGT	1.000	0.008	0.011	0.091	0.008	0.027	0.051
GV_DVLAT	0.008	1.000	-0.003	-0.056	-0.114	0.001	-0.053
GV_DVLONG	0.011	-0.003	1.000	-0.277	-0.001	0.020	-0.019
GV_ENERGY	0.091	-0.056	-0.277	1.000	-0.028	0.092	0.124
GV_LANES	0.008	-0.114	-0.001	-0.028	1.000	-0.007	0.092
GV_OTVEHWGT	0.027	0.001	0.020	0.092	-0.007	1.000	0.066

GV_SPLIMIT	0.051	-0.053	-0.019	0.124	0.092	0.066	1.000
OA_AGE	0.085	0.027	0.044	-0.007	-0.021	0.007	-0.009
OA_HEIGHT	0.155	0.009	-0.017	0.051	-0.002	-0.001	0.038
OA_WEIGHT	0.155	0.022	0.001	0.050	-0.013	0.004	0.031
VE_ORIGAVTW	0.764	0.017	0.023	0.083	0.006	0.033	0.040
VE_WHEELBAS	0.768	0.007	-0.009	0.111	0.001	0.031	0.057
VE_PDOF_TR	-0.032	-0.448	0.583	0.004	0.080	0.044	0.154
GV_FOOTPRINT	0.819	0.011	0.000	0.110	0.003	0.033	0.056

Table 5: Correlation between numerical variables (8-14)

	OA_AGE	OA_HEIGHT	OA_WEIGHT	VE_ORIGAVTW	VE_WHEELBAS	VE_PDOF_TR	GV_FOOTPRINT
GV_CURBWGT	0.085	0.155	0.155	0.764	0.768	-0.032	0.819
GV_DVLAT	0.027	0.009	0.022	0.017	0.007	-0.448	0.011
GV_DVLONG	0.044	-0.017	0.001	0.023	-0.009	0.583	0.000
GV_ENERGY	-0.007	0.051	0.050	0.083	0.111	0.004	0.110
GV_LANES	-0.021	-0.002	-0.013	0.006	0.001	0.080	0.003
GV_OTVEHWGT	0.007	-0.001	0.004	0.033	0.031	0.044	0.033
GV_SPLIMIT	-0.009	0.038	0.031	0.040	0.057	0.154	0.056
OA_AGE	1.000	-0.041	0.130	0.111	0.079	-0.002	0.092
OA_HEIGHT	-0.041	1.000	0.489	0.132	0.180	-0.004	0.176
OA_WEIGHT	0.130	0.489	1.000	0.144	0.184	-0.003	0.183
VE_ORIGAVTW	0.111	0.132	0.144	1.000	0.701	-0.024	0.852
VE_WHEELBAS	0.079	0.180	0.184	0.701	1.000	-0.030	0.969
VE_PDOF_TR	-0.002	-0.004	-0.003	-0.024	-0.030	1.000	-0.030
GV_FOOTPRINT	0.092	0.176	0.183	0.852	0.969	-0.030	1.000

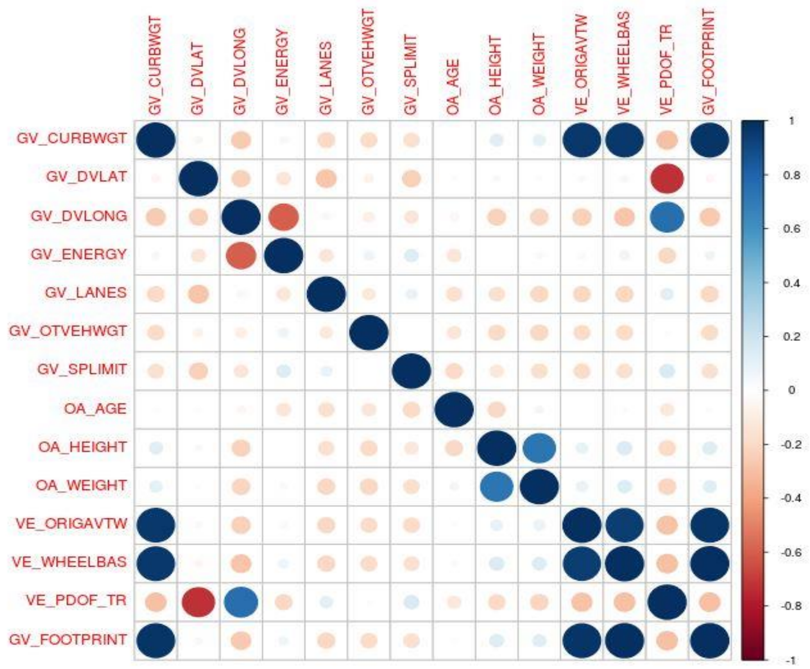


Figure 2: Correlation between numerical variables (easy interpretation)

The results of **correlation matrix** show that the following pair of variables have **high correlation**:

- **GV_CURBWGT** and **GV_FOOTPRINT**
- **GV_FOOTPRINT, VE_ORIGAVTW** and **VE_WHEELBAS**
- **OA_HEIGHT and OA_WEIGHT**

Based on these results, we can safely take only **GV_CURBWGT** out of all these **GV_CURBWGT, GV_FOOTPRINT, VE_ORIGAVTW and VE_WHEELBAS** variables for analysis using **Naïve Bayes** where **independence** of variables is an **assumption** but not in **TAN** where **dependencies** among the attributes are incorporated in the model itself. Also, because of evident correlation, **OA_HEIGHT** and **OA_WEIGHT** variables are **transformed** into a single calculated variable **OA_BMI** which signifies **Body Mass Index.** The formula used for calculating BMI is: -

$$\boldsymbol{BMI = \frac{Weight}{(Height)2}}$$

➢ DATA DISCRETIZATION

1. Binning using Genie

Before proceeding with classification using Bayesian networks, if the dataset has attributes that are **continuous valued**, then those attributes are first **discretized** and then used by the classifier. The method used to discretize the data in this project is called **equal width binning or equal width partitioning** using **Genie Discretization tool**.

Table 2: Binning details tabulated

Variable	Datatype	No. of Bins	Binning Methodology
OA_BAGDEPLY	Categorical	2	Already categorized in dataset
GV_WGTCDTR	Categorical	3	Already categorized in dataset
OA_SEX	Categorical	2	Already categorized in dataset
VE_GAD1	Categorical	4	Already categorized in dataset
OA_MANUSE	Categorical	2	Already categorized in dataset
OA_MAIS	Categorical	4	Intuitive, divided into 4 injury levels
GV_MODELYR	Numerical	5	Intuitive, uniform distribution of years
GV_CURBWGT	Numerical	5	Equal width binning
GV_DVLAT	Numerical	5	Equal width binning
GV_DVLONG	Numerical	4	Equal width binning
GV_ENERGY	Numerical	4	Equal width binning
GV_LANES	Numerical	3	Intuitive, division of lanes
GV_OTVEHWGT	Numerical	4	Equal width binning
GV_SPLIMIT	Numerical	5	Equal width binning
OA_AGE	Numerical	7	Intuitive, uniform distribution of age
OA_HEIGHT	Numerical	5	Intuitive, uniform distribution of height
OA_WEIGHT	Numerical	7	Intuitive, uniform distribution of Weight
VE_ORIGAVTW	Numerical	-	Redundant variable
VE_WHEELBAS	Numerical	-	Redundant variable
VE_PDOF_TR	Numerical	6	Equal width binning
GV_FOOTPRINT	Numerical	6	Equal width binning
OA_BMI	Numerical	3	Intuitive, 3 categories of people

The details of binning process of different numerical variables is described below:

OA_MAIS (Class Variable)

Four different bins are created intuitively for the given variable. The probability of death has been kept in mind while classifying the MAIS into bins, the details have been summarized below: -

- Minor injury: Class 0 and Class 1 (negligible probability of death)
- Moderate Injury: Class 2 and Class 3 (less probability of death (1-10%))
- Severe Injury: Class 4 and Class 5 (medium probability of death (5-50%))
- Death: Class 6 (No chance of Survival)

GV_MODELYR

Separate bins are created for cars manufactured before 2002 and after 2008. The cars manufactured between 2002 to 2008 are uniformly distributed at an interval span of two years.

s1_below_2002	s2_2002_2004	s3_2004_2006	s4_2006_2008	s5_2008_up

GV_CURBWGT

Equal width binning algorithm. Separate bins are created for vehicle weighing below 1000 Kg and above 2500 kg. The vehicle weighing between 1000 kg to 2500 kg are uniformly distributed at an interval of 500 Kg.

s1_below_1000	s2_1000_1500	s3_1500_2000	s4_2000_2500	s5_2500_up

GV_DVLAT

The given variable tells about the Lateral component of Delta V. We followed the original distribution of dataset and applied equal width binning algorithm. Lateral component of Delta V having negative and positive values of 30 are distributed into separate bins. Three bins are created for values ranging from -30 Kmph to 30 kmph which are uniformly distributed at intervals of 20Kmph.

s1_below_n30	s2_n30_n10	s3_n10_10	s4_10_30	s5_30_up

GV_DVLONG

The given variable tells about the longitudinal component of Delta V. We followed the original distribution of dataset and applied equal width binning algorithm. Longitudinal component of Delta V having negative and positive values of 25 are distributed into separate bins. Two bins are created for the values ranging from -25 Kmph to 25 kmph which are uniformly distributed at an interval of 25Kmph.

s1_below_n25	s2_n25_0	s3_0_25	s4_25_up

GV_ENERGY

The given variable tells about the energy absorption. We followed the original distribution of dataset and applied equal width binning algorithm. Three bins are created for values ranging from 0 to 1500. All the values above 1500 are created in a different bin.

s1_below_500	s2_500_1000	s3_1000_1500	s4_1500_up

GV_LANES

The given variable tells the number of lanes present. Binning was done and record was distributed into three categories. The first bin contains the records having less than three lanes generally for roadways and small city roads. The second bin contain the records having three to five lanes generally state Highways. The last bin is created where we have more than five lanes. These are generally the expressway and major highways.

s1_below_3	s2_3_5	s3_5_up

GV_OTVEHWGT

The given variable tells about the weight of the other vehicle during crash. We followed the original distribution of dataset and applied equal width binning algorithm. Vehicles weighing below 1250Kg and above 2250Kg are created in separate bins and rest other records are uniformly distributed at an equal width of 500.

s1_below_1250	s2_1250_1750	s3_1750_2250	s4_2250_up

GV_SPLIMIT

This variable explains the speed limit of the vehicle. We followed the original distribution of dataset and applied equal width binning algorithm. Two bins were created for speed less than 30mph and greater than 60 mph. Other speed values were equally distributed with the interval of 10 mph.

s1_below_30	s2_30_40	s3_40_50	s4_50_60	s5_60_up

OA_AGE

This variable defines the age of the occupant during the crash. As with other continuous variables, age was also binned into 7 groups using equal width binning system taking the original distribution of the variable into account. Two separate bins were created for age groups below 20 and above 70. Other values were equally binned with the interval of 10 years

s1_below_20	s2_20_30	s3_30_40	s4_40_50	s5_50_60	s6_60_70	s7_70_up

OA_BMI

This column, which was calculated using the occupant's weight and height, was binned into 3 categories based on the information from BMI Chart. The category <20 was under weight, 20<BMI<35 was the normal range and >35 were categorised under overweight.

s1_below_20	s2_20_35	s3_35_up

VE_PDOF_TR

This variable describes the clock direction for principle direction of force. Measured in angles, in the confines of collision reconstruction, the PDOF is used to describe the direction of the force that was applied to the vehicle during the collision. Separate bins were created for values below 50 and above 300. The original distribution of the variable was considered and 6 bins were formed using the equal width distribution algorithm.

s1_below_50	s2_50_100	s3_100_160	s4_160_220	s5_220_300	s6_300_up

2. Checking correlation using RStudio

Two of suspected variables, both of which are related to the direction of impact were subjected to Chi-square test to find out the correlation. The results are shown in the figure below.

```
        Pearson's Chi-squared test

data:  CarCrash$VE_PDOF_TR and CarCrash$VE_GAD
X-squared = 29557, df = 12, p-value < 2.2e-16
```

Figure 3: Chi-Sq test in Rstudio

It can be inferred that variables VE_GAD1 and VE_PDOF_TR are highly correlated and both are signifying the direction of impact. So, out of these two variables, only one can be used for Naïve Bayes modelling. Here we choose VE_GAD1 which directly tells the direction of impact from four directions (Left, Right, Front and Centre).

➢ NAÏVE BAYES MODEL (NB)

The Naive Bayes model is a special form of Bayesian network. This model is mainly used for classification problems. The important feature of Naive Bayes model is that, it has very strong independence assumptions. The final dataset **after imputation** and **binning** is used for modelling [1].

a. Naïve Model

Naïve Base Model is created using **Genie** software. The option **Learn New Network** is used for learning new network like **Naïve Bayes** or **TAN**.

Class variable: OA_MAIS

Predictors: All variables except for the correlated variables.

The Naïve Bayes network is shown below: -

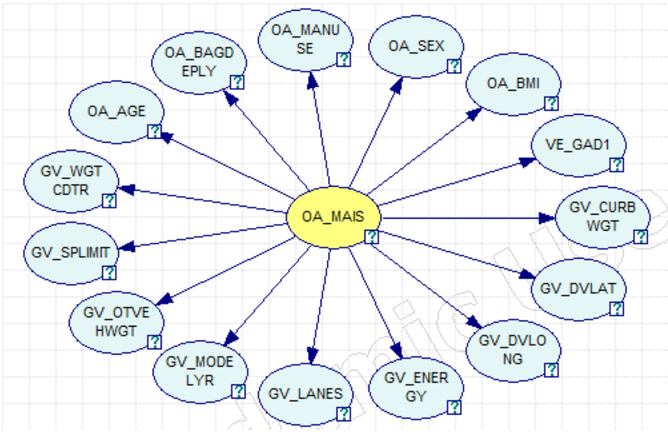


Figure 4: Naive Bayes Model

b. Training set and Test set – K fold cross validation

In a k-fold cross validation, the original data set is divided into k equal parts. Out of those k-parts, one part of the dataset is used for validation or testing and the remaining k-1 parts are used for training the classifier. This process is then repeated k-times and each of the parts is used as testing data, exactly once.

The advantage of this method is that it ensures each instance in the dataset is used both, as a training and testing sample and every instance is used exactly once as a testing sample.

Here, 5-fold cross validation has been used as shown in figure below: -

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5

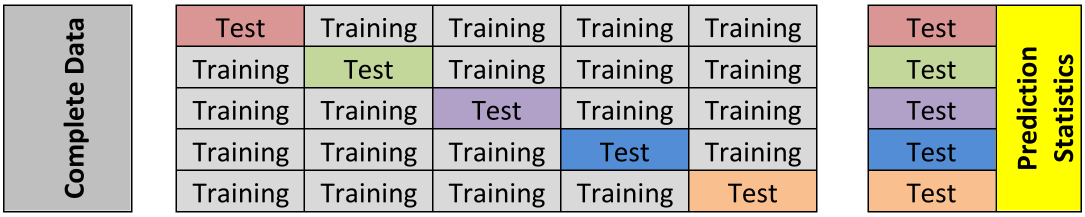


Figure 5: Cross Validation using 5-fold technique

The figure given below explains 95% confidence interval has been used and the results are summarized as given in the section below.

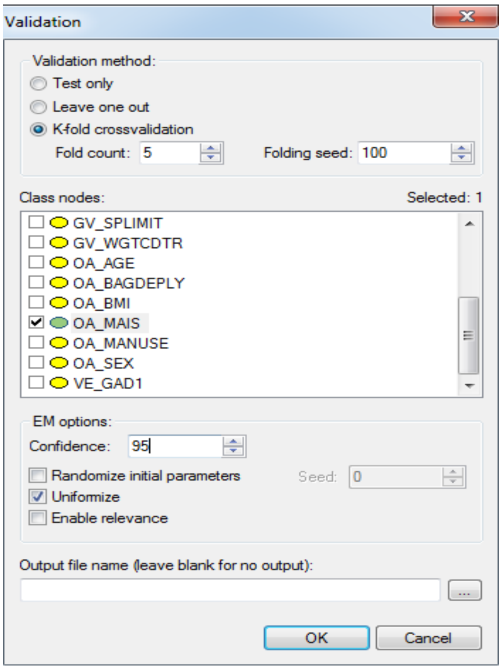


Figure 6: Cross validation utilized in Genie

c. **Accuracy:** Overall accuracy being **83.51%** while the accuracy of individual classes has been shown in the figure below.

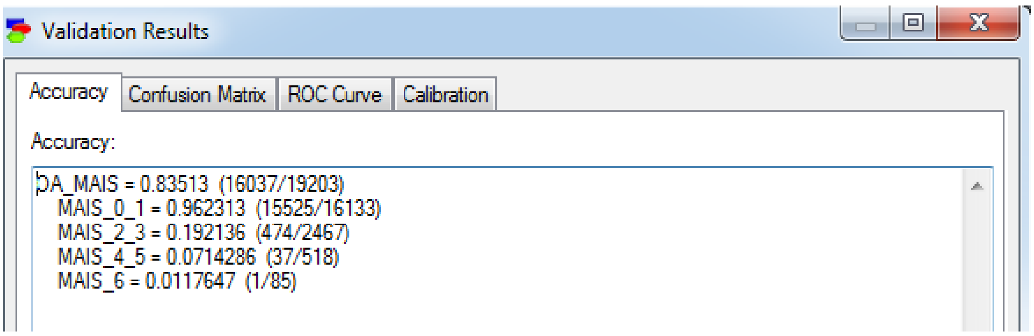
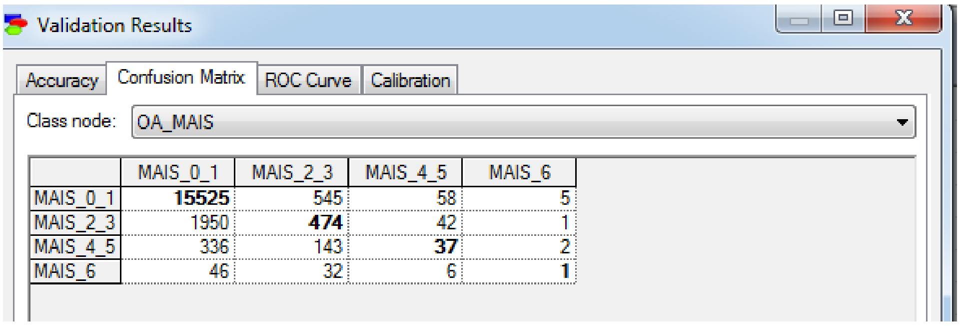


Figure 7: Naive Bayes Accuracy

d. Confusion Matrix: The model can predict the minor injuries and moderate injury with good accuracy and has low accuracy for predicting deaths.



	MAIS_0_1	MAIS_2_3	MAIS_4_5	MAIS_6
MAIS_0_1	**15525**	545	58	5
MAIS_2_3	1950	**474**	42	1
MAIS_4_5	336	143	**37**	2
MAIS_6	46	32	6	**1**

Figure 8: Naive Bayes Confusion Matrix

e. ROC curves

The **receiving operating characteristic** is a measure of classifier performance. Using the proportion of positive data points that are correctly considered as positive and the proportion of negative data points that are mistakenly considered as positive, a graph is generated that shows the trade-off between the rate at which something is predicted correctly with the rate of something predicted incorrectly. The ROC curve for the **Naïve Bayes** model has been shown in Figure below.

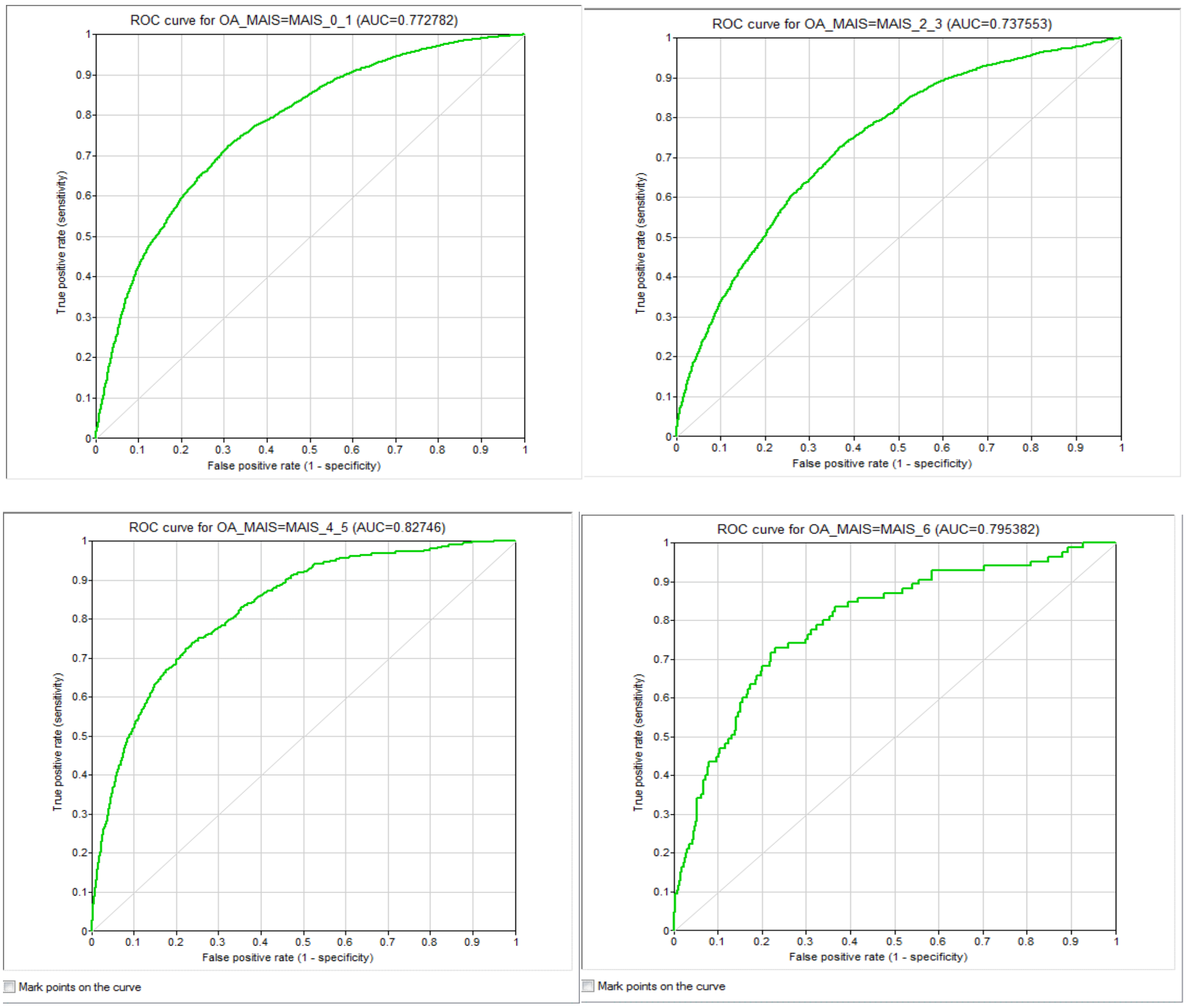


Figure 9: ROC Curve for Naive Bayes

f. Calibration curves

The calibration curves for all the four classes of OA_MAIS variables are very close to ideal calibration curve and have different regions of good and bad calibration results.

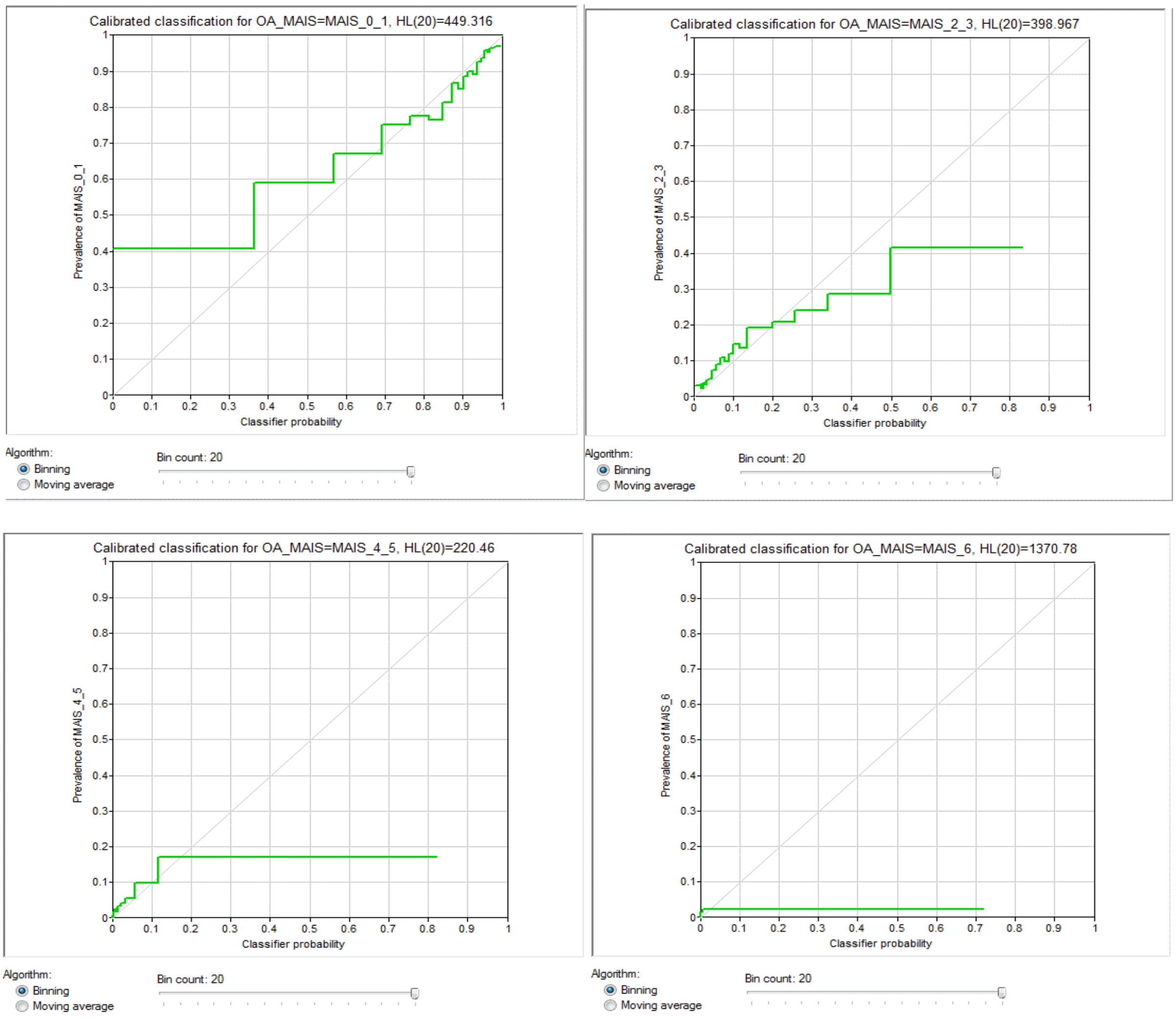


Figure 10: Naive Bayes Calibration curve

➢ TREE AUGMENTED NAÏVE MODEL (TAN)

The **Naive Bayes** model discussed in the previous section, encodes incorrect independence assumptions that, given the class label, the attributes are independent of each other. But in the real world, the attributes of any system are mostly correlated and the case as in Naive Bayes rarely happens. Despite such incorrect independent assumptions, the Naive Bayes model seems to perform well. So, if the model also considers the correlations between the attributes, then the classification accuracy can be improved [1].

a. TAN Model

TAN Model is also created using **Genie** software. The option **Learn New Network** is used for learning new network like **Naïve Bayes** or **TAN**.

Class variable: OA_MAIS

Predictors: All variables along with the correlated variables. Missing values in categorical variables as shown in Table 3 have been removed from the analysis.

The TAN network is shown below: -

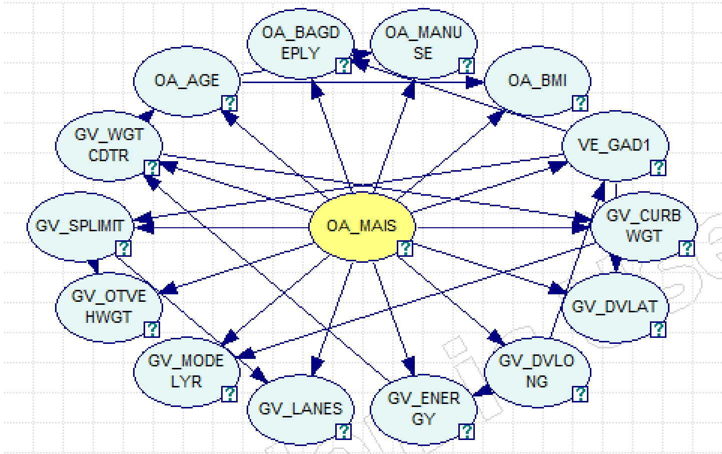


Figure 11: TAN Model

b. Training set and Test set – K fold cross validation

The same concept which has been used in Naïve Bayes has been used for cross validation in TAN also.

c. Accuracy: Overall accuracy 83.64% while the accuracy of individual classes has been shown in the figure below.

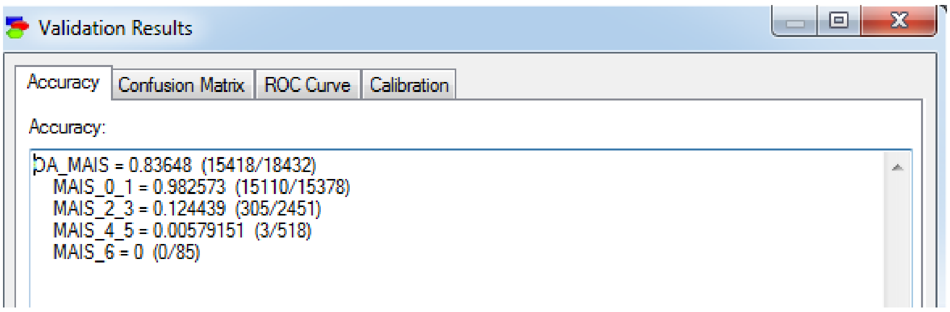
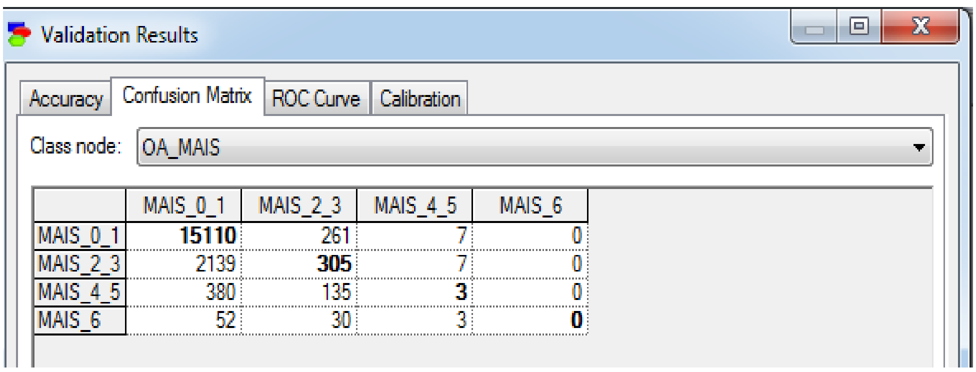


Figure 12: TAN Accuracy results

There is slight improvement in the overall accuracy of TAN model as compared to Naïve Bayes model.

d. Confusion Matrix

The confusion matrix results are somewhat similar to Naïve Bayes with the difference in MAIS_6 prediction results. The model is not able to classify the death class, but the accuracy to class the other three classes is improved as compared to Naïve Bayes.



	MAIS_0_1	MAIS_2_3	MAIS_4_5	MAIS_6
MAIS_0_1	**15110**	261	7	0
MAIS_2_3	2139	**305**	7	0
MAIS_4_5	380	135	**3**	0
MAIS_6	52	30	3	**0**

Figure 13: TAN Confusion Matrix

e. ROC curves

The ROC curves have slightly better AUC values for the different classes of OA_MAIS as compared to Naïve Bayes.

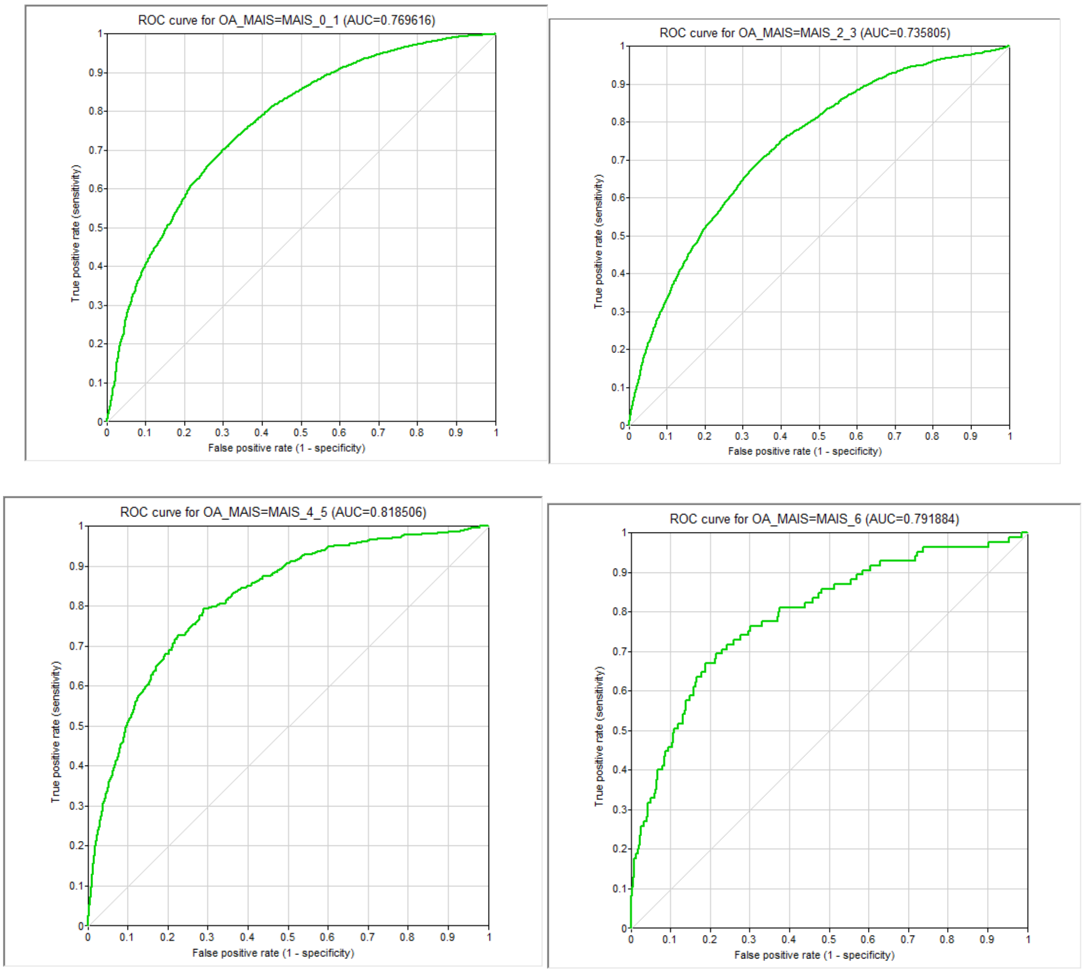
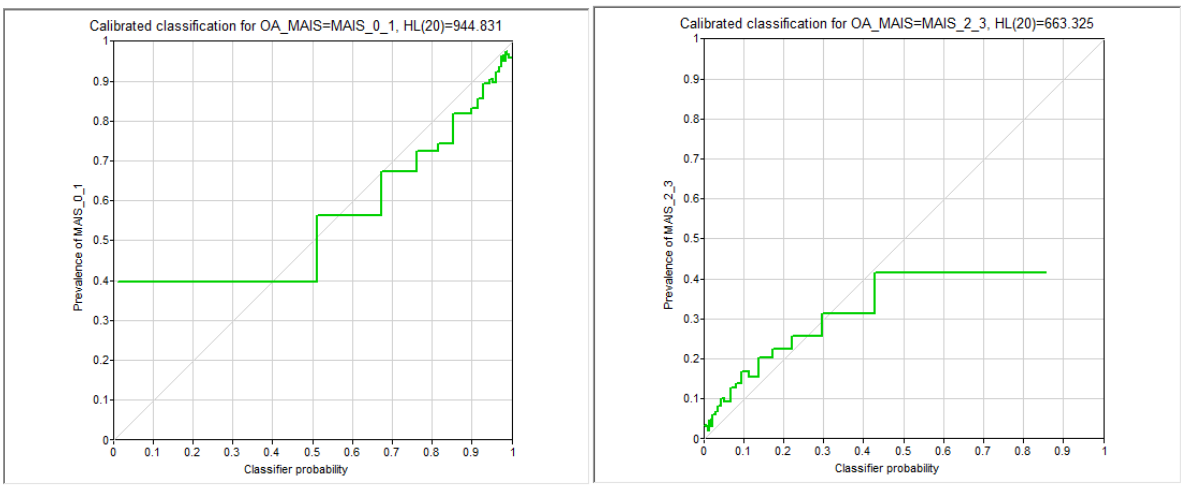


Figure 14: ROC Curve for TAN

f. Calibration curves

The calibration curves for all the four classes of OA_MAIS variables are very close to ideal calibration curve and have different regions of good and bad calibration results.



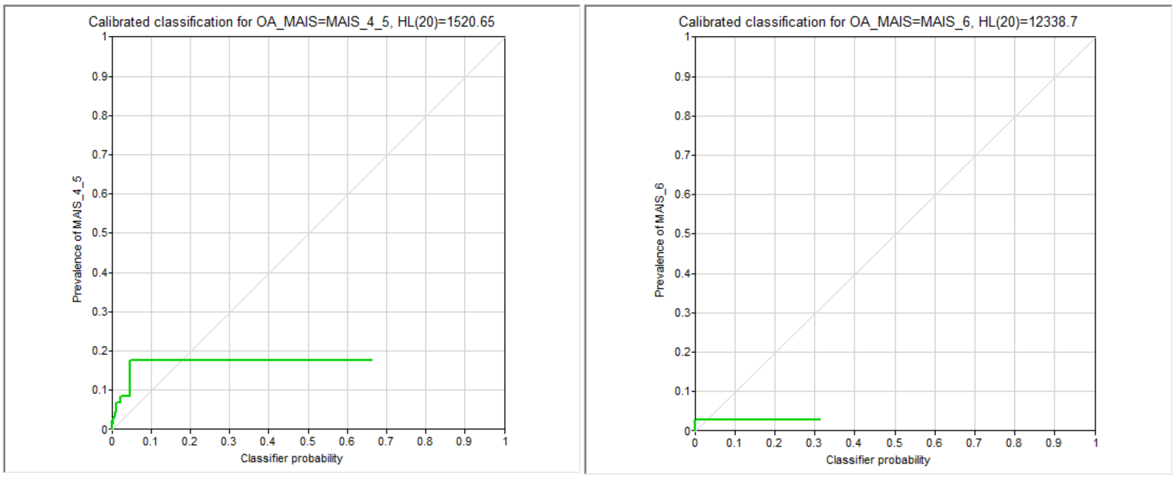


Figure 15: Calibration results for TAN

➢ EXPLORATORY ANALYSIS SUMMARY

- While trying to explore the network using GV_CURBWGT, OA_BMI and OA_AGE as shown in figure 16 below, it was found that for an Old person driving a heavy vehicle, if the BMI of the Old person is normal then his probability of death (MAIS_6) is low. But, if the BMI is high or very low, then the probability of death is high.

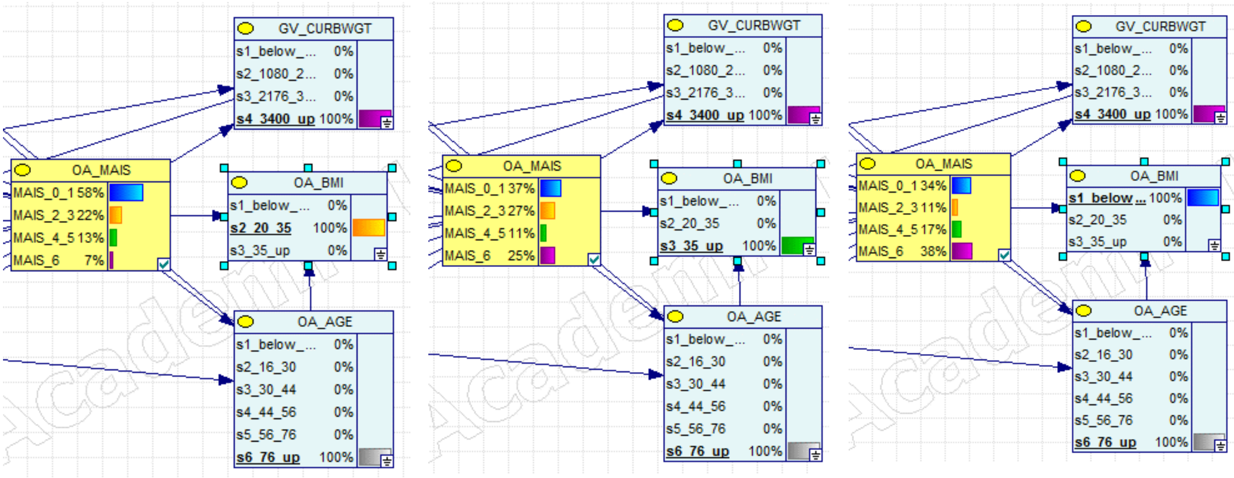


Figure 16: Exploratory analysis - 1

- While trying to explore the network using GV_CURBWGT, GV_LANES and GV_SPLIMIT as shown in figure 17 below, it was found that on a 6 lane road, if the vehicle has lower speed and higher speed then the probability of death rate (MAIS_6) is high and when the speed is medium then probability of death is low.

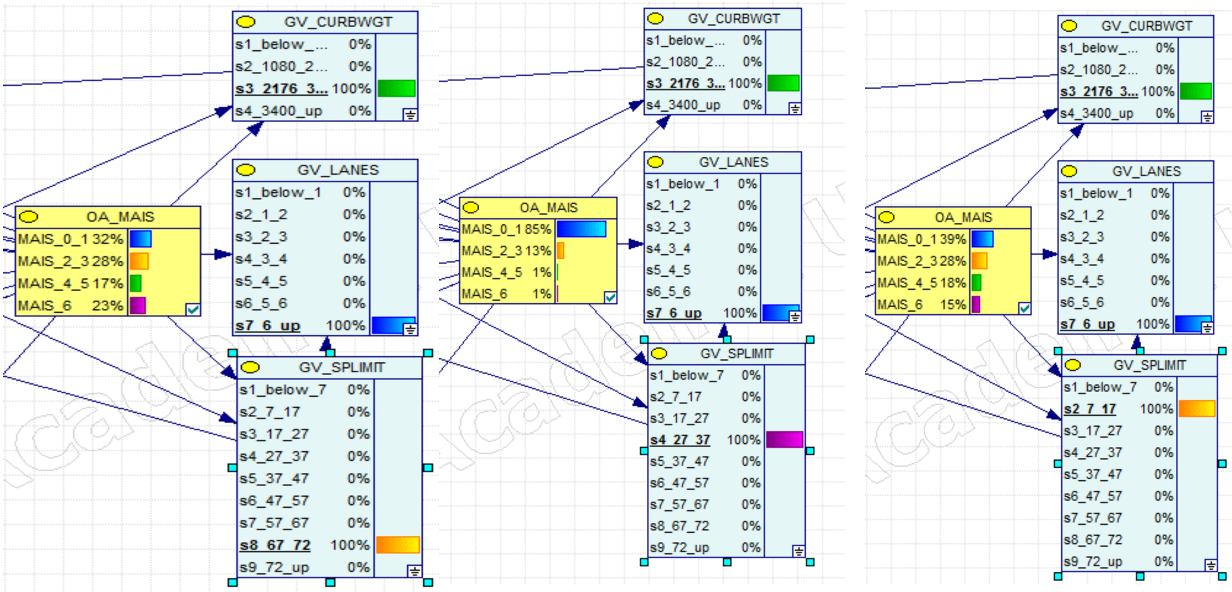


Figure 17: Exploratory analysis - 2

- While trying to explore the network using VE_GAD1, GV_SPLIMIT and GV_MODELYR as shown in figure 18 below, it was found that whenever vehicle is at high speed and the impact is from the left i.e. from the driver side, then if the vehicle is of old model than the probability of death is low and if the vehicle is of new model than the probability of death is high. This suggests that vehicles made these days are not rugged and robust as far safety of driver is concerned.

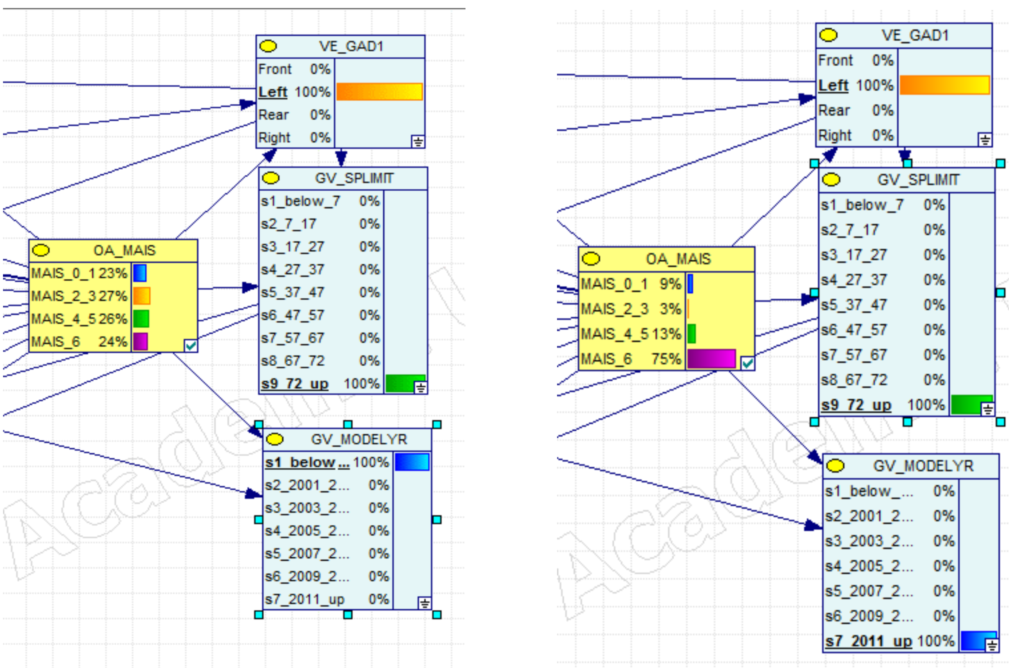


Figure 18: Exploratory analysis - 3

➢ COMPARATIVE ANALYSIS AND CONCLUSION

The **comparative analysis** on the performance of **TAN and NB models** shows that, the **TAN** model **outperforms NB** in this vehicle safety dataset. Even if there are some **correlations** between the variables in the dataset, the **accuracy** results of the **NB** model are **close** to the **TAN** model. This implies that, **considering** the **correlations** between the **variables** in a system, would lead to **better performance**. But we also need to take into consideration the complexity of the model. The **NB** model is very **simple** and **less complex** with almost the **same accuracy results.**

Thus, in this project, by **adding** one level of **dependency** among the **attributes** has given **better accuracy** results for this dataset. Hence, by increasing the level of **interaction** among the **attributes** we can achieve **performance gains**.

➢ REFERENCES

[1] Padmanaban, Harini, "Comparative Analysis of Naive Bayes and Tree Augmented Naive Bayes Models" (2014). Master's Projects. Paper 356.