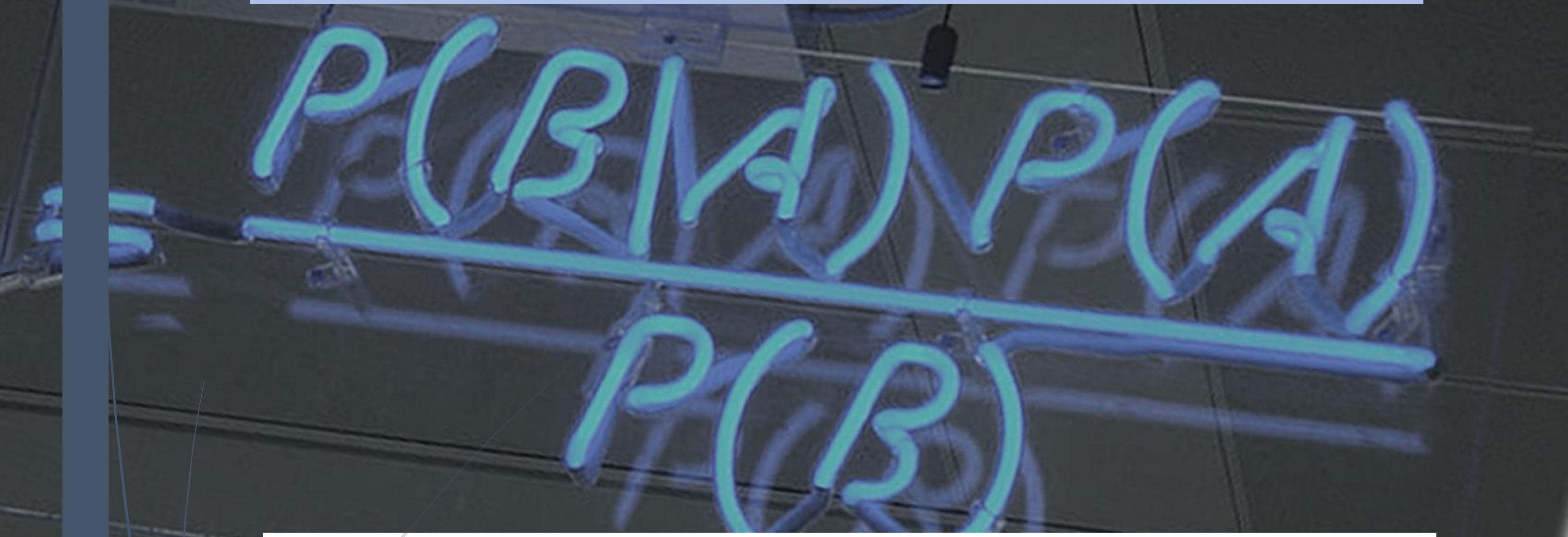
3/22/2018

VEHICLE SAFETY ANALYSIS USING BAYESIAN APPROACH



SUBMITTED TO

Dr. Barry Shepherd

INSITITUTE OF SYSTEMS SCIENCE

NATIONAL UNIVERSITY OF SINGAPORE

PREPARED BY

GOPALAKRISHNAN SAISUBRAMANIAM(A0178249N)
MADAN KUMAR MANJUNATH(A0178237W)
GANANATHAN KHOTEESWARUN(A0178328U)
CHOKKALINGAM SHANMUGASIVA(A0178230J)
ANJALI SINHA(A0178476L)
KENNETH RITHVIK(A0178448M)

MASTER OF TECHNOLOGY IN KNOWLEDGE ENGINEERING
BATCH KE-30(2018)

Table of Contents

1. Business Understanding

1.1. Business Objective

There are a lot of factors which contribute to a vehicle accident and injury incurred by the occupant, but we want to understand the level of injury of the vehicle occupant in terms of vehicle characteristics, accident attributes and occupant's quantitative attributes. Here we do not take into consideration the occupant's driving characteristics. This will give us an insight into the vehicle's susceptibility to accidents.

1.2. Assess Situation

We take into consideration, several vehicle characteristics for understanding the vehicle occupant injury level during and accident, such as vehicle weight and height, braking distance etc. We also look at occupant's quantitative characteristics such as occupant age, weight, height etc., for our prediction.

1.3. Goals

Our goals are to build and compare Bayesian Network prediction models, to predict the likely injury level for vehicle occupants. We will be using GeNIe Bayesian Net to understand the factors which impact vehicle safety.

2. Dataset

The Dataset provided by NASS (National Automotive Sampling Systems) [1] contains attributes which has various parameters to understand the nature of accidents and the fatality rates. Initially there were more than 400 variables, later the selection was limited to 21 variables which were relevant to the variables in EPA/NHSTA research.

VARIABLE NAME	VARIABLE TYPE	VARIABLE DESCRIPTION
GV_CURBWGT	Numerical	Given vehicle's curb weight
GV_DVLAT	Numerical	Vehicle's difference in latitudinal velocity
GV_DVLONG	Numerical	Vehicle's difference in longitudinal velocity
GV_ENERGY	Numerical	Energy absorbed by the Vehicle
GV_LANES	Numerical	Number of lanes where accident took place
GV_MODELYR	Numerical	Model manufacturing year
GV_OTVEHWGT	Numerical	Other vehicle's weight
GV_SPLIMIT	Numerical	Speed Limit in that lane
GV_WGTCDTR	Categorical	Truck weight code
OA_AGE	Numerical	Occupant's age
OA_BAGDEPLY	Categorical	Variable to indicate whether Airbag is deployed or not
OA_HEIGHT	Numerical	Occupant's height
OA_MAIS	Numerical	Occupant's Injury scale
OA_MANUSE	Categorical	Occupant's seat belt was usage

OA_SEX	Categorical	Occupant's gender
OA_WEIGHT	Numerical	Occupant's weight
VE_GAD1	Categorical	Vehicle's deformation
VE_ORIGAVTW	Numerical	Vehicle's average track width
VE_WHEELBAS	Numerical	Vehicle's wheelbase
VE_PDOF_TR	Numerical	Vehicle's principal direction of force
GV_FOOTPRINT	Numerical	Vehicle's footprint

[Table 1: Data Description]

3. Data Pre-processing

The data set considered for building the Naïve Bayes' network is having 20247 records in total against 21 shortlisted columns.

3.1. Data Cleaning : Removal

OA-MAIS which measures the injury is our target variable. It is observed form the dataset that there are 1044 records against the target variable which either have 'NA' or empty values. These records constitute only 5% of the data set and hence removed it by deleting.

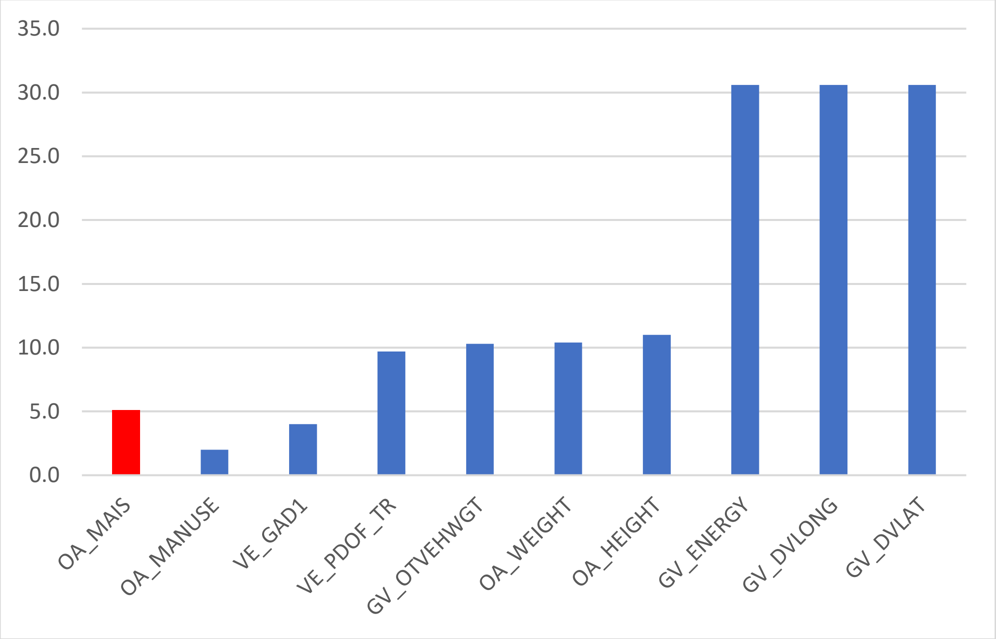


Fig 1. OA_MAIS Missing Data %

3.2. Data Cleaning : Imputation

We analysed all continuous variables' missing data and its pattern in the data set. Below is the plot depicting the same.

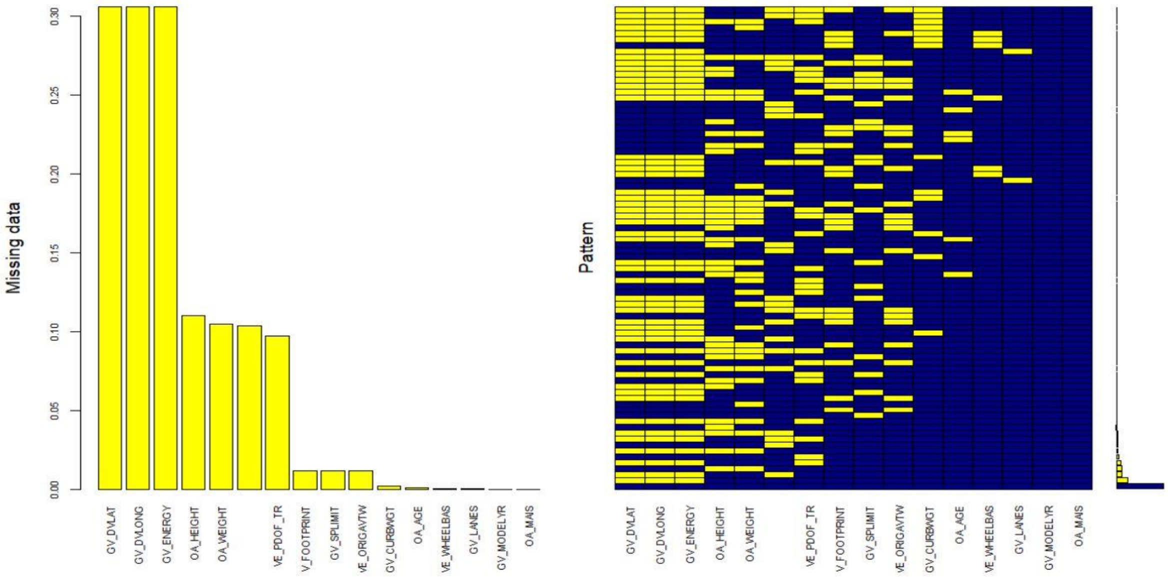


Fig 2. Missing Data and Pattern

From the above histogram we can see that the variables GV_DVLAT, GV_DVLONG and GV_ENERGY are having higher influence of its missing data. We then followed two different approaches to impute the missing values in the data set.

KNN Imputation

KNN imputation uses k-Nearest Neighbours approach to impute missing values. For every observation to be imputed, it identifies 'k' closest observations based on a distance function and computes the weighted average (weighted based on distance) of these 'k' observations.

The VIM package in R contains a function called KNN that makes use of Gowers distance [2] to determine the k nearest neighbours. Gower's distance between two records labelled i and j is defined as

$$d_g(i,j)\frac{\sum_k w_{ijk} d_k(i,j)}{\sum_k w_{ijk}},$$

The sum is over all the variables in the record and d_k (i, j) is the distance is between the k value in record i and record j. The weight $w_{ijk} = 0$, when the kth variable is missing in the record i or j and otherwise 1.

The KNN function determines the k (default: 5) nearest neighbours of a record with missing values. For numerical variables the median of the k nearest neighbours is used as the imputation value, and for categorical variables the mode category in the k nearest neighbours is used.

kNN_Imputation.Rmd

➔ R-Code file for KNN Imputation

IMPUTATION USING PMM (PREDICTIVE MEAN MATCHING)

With MICE (Multivariate Imputation via Chained Equations) [3] package, we create multiple imputations instead of a lone imputation such as mean. This takes care of the uncertainty of the missing data. The missing data are assumed to be missing at random and it imputes the data one variable at a time by following PMM (Predictive Mean Matching) model.

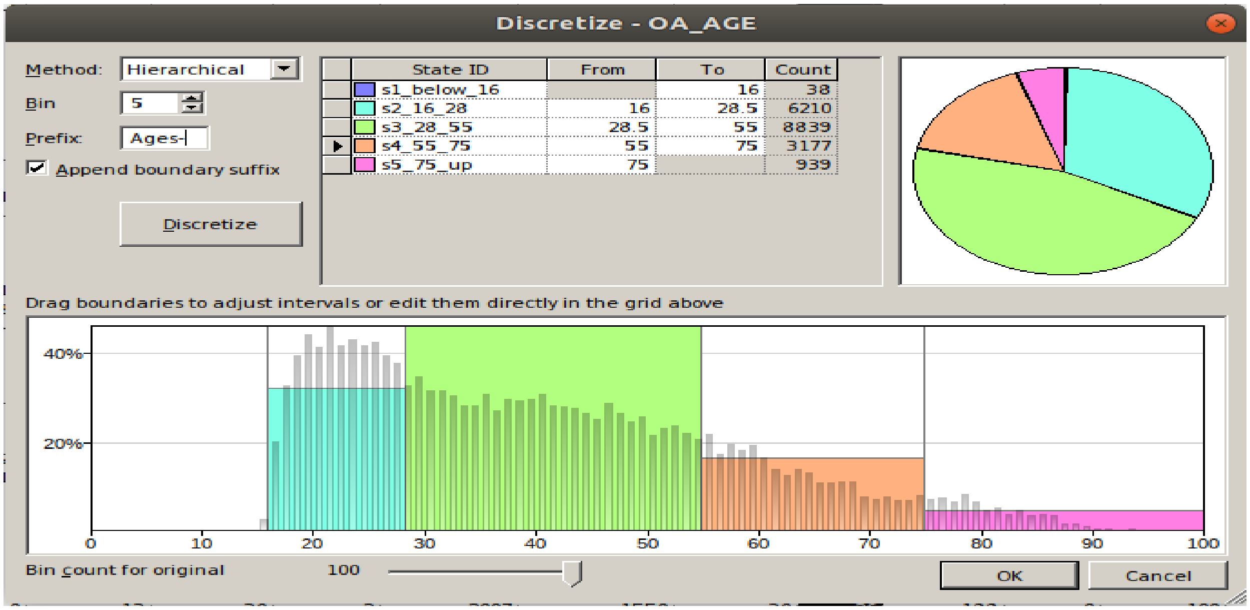
R Code - Cleaning & Imputation.Rmd

➔ R-Code file for PMM Imputation

3.3. DATA BINNING : INPUT VARIABLES

The naive and tree augmented Bayes network that we are building does not accept continuous variables as input, therefore we must group the continuous variables present in our dataset into distinct ranges called bins. We have thirteen continuous variables that we must discretize to bins. We can choose the bin sizes based on either equal number of observations in each bin, equal range for all bins (quartile, percentile, etc.) or in a hierarchical way or we can also use custom bin sizes.

Binning helps us in figuring the outliers and anomalous data in our data-set and to reduce the noise and non-linearity in our data. We use the hierarchical method of binning as it groups together similar data-points that can be categorized under one bin. We later use the customization features in Genie to re-order the ranges based on our preferences. We then assign the label prefix to them to make the ranges more meaningful. In the example below, we have binned the passenger age as described above.



Discretize - OA_AGE

Method: Hierarchical

Bin 5

Prefix: Ages-

☑ Append boundary suffix

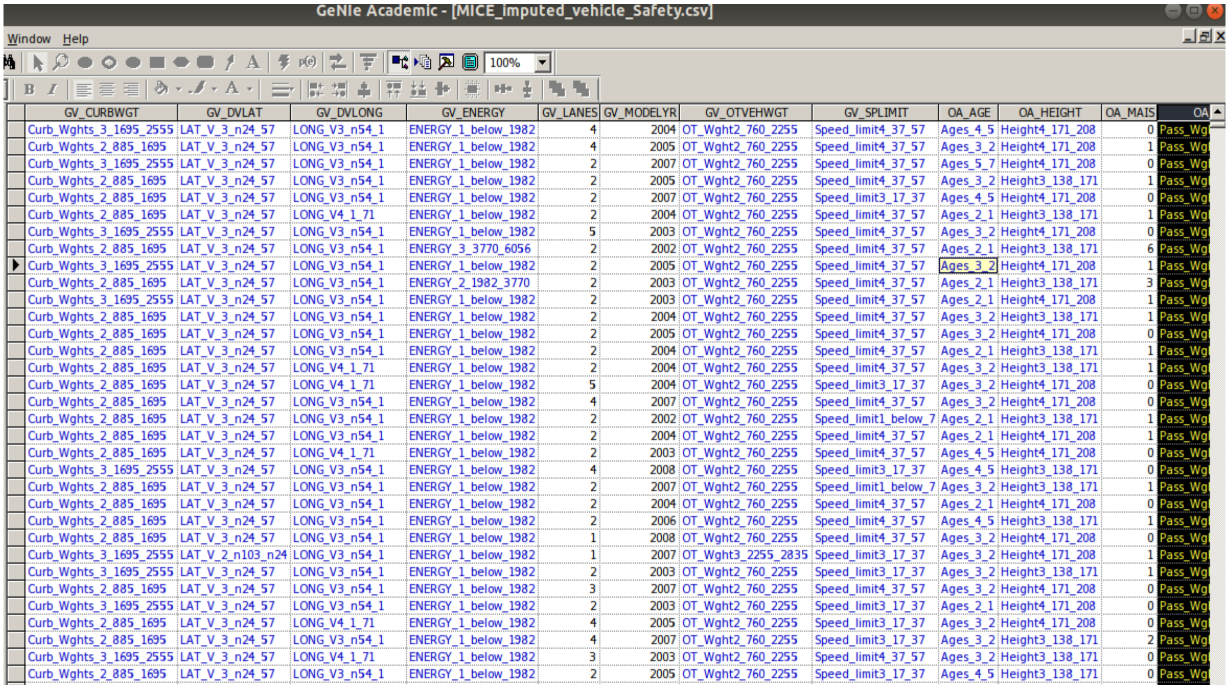
	State ID	From	To	Count
	s1_below_16		16	38
	s2_16_28	16	28.5	6210
	s3_28_55	28.5	55	8839
▶	s4_55_75	55	75	3177
	s5_75_up	75		939

Drag boundaries to adjust intervals or edit them directly in the grid above

Bin count for original 100

Fig 3. Passenger age binned based on hierarchical binning.

After binning all the continuous input variables our data-set attains the following structure as showed in the image below. All the variables have been labelled accordingly.



	GV_CURBWGT	GV_DVLAT	GV_DVLONG	GV_ENERGY	GV_LANES	GV_MODELYR	GV_OTVEHWGT	GV_SPLIMIT	OA_AGE	OA_HEIGHT	OA_MAIS	OA
	Curb_Wghts_3_1695_2555	LAT_V_3_n24_57	LONG_V3_n54_1	ENERGY_1_below_1982	4	2004	OT_Wght2_760_2255	Speed_limit4_37_57	Ages_4_5	Height4_171_208	0	Pass_Wg
	Curb_Wghts_2_885_1695	LAT_V_3_n24_57	LONG_V3_n54_1	ENERGY_1_below_1982	4	2005	OT_Wght2_760_2255	Speed_limit4_37_57	Ages_3_2	Height4_171_208	1	Pass_Wg
	Curb_Wghts_3_1695_2555	LAT_V_3_n24_57	LONG_V3_n54_1	ENERGY_1_below_1982	2	2007	OT_Wght2_760_2255	Speed_limit4_37_57	Ages_5_7	Height4_171_208	0	Pass_Wg
	Curb_Wghts_2_885_1695	LAT_V_3_n24_57	LONG_V3_n54_1	ENERGY_1_below_1982	2	2005	OT_Wght2_760_2255	Speed_limit4_37_57	Ages_3_2	Height3_138_171	1	Pass_Wg
	Curb_Wghts_2_885_1695	LAT_V_3_n24_57	LONG_V3_n54_1	ENERGY_1_below_1982	2	2007	OT_Wght2_760_2255	Speed_limit3_17_37	Ages_4_5	Height4_171_208	0	Pass_Wg
	Curb_Wghts_2_885_1695	LAT_V_3_n24_57	LONG_V4_1_71	ENERGY_1_below_1982	2	2004	OT_Wght2_760_2255	Speed_limit4_37_57	Ages_2_1	Height3_138_171	1	Pass_Wg
	Curb_Wghts_3_1695_2555	LAT_V_3_n24_57	LONG_V3_n54_1	ENERGY_1_below_1982	5	2003	OT_Wght2_760_2255	Speed_limit4_37_57	Ages_3_2	Height4_171_208	0	Pass_Wg
	Curb_Wghts_2_885_1695	LAT_V_3_n24_57	LONG_V3_n54_1	ENERGY_3_3770_6056	2	2002	OT_Wght2_760_2255	Speed_limit4_37_57	Ages_2_1	Height3_138_171	6	Pass_Wg
▶	Curb_Wghts_3_1695_2555	LAT_V_3_n24_57	LONG_V3_n54_1	ENERGY_1_below_1982	2	2005	OT_Wght2_760_2255	Speed_limit4_37_57	Ages_3_2	Height4_171_208	1	Pass_Wg
	Curb_Wghts_2_885_1695	LAT_V_3_n24_57	LONG_V3_n54_1	ENERGY_2_1982_3770	2	2003	OT_Wght2_760_2255	Speed_limit4_37_57	Ages_2_1	Height3_138_171	3	Pass_Wg
	Curb_Wghts_3_1695_2555	LAT_V_3_n24_57	LONG_V3_n54_1	ENERGY_1_below_1982	2	2003	OT_Wght2_760_2255	Speed_limit4_37_57	Ages_2_1	Height4_171_208	1	Pass_Wg
	Curb_Wghts_2_885_1695	LAT_V_3_n24_57	LONG_V3_n54_1	ENERGY_1_below_1982	2	2004	OT_Wght2_760_2255	Speed_limit4_37_57	Ages_3_2	Height3_138_171	1	Pass_Wg
	Curb_Wghts_2_885_1695	LAT_V_3_n24_57	LONG_V3_n54_1	ENERGY_1_below_1982	2	2005	OT_Wght2_760_2255	Speed_limit4_37_57	Ages_3_2	Height4_171_208	0	Pass_Wg
	Curb_Wghts_2_885_1695	LAT_V_3_n24_57	LONG_V3_n54_1	ENERGY_1_below_1982	2	2004	OT_Wght2_760_2255	Speed_limit4_37_57	Ages_2_1	Height3_138_171	1	Pass_Wg
	Curb_Wghts_2_885_1695	LAT_V_3_n24_57	LONG_V4_1_71	ENERGY_1_below_1982	2	2004	OT_Wght2_760_2255	Speed_limit4_37_57	Ages_3_2	Height3_138_171	1	Pass_Wg
	Curb_Wghts_2_885_1695	LAT_V_3_n24_57	LONG_V4_1_71	ENERGY_1_below_1982	5	2004	OT_Wght2_760_2255	Speed_limit3_17_37	Ages_3_2	Height4_171_208	0	Pass_Wg
	Curb_Wghts_2_885_1695	LAT_V_3_n24_57	LONG_V3_n54_1	ENERGY_1_below_1982	4	2007	OT_Wght2_760_2255	Speed_limit4_37_57	Ages_3_2	Height4_171_208	0	Pass_Wg
	Curb_Wghts_2_885_1695	LAT_V_3_n24_57	LONG_V3_n54_1	ENERGY_1_below_1982	2	2002	OT_Wght2_760_2255	Speed_limit1_below_7	Ages_2_1	Height3_138_171	1	Pass_Wg
	Curb_Wghts_2_885_1695	LAT_V_3_n24_57	LONG_V3_n54_1	ENERGY_1_below_1982	2	2004	OT_Wght2_760_2255	Speed_limit4_37_57	Ages_2_1	Height4_171_208	1	Pass_Wg
	Curb_Wghts_2_885_1695	LAT_V_3_n24_57	LONG_V4_1_71	ENERGY_1_below_1982	2	2003	OT_Wght2_760_2255	Speed_limit4_37_57	Ages_4_5	Height4_171_208	0	Pass_Wg
	Curb_Wghts_3_1695_2555	LAT_V_3_n24_57	LONG_V3_n54_1	ENERGY_1_below_1982	4	2008	OT_Wght2_760_2255	Speed_limit3_17_37	Ages_4_5	Height3_138_171	0	Pass_Wg
	Curb_Wghts_2_885_1695	LAT_V_3_n24_57	LONG_V3_n54_1	ENERGY_1_below_1982	2	2007	OT_Wght2_760_2255	Speed_limit1_below_7	Ages_3_2	Height3_138_171	1	Pass_Wg
	Curb_Wghts_2_885_1695	LAT_V_3_n24_57	LONG_V3_n54_1	ENERGY_1_below_1982	2	2004	OT_Wght2_760_2255	Speed_limit4_37_57	Ages_2_1	Height4_171_208	0	Pass_Wg
	Curb_Wghts_2_885_1695	LAT_V_3_n24_57	LONG_V3_n54_1	ENERGY_1_below_1982	2	2006	OT_Wght2_760_2255	Speed_limit4_37_57	Ages_4_5	Height3_138_171	1	Pass_Wg
	Curb_Wghts_2_885_1695	LAT_V_3_n24_57	LONG_V3_n54_1	ENERGY_1_below_1982	1	2008	OT_Wght2_760_2255	Speed_limit4_37_57	Ages_3_2	Height4_171_208	0	Pass_Wg
	Curb_Wghts_3_1695_2555	LAT_V_2_n103_n24	LONG_V3_n54_1	ENERGY_1_below_1982	1	2007	OT_Wght3_2255_2835	Speed_limit3_17_37	Ages_3_2	Height4_171_208	1	Pass_Wg
	Curb_Wghts_3_1695_2555	LAT_V_3_n24_57	LONG_V3_n54_1	ENERGY_1_below_1982	2	2003	OT_Wght2_760_2255	Speed_limit3_17_37	Ages_3_2	Height3_138_171	1	Pass_Wg
	Curb_Wghts_2_885_1695	LAT_V_3_n24_57	LONG_V3_n54_1	ENERGY_1_below_1982	3	2007	OT_Wght2_760_2255	Speed_limit4_37_57	Ages_3_2	Height4_171_208	0	Pass_Wg
	Curb_Wghts_3_1695_2555	LAT_V_3_n24_57	LONG_V3_n54_1	ENERGY_1_below_1982	2	2003	OT_Wght2_760_2255	Speed_limit3_17_37	Ages_2_1	Height4_171_208	0	Pass_Wg
	Curb_Wghts_2_885_1695	LAT_V_3_n24_57	LONG_V4_1_71	ENERGY_1_below_1982	4	2005	OT_Wght2_760_2255	Speed_limit3_17_37	Ages_3_2	Height4_171_208	0	Pass_Wg
	Curb_Wghts_2_885_1695	LAT_V_3_n24_57	LONG_V3_n54_1	ENERGY_1_below_1982	4	2007	OT_Wght2_760_2255	Speed_limit3_17_37	Ages_3_2	Height3_138_171	2	Pass_Wg
	Curb_Wghts_3_1695_2555	LAT_V_3_n24_57	LONG_V4_1_71	ENERGY_1_below_1982	3	2003	OT_Wght2_760_2255	Speed_limit4_37_57	Ages_3_2	Height3_138_171	0	Pass_Wg
	Curb_Wghts_2_885_1695	LAT_V_3_n24_57	LONG_V3_n54_1	ENERGY_1_below_1982	2	2005	OT_Wght2_760_2255	Speed_limit3_17_37	Ages_4_5	Height3_138_171	0	Pass_Wg

Fig 4. Data-set after carrying out Hierarchical and custom binning

We experimented with various other unsupervised binning strategies as well like equal frequency, equal width and employed the Freedman–Diaconis rule in calculating the bin range and size but settled with the hierarchical approach as it gave us optimum results.

3.4. DATA BINNING : OUTPUT VARIABLE

The target variable OA_MAIS, rates injury by body zone and according to relative importance. It is a 6-point ordinal scale. Where the lowest score of 0 indicates a no injury and max score of 6 indicates a severe/fatal injury.

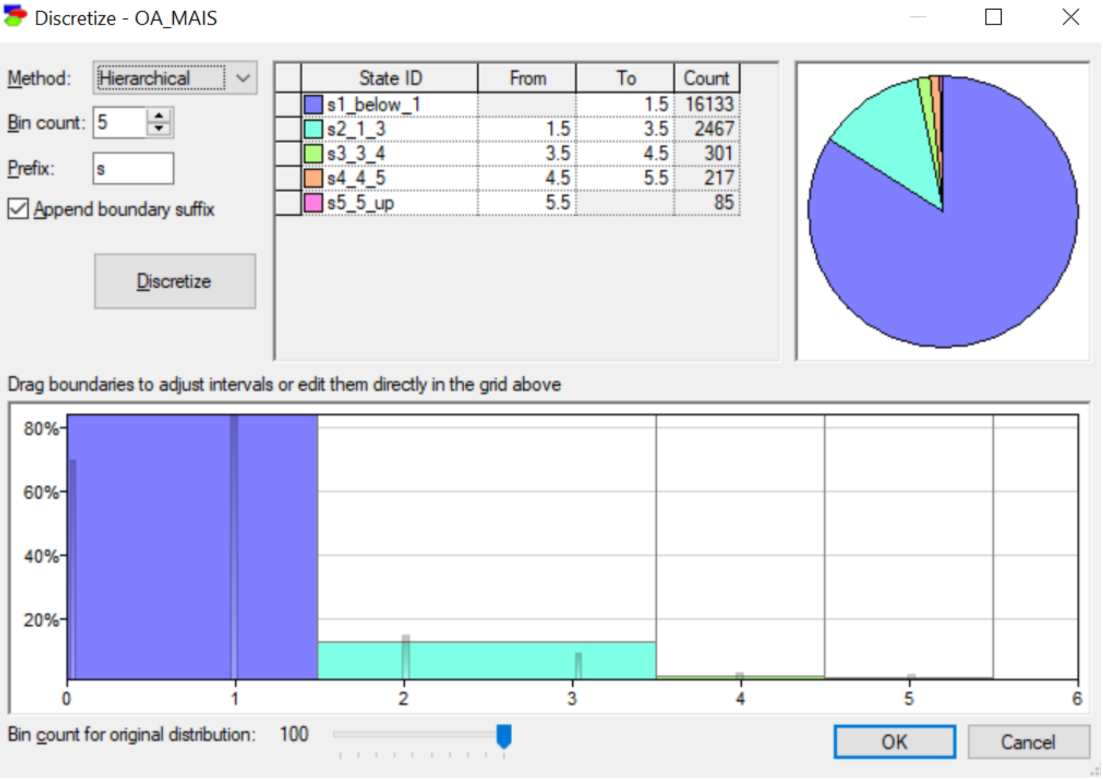
Each injury level has a certain probability of death to it as follows:

No Injury (0)	- NIL
Minor Injury (1)	- 0%
Moderate (2)	- 2%
Serious Injury (3)	- 10%
Severe/Injury (4)	- 50%
Critical/Injury (5)	- 50%
Maximum/Injury (6)	- 100%

So according to the probability of death associated with each injury level, we have binned the data. The bins are as follows:

[0]	Class 1
[1-3]	Class 2
[3-4]	Class 3
[4-5]	Class 4

[5-6] Class 5



Discretize - OA_MAIS

Method: Hierarchical

Bin count: 5

Prefix: s

☑ Append boundary suffix

State ID	From	To	Count
■ s1_below_1		1.5	16133
■ s2_1_3	1.5	3.5	2467
■ s3_3_4	3.5	4.5	301
■ s4_4_5	4.5	5.5	217
■ s5_5_up	5.5		85

Drag boundaries to adjust intervals or edit them directly in the grid above

Bin count for original distribution: 100

Fig 5. Target Variable Binning

4. MODELLING

We used GeNIe to build two types of Bayesian network models.

1) Naïve Bayes Network
2) Tree Augmented Naïve Bayes (TAN) Network

4.1. NAÏVE BAYES

4.1.1. NETWORK DIAGRAM

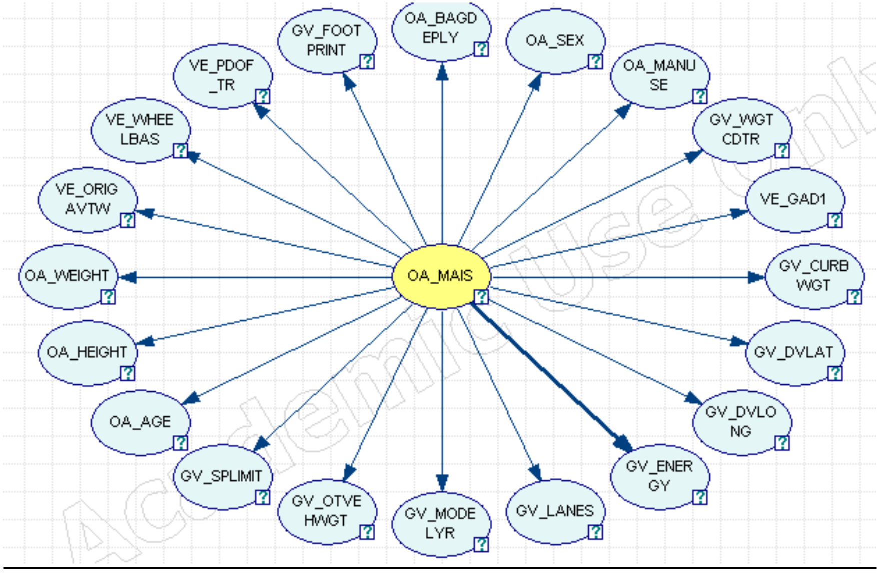
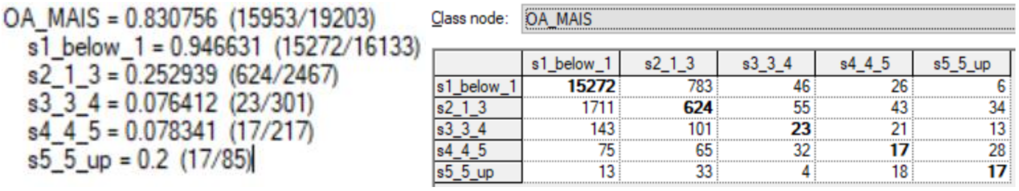


Fig 6: Naïve Bayes Network

4.1.2. Accuracy

The confusion matrix of the naïve bayes network shown below depicts that the overall accuracy of the naïve bayes model is a good 83%. As most of the records are of cases where there is no/minor injury our model is able to predict this class with an accuracy of 94.6%.



OA_MAIS = 0.830756 (15953/19203)
s1_below_1 = 0.946631 (15272/16133)
s2_1_3 = 0.252939 (624/2467)
s3_3_4 = 0.076412 (23/301)
s4_4_5 = 0.078341 (17/217)
s5_5_up = 0.2 (17/85)

Class node: OA_MAIS

	s1_below_1	s2_1_3	s3_3_4	s4_4_5	s5_5_up
s1_below_1	**15272**	783	46	26	6
s2_1_3	1711	**624**	55	43	34
s3_3_4	143	101	**23**	21	13
s4_4_5	75	65	32	**17**	28
s5_5_up	13	33	4	18	**17**

Fig 6: TAN Confusion Matrix

As an alternative to the confusion matrix we can also determine the accuracy by plotting the ROC curve.

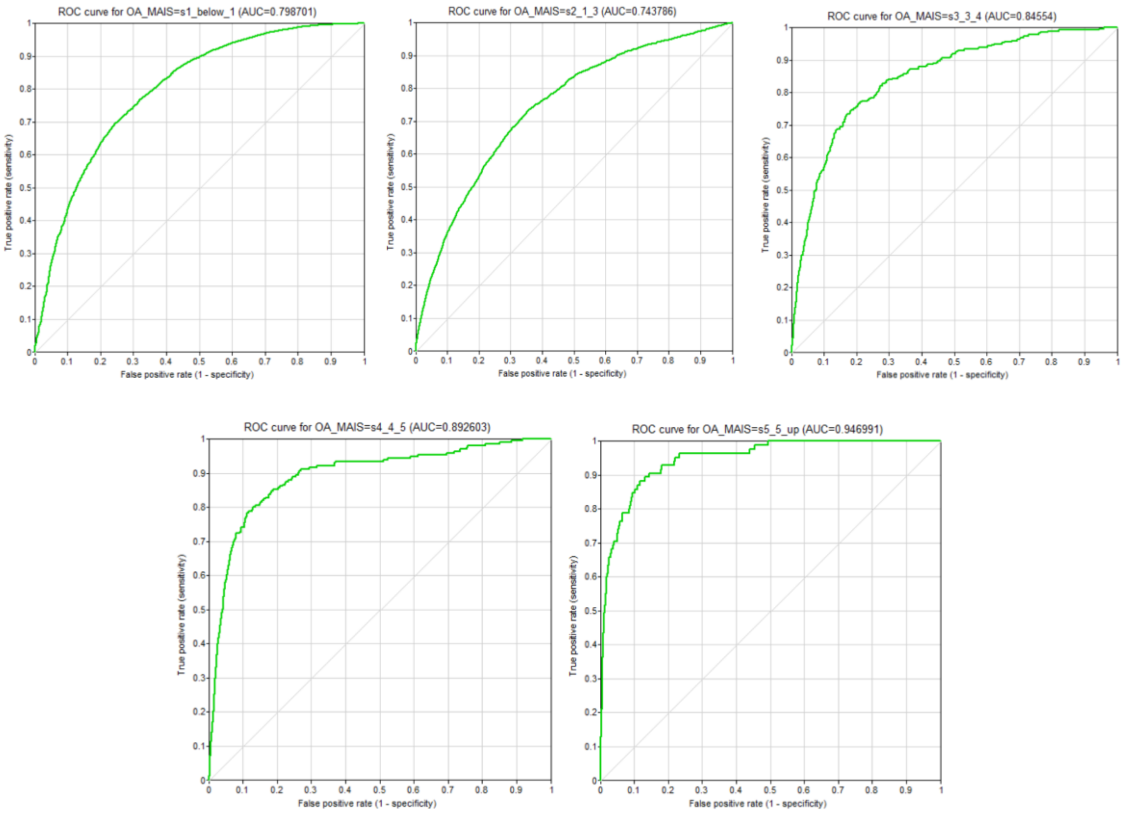


Fig 8: Naïve Bayes ROC Curves

Based on our binning the AUC readings are as below for Naïve Bayes network

Class 1 – 79.8%
Class 2 – 74.4%
Class 3 – 84.6%
Class 4 – 89.3%
Class 5 – 94.7%

4.2. Tree Augmented Naïve Bayes

4.2.1. Network Diagram

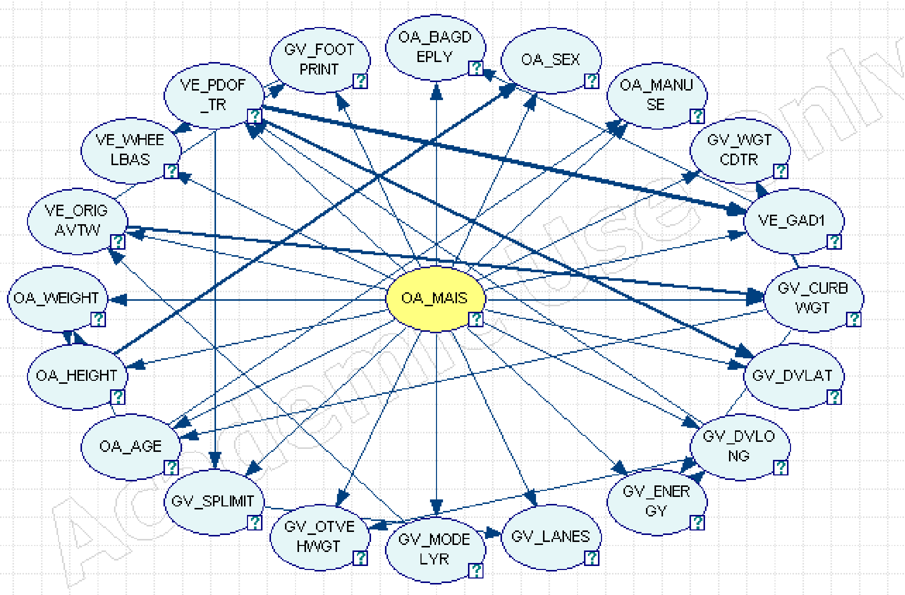
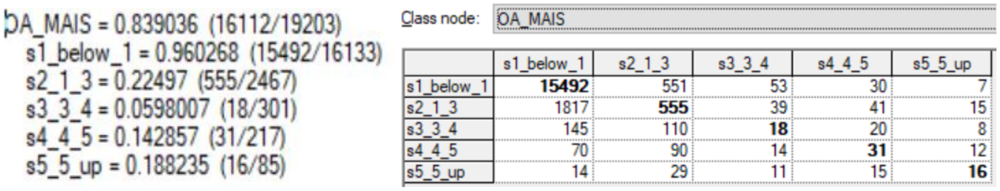


Fig 9: Tree Augmented Naïve Bayes Network

4.2.2. Accuracy

The confusion matrix of the TAN network shown below depicts that the overall accuracy of the TAN model is a good 83.9%. As most of the records are of cases where there is no/minor injury our model is able to predict this class with an accuracy of 96%.



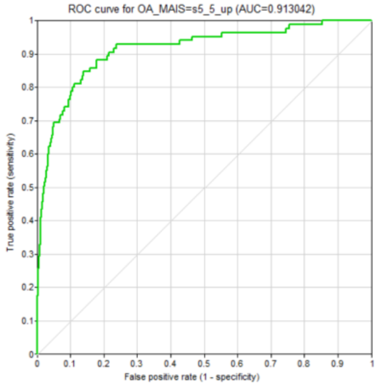
OA_MAIS = 0.839036 (16112/19203)
s1_below_1 = 0.960268 (15492/16133)
s2_1_3 = 0.22497 (555/2467)
s3_3_4 = 0.0598007 (18/301)
s4_4_5 = 0.142857 (31/217)
s5_5_up = 0.188235 (16/85)

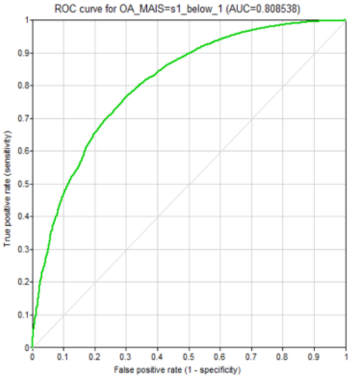
Class node: OA_MAIS

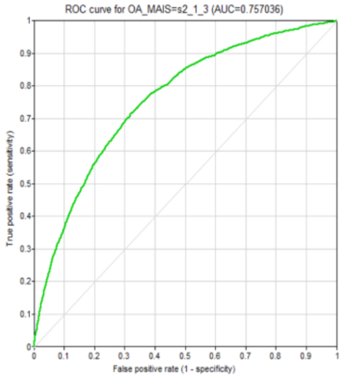
	s1_below_1	s2_1_3	s3_3_4	s4_4_5	s5_5_up
s1_below_1	**15492**	551	53	30	7
s2_1_3	1817	**555**	39	41	15
s3_3_4	145	110	**18**	20	8
s4_4_5	70	90	14	**31**	12
s5_5_up	14	29	11	15	**16**

Fig 10: TAN Confusion Matrix

As an alternative to the confusion matrix we can also determine the accuracy by plotting the ROC curve.







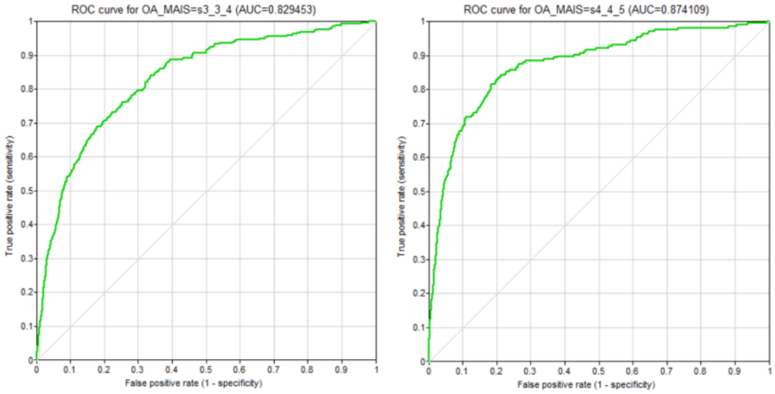


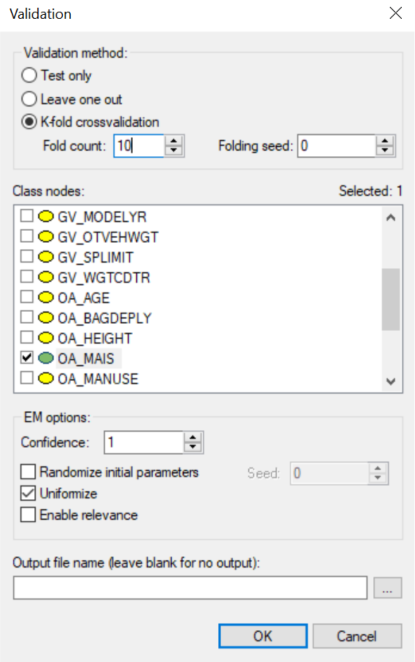
Fig 11: TAN ROC Curves

Based on our binning the AUC readings are as below for TAN network

Class 1 – 80.9%
Class 2 – 75.7%
Class 3 – 82.9%
Class 4 – 87.4%
Class 5 – 91.3%

NETWORK VALIDATION

We used 10-fold cross validation technique while learning the above two network models. It helps in evaluating our predictive models by partitioning the binned data set into training set and test set. The data set will be randomly sampled into 10 equal sized subsamples. Out of these single subsample is retained as test data to evaluate the model and the remaining subsamples are used to train the model. This process is repeated 10 times and the results from this are averaged to give out a single estimation. The advantage of this method is that all the data in the data set are being considered while training as well as testing the model.



Validation

Validation method:
○ Test only
○ Leave one out
● K-fold crossvalidation
Fold count: 10
Folding seed: 0

Class nodes: Selected: 1
□ GV_MODELYR
□ GV_OTVEHWGT
□ GV_SPLIMIT
□ GV_WGTCDTR
□ OA_AGE
□ OA_BAGDEPLY
□ OA_HEIGHT
☑ OA_MAIS
□ OA_MANUSE

EM options:
Confidence: 1
□ Randomize initial parameters
Seed: 0
☑ Uniformize
□ Enable relevance

Output file name (leave blank for no output):

Fig 12: k=10 Fold Cross Validation

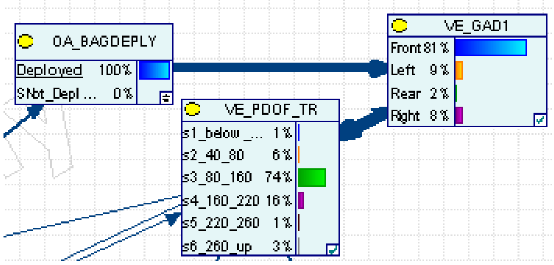
5. NETWORK ANALYSIS AND INSIGHTS

We framed some questions and analysed the network to discover the following interesting insights. All analysis has been made based on the evidences from the TAN network. As we have retained the actual distribution across classes and not performed any up/down sampling, the values represent the actual % split and more cases of class 0 (no injury) are observed.

The following features exhibit influence one another (thicker blue arrows) –

5.1 DIRECTION (DEGREES) FOR PRINCIPAL DIRECTION OF FORCE, DEFORMATION LOCATION AND AIR BAG DEPLOYMENT

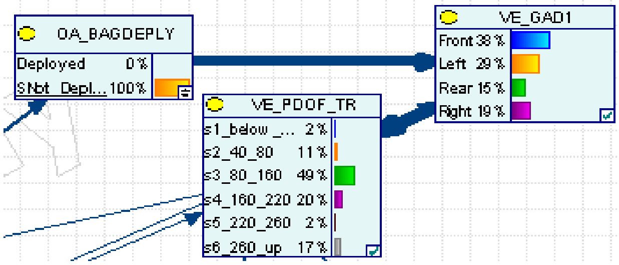
Impact in the front is a major influencer of airbag deployment in the vehicle. In observations where the air bag was not deployed, the impact location is distributed across all directions.


OA_BAGDEPLY
Deployed 100%
SNot_Depl ... 0%

VE_PDOF_TR
s1_below_... 1%
s2_40_80 6%
s3_80_160 74%
s4_160_220 16%
s5_220_260 1%
s6_260_up 3%

VE_GAD1
Front 81%
Left 9%
Rear 2%
Right 8%

Fig 13: Air bag deployed

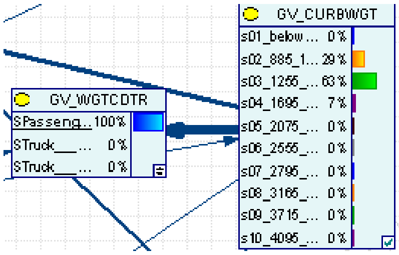

OA_BAGDEPLY
Deployed 0%
SNot_Depl... 100%

VE_PDOF_TR
s1_below_... 2%
s2_40_80 11%
s3_80_160 49%
s4_160_220 20%
s5_220_260 2%
s6_260_up 17%

VE_GAD1
Front 38%
Left 29%
Rear 15%
Right 19%

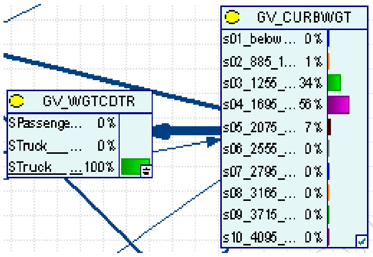
Fig 14: Air bag not deployed

5.2 VEHICLE CURB WEIGHT AND THE TYPE OF VEHICLE (PASSENGER/LIGHT TRUCK/ HEAVY TRUCK)


GV_WGTCDTR
SPasseng... 100%
STruck___... 0%
STruck___... 0%

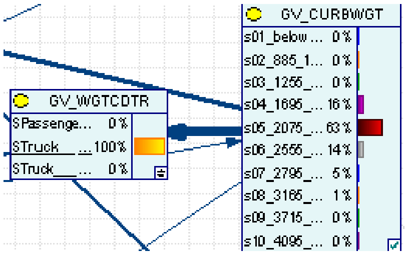
GV_CURBWGT
s01_below... 0%
s02_885_1... 29%
s03_1255_... 63%
s04_1695_... 7%
s05_2075_... 0%
s06_2555_... 0%
s07_2795_... 0%
s08_3165_... 0%
s09_3715_... 0%
s10_4095_... 0%

Fig 15: Passenger vehicle


GV_WGTCDTR
SPassenge... 0%
STruck___... 0%
STruck___...100%

GV_CURBWGT
s01_below... 0%
s02_885_1... 1%
s03_1255_... 34%
s04_1695_... 56%
s05_2075_... 7%
s06_2555_... 0%
s07_2795_... 0%
s08_3165_... 0%
s09_3715_... 0%
s10_4095_... 0%

Fig 16: Light trucks <=6k lbs

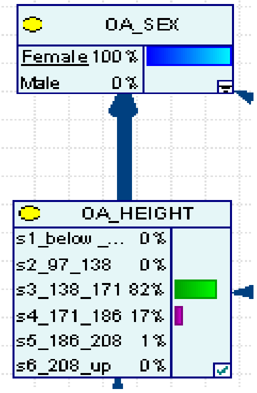

GV_WGTCDTR
SPassenge... 0%
STruck___...100%
STruck___... 0%

GV_CURBWGT
s01_below... 0%
s02_885_1... 0%
s03_1255_... 0%
s04_1695_... 16%
s05_2075_... 63%
s06_2555_... 14%
s07_2795_... 5%
s08_3165_... 1%
s09_3715_... 0%
s10_4095_... 0%

Fig 17: Heavy trucks 6k-10k lbs

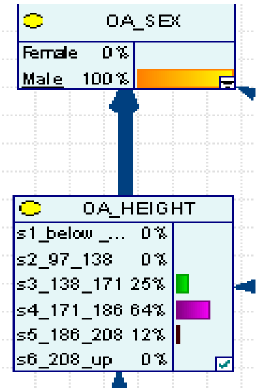
5.3 OCCUPANT SEX AND HEIGHT

The notion that males are generally taller than females holds here.


OA_SEX
Female 100%
Male 0%

OA_HEIGHT
s1_below_... 0%
s2_97_138 0%
s3_138_171 82%
s4_171_186 17%
s5_186_208 1%
s6_208_up 0%

Fig 18: Female height


OA_SEX
Female 0%
Male 100%

OA_HEIGHT
s1_below_... 0%
s2_97_138 0%
s3_138_171 25%
s4_171_186 64%
s5_186_208 12%
s6_208_up 0%

Fig 19: Male height

5.4 If low curb weight passenger vehicles were to be driven at very fast speeds and get hit, is the injury likely to be fatal?

1) Deformation Location:

- Front Section

  With just the above considerations, the chance of facing category 4 and category 5-6 injury levels is 1%. Both male and female appear to have equal split. The airbag was deployed in 68% of the cases and 86% of the time, people wore their seat belts. As such, though it was a head-on collision, there is a very small chance the injury will be fatal.

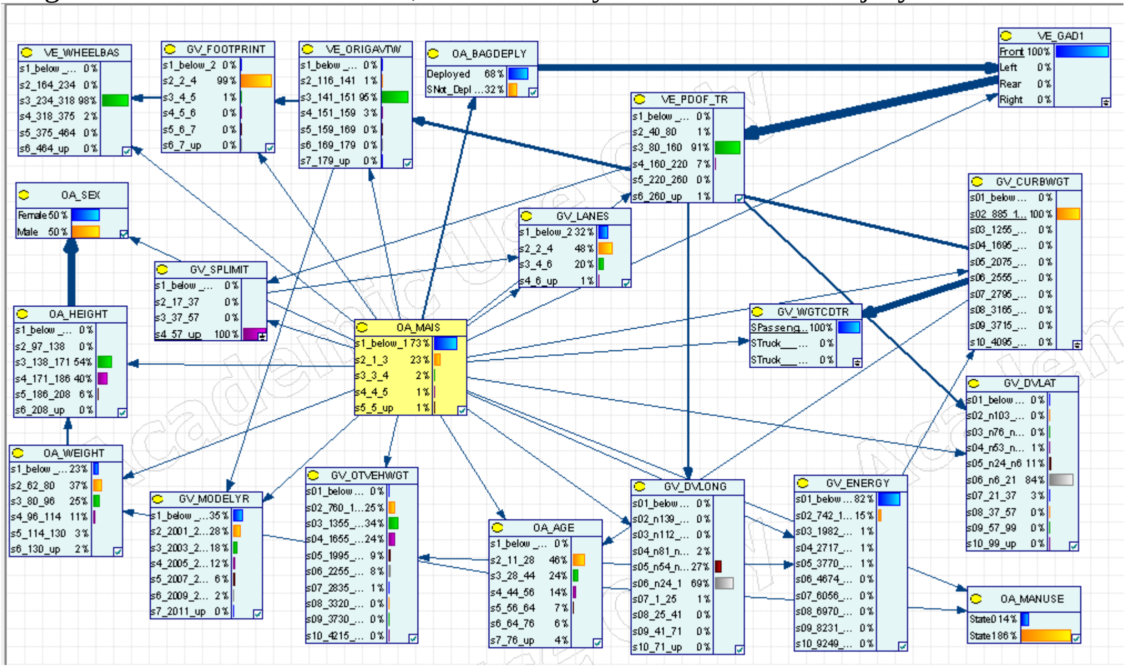


Fig 20: Low curb passenger vehicle driving at high speed and accident impact in front location

If we consider the observations where airbags deployed all the time and seatbelts were not worn, then the probability of minor injuries goes up from 23% to 45% and severe, critical and maximum injuries rise to 3-5%.

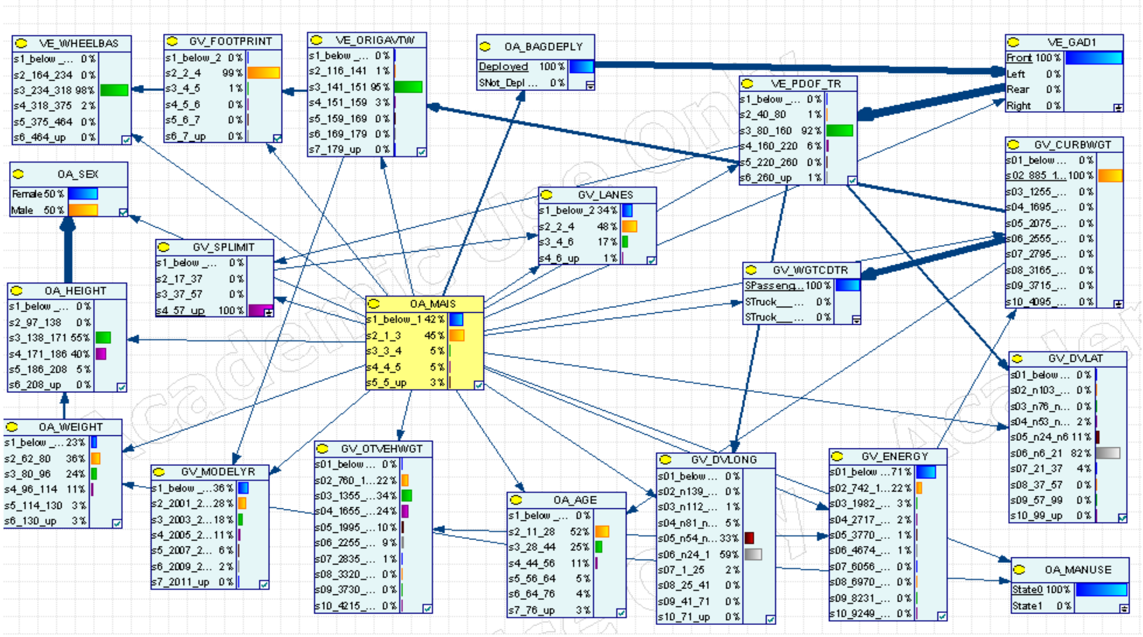


Fig 21: Low curb passenger vehicle driving at high speed. accident impact in front location - seat belt worn & airbag deployed

- <u>LEFT SECTION</u>

  When we shift our focus to the impact in the left direction, i.e. where the driver is seated, the distribution of injury increases towards higher levels. The accident took place perpendicular to the direction of travel. Category 4 reads 17%, more than 3x than when the damage was in the front.

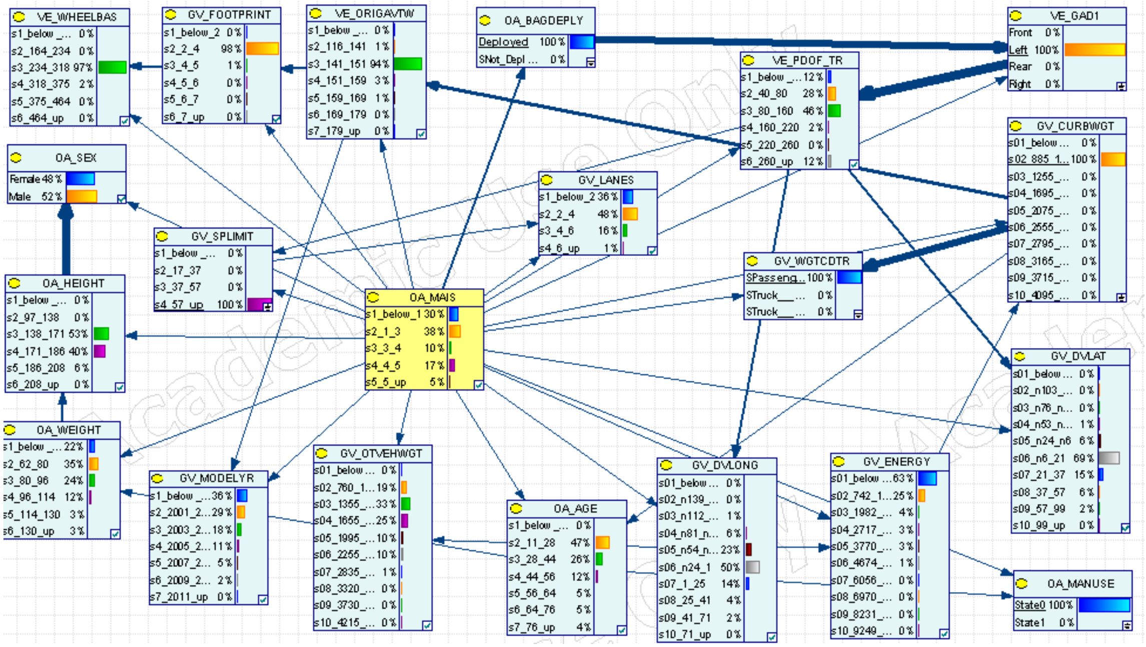


Fig 22: Low curb passenger vehicle driving at high speed. accident impact in left location - seat belt worn & airbag deployed

- <u>RIGHT AND REAR SECTION</u>

  An accident impact in the right section has slightly higher chance of being severe (7%) as compared to the rear section (5%), while in the rear section, close to half of the occupants (49%) faced minor injuries with 1% of the population being unlucky of survival.

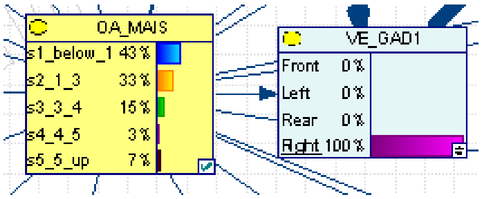


Fig 23: Impact of Right Section

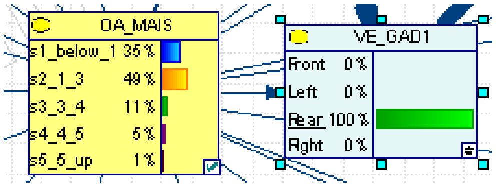
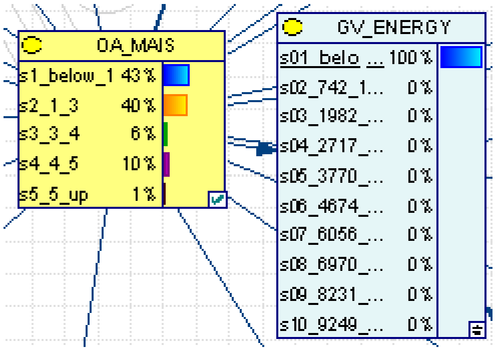


Fig 24: Impact of Rear Section

2) Impact of damage- Energy Absorption

   There is no clear definition of Energy Absorption (GV_ENERGY) in the documentation or reference, so our assumption is that the column indicates the amount of damage dealt and absorbed by the vehicle body during the accident.

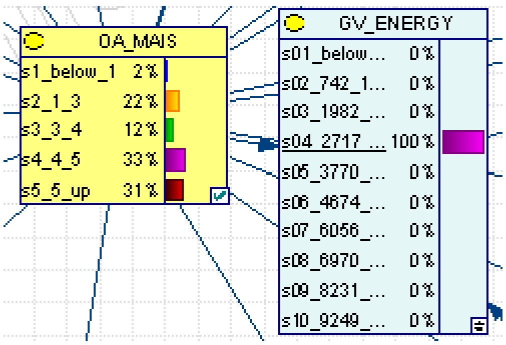
   A trend is observed – the increasing value of this factor forces the distribution of injury level falls towards the danger zone.



```
OA_MAIS
s1_below_1 43%
s2_1_3     40%
s3_3_4      6%
s4_4_5     10%
s5_5_up     1%

GV_ENERGY
s01_belo ... 100%
s02_742_1...   0%
s03_1982_...   0%
s04_2717_...   0%
s05_3770_...   0%
s06_4674_...   0%
s07_6056_...   0%
s08_6970_...   0%
s09_8231_...   0%
s10_9249_...   0%
```

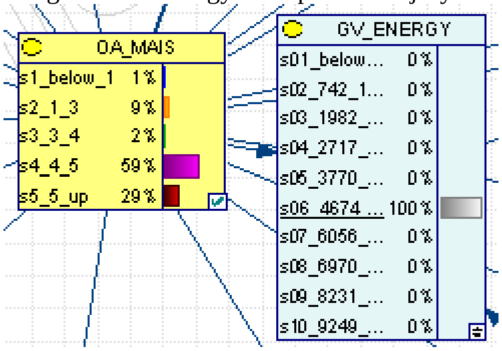
Fig 25: Low energy absorption vs injury level



```
OA_MAIS
s1_below_1  2%
s2_1_3     22%
s3_3_4     12%
s4_4_5     33%
s5_5_up    31%

GV_ENERGY
s01_below...   0%
s02_742_1...   0%
s03_1982_...   0%
s04_2717 ... 100%
s05_3770_...   0%
s06_4674_...   0%
s07_6056_...   0%
s08_6970_...   0%
s09_8231_...   0%
s10_9249_...   0%
```

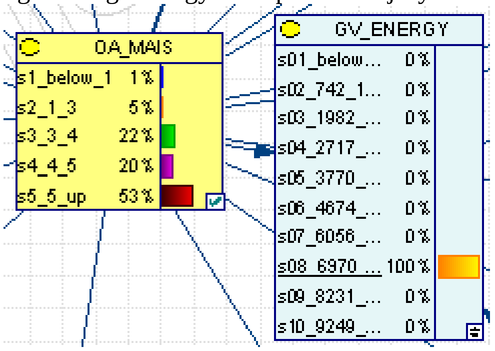
Fig 26: High energy absorption vs injury level



```
OA_MAIS
s1_below_1  1%
s2_1_3      9%
s3_3_4      2%
s4_4_5     59%
s5_5_up    29%

GV_ENERGY
s01_below...   0%
s02_742_1...   0%
s03_1982_...   0%
s04_2717_...   0%
s05_3770_...   0%
s06_4674 ... 100%
s07_6056_...   0%
s08_6970_...   0%
s09_8231_...   0%
s10_9249_...   0%
```

Fig 27: Medium energy absorption vs injury level

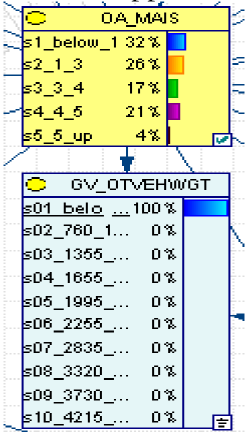


```
OA_MAIS
s1_below_1  1%
s2_1_3      5%
s3_3_4     22%
s4_4_5     20%
s5_5_up    53%

GV_ENERGY
s01_below...   0%
s02_742_1...   0%
s03_1982_...   0%
s04_2717_...   0%
s05_3770_...   0%
s06_4674_...   0%
s07_6056_...   0%
s08_6970 ... 100%
s09_8231_...   0%
s10_9249_...   0%
```

Fig 28: Very high energy absorption vs injury level

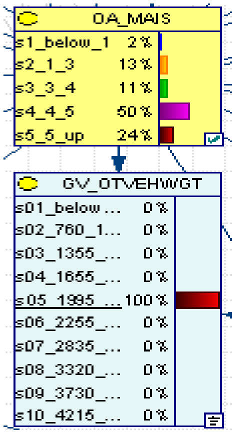
3) Strength of the other vehicle – curb weight
The weight of the other vehicle involved in the accident is also a factor in determining the impact of injury faced. Lighter vehicles cause minor injuries and damage while heavier vehicles appear to cause 51-53% of the critical and maximum injuries.



```
OA_MAIS
s1_below_1 32%
s2_1_3     26%
s3_3_4     17%
s4_4_5     21%
s5_5_up     4%

GV_OTVEHWGT
s01_belo ... 100%
s02_760_1...   0%
s03_1355_...   0%
s04_1655_...   0%
s05_1995_...   0%
s06_2255_...   0%
s07_2835_...   0%
s08_3320_...   0%
s09_3730_...   0%
s10_4215_...   0%
```

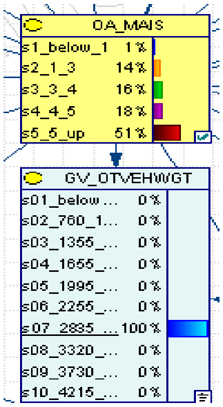
Fig 29: Other vehicle – low curb weight



```
OA_MAIS
s1_below_1  2%
s2_1_3     13%
s3_3_4     11%
s4_4_5     50%
s5_5_up    24%

GV_OTVEHWGT
s01_below...   0%
s02_760_1...   0%
s03_1355_...   0%
s04_1655_...   0%
s05_1995 ... 100%
s06_2255_...   0%
s07_2835_...   0%
s08_3320_...   0%
s09_3730_...   0%
s10_4215_...   0%
```

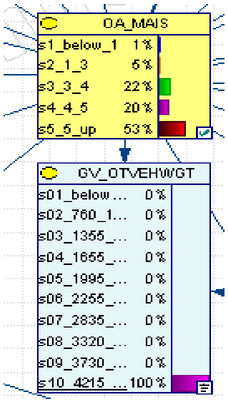
Fig 30: Other vehicle – moderate curb weigh



```
OA_MAIS
s1_below_1  1%
s2_1_3     14%
s3_3_4     16%
s4_4_5     18%
s5_5_up    51%

GV_OTVEHWGT
s01_below...   0%
s02_760_1...   0%
s03_1355_...   0%
s04_1655_...   0%
s05_1995_...   0%
s06_2255_...   0%
s07_2835 ... 100%
s08_3320_...   0%
s09_3730_...   0%
s10_4215_...   0%
```

Fig 31: Other vehicle – heavy curb weight



```
OA_MAIS
s1_below_1  1%
s2_1_3      5%
s3_3_4     22%
s4_4_5     20%
s5_5_up    53%

GV_OTVEHWGT
s01_below...   0%
s02_760_1...   0%
s03_1355_...   0%
s04_1655_...   0%
s05_1995_...   0%
s06_2255_...   0%
s07_2835_...   0%
s08_3320_...   0%
s09_3730_...   0%
s10_4215 ... 100%
```

Fig 32: Other vehicle – very heavy curb weight

5.5 WHICH GENDER HAS HIGHER PROBABILITY TO FACE MAJOR INJURIES? ARE SEATBELTS USEFUL IN MINIMIZING THE INTENSITY OF INJURY?

With the evidence set for different bins of injury level (OA_MAIS), the observations are as follows:

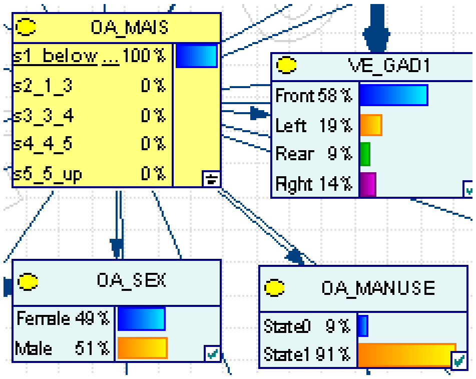


Fig 33: Injury level 0

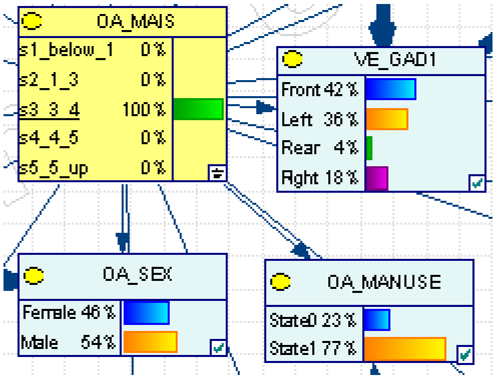


Fig 34: Injury level 3

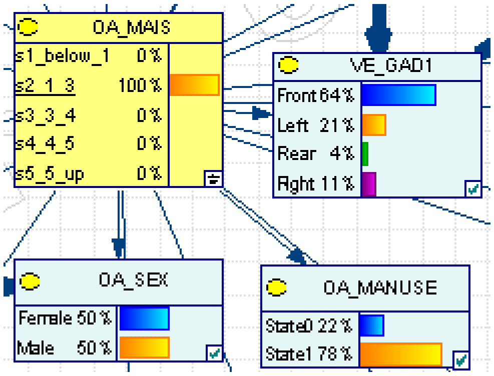


Fig 35: Injury level 1 and 2

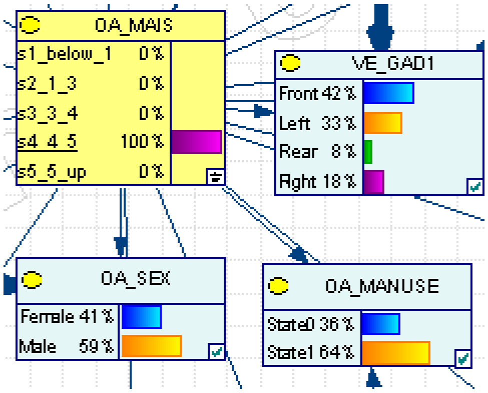


Fig 36: Injury level 4

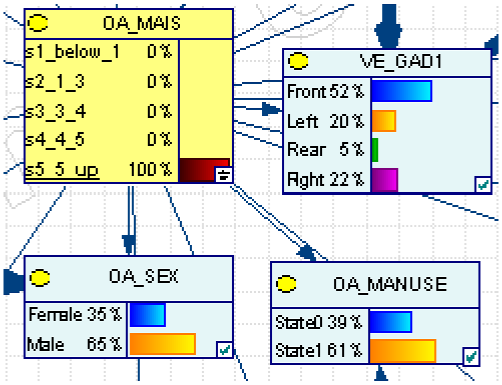


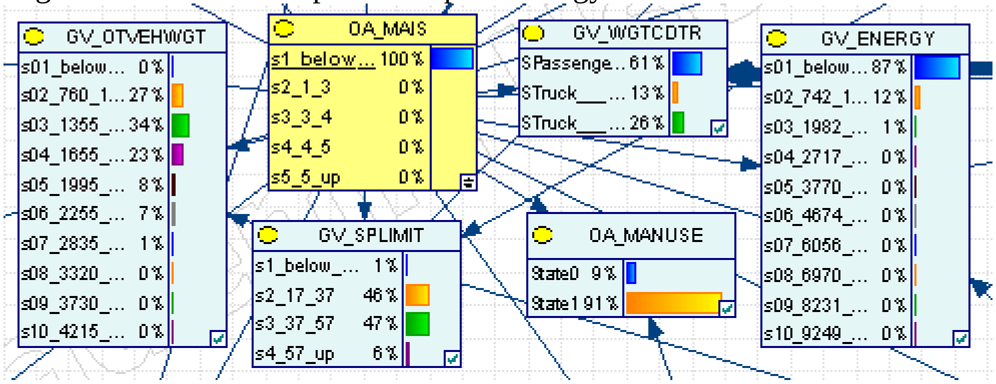
Fig 37: Injury level 5 and above

The ratio of Male : Female with no or minor injuries is equal (50-50) %. As the intensity of injury level is more, males are more prone to suffer from such damages. The percentage shifts from 50 to 54-59% for serious and severe injuries and up to 65% for critical and maximum/fatal injuries. Consequently, the decrease in the percentage of people wearing seat belts (State0 in OA_MANUSE) is an added factor, i.e. injuries are less likely to be above serious if the occupant wears a seat belt 77% of the time.

5.6 WHAT FACTORS CONTRIBUTE TO NO INJURY IN THE EVENT OF AN ACCIDENT?

In the event of an accident, the following factors have shown to prevent injuries –

- 91% of the people wore seat belts at the time of accident
- The average speed limit was around 17-57 mph
- 61% of the occupants, despite being in passenger vehicles did not suffer injuries
- The weight of the other vehicle in the accident was predominately between 760-1995 kg and the collision impact/absorption energy was below 742 J for 87% of the cases.



GV_OTVEHWGT
s01_below… 0%
s02_760_1… 27%
s03_1355_… 34%
s04_1655_… 23%
s05_1995_… 8%
s06_2255_… 7%
s07_2835_… 1%
s08_3320_… 0%
s09_3730_… 0%
s10_4215_… 0%

OA_MAIS
s1_below… 100%
s2_1_3 0%
s3_3_4 0%
s4_4_5 0%
s5_5_up 0%

GV_WGTCDTR
$Passenge… 61%
$Truck___… 13%
$Truck___… 26%

GV_ENERGY
s01_below… 87%
s02_742_1… 12%
s03_1982_… 1%
s04_2717_… 0%
s05_3770_… 0%
s06_4674_… 0%
s07_6056_… 0%
s08_6970_… 0%
s09_8231_… 0%
s10_9249_… 0%

GV_SPLIMIT
s1_below_… 1%
s2_17_37 46%
s3_37_57 47%
s4_57_up 6%

OA_MANUSE
State0 9%
State1 91%

Fig 38: Factors influencing no injury

6. Conclusion

We compared two types of Bayesian networks – i. Naïve Bayes, ii. Tree Augmented Network (TAN) and reported the results in the modelling section. The NB model assumes independence between attributes given the class, whereas TAN (belonging to the augmented network type) considers dependence between its class and other attributes. As seen from the network and its analysis there are some features with mutual influence and this information captured by TAN is reflected in its slight performance improvement. We also posed some questions and using Exploratory Network Analysis, found some insights on the different factors influencing vehicle safety.

7. References

1. Bayesian Lab
2. C. Gower. A general coefficient of similarity and some of its properties. Biometrics, 27:857--874, 1971.
3. PMM Imputation