Bayesian Networks

Butterfly Effects

By

Bharat Nagaraju (A0178258N)

Edwin Tam (A0178396J)

Vigneshram Andiappan Selvaraj (A0178215A)

Contents

[Domain & Approaches 1](#_Toc6088142)

[Just what are the factors impact vehicle safety? 1](#_Toc6088143)

[Preparing the Data 3](#_Toc6088144)

[Our Network Models 8](#_Toc6088145)

[How Good Are Our Network Models? 12](#_Toc6088146)

# Domain & Approaches

## **Just what are the factors impact vehicle safety?**

We are given a car crash dataset from Bayesia Website to clean, review, and infer from using 2 Bayesian Networks.

The dataset contains 19 usable features, of which OA\_MAIS is the target variable (it measures injury severity). Before we describe the preparation, network generation, and results etc. – let’s have a looksee into the dataset. We’ll use a mixture of Genie & Excel to get the work done.

**Dataset Shape:** 21 columns X 20,247 observations

**Continuous Variables**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Mean** | **Variance** | **StdDev** | **Min** | **Max** | **Count** | **Missing %** | **Remarks** |
| GV\_CURBWGT | 1617.3 | **154900.5** | 393.6 | 670 | 4310 | 20204 | 0.2% |  |
| GV\_DVLAT | 0.0 | 169.5 | 13.0 | -114 | 118 | 14049 | 30.6% |  |
| GV\_DVLONG | -14.8 | 311.8 | 17.7 | -145 | 84 | 14049 | 30.6% |  |
| GV\_ENERGY | 505.2 | **416986.1** | 645.7 | 4 | 9852 | 14049 | 30.6% |  |
| GV\_LANES | 3.3 | 1.8 | 1.4 | 1 | 7 | 20244 | 0.0% |  |
| GV\_MODELYR | 2003.6 | 7.6 | 2.8 | 2000 | 2012 | 20247 | 0.0% |  |
| GV\_OTVEHWGT | 1630.2 | **169212.7** | 411.4 | 640 | 4540 | 18147 | 10.4% |  |
| GV\_SPLIMIT | 40.7 | 126.4 | 11.2 | 0 | 75 | 20016 | 1.1% |  |
| OA\_AGE | 40.2 | 301.8 | 17.4 | 0 | 97 | 20190 | 0.3% |  |
| OA\_HEIGHT | 170.8 | 115.5 | 10.7 | 59 | 216 | 17508 | 13.5% |  |
| OA\_MAIS | 0.9 | 1.1 | 1.0 | 0 | 6 | 19203 | 5.2% |  |
| OA\_MANUSE | 0.9 | 0.1 | 0.3 | 0 | 1 | 19774 | 2.3% |  |
| OA\_WEIGHT | 78.7 | 385.9 | 19.6 | 28 | 150 | 17599 | 13.1% |  |
| VE\_ORIGAVTW | 154.7 | 58.6 | 7.7 | 105 | 185 | 20014 | 1.2% | Not Used |
| VE\_WHEELBAS | 281.0 | 824.7 | 28.7 | 141 | 481 | 20238 | 0.0% | Not Used |
| VE\_PDOF\_TR | 152.6 | 4557.2 | 67.5 | 5 | 355 | 18298 | 9.6% |  |
| GV\_FOOTPRINT | 4.4 | 0.4 | 0.6 | 2.4684 | 7.7952 | 20010 | 1.2% |  |

**Categorical Variables**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **GV\_WGTCDTR** | **Count** | **Percentage** |  | **OA\_BAGDEPLY** | **Count** | **Percentage** |
| Passenger Car | 12474 | 62% | Deployed | 9593 | 47% |
| Truck (<=10000 lbs.) | 2542 | 13% | Not Deployed | 10654 | 53% |
| Truck (<=6000 lbs.) | 5231 | 26% |  |  |  |
|  | | | | | | |
| **OA\_SEX** | **Count** | **Percentage** |  | **VE\_GAD1** | **Count** | **Percentage** |
| (missing) | 234 | 1% | (missing) | 789 | 4% |
| Female | 9938 | 49% | Front | 11793 | 58% |
| Male | 10075 | 50% | Left | 3145 | 16% |
|  |  |  | Rear | 1741 | 9% |
|  |  |  | Right | 2779 | 14% |

# Preparing the Data

The Dataset provided by NASS (National Automotive Sampling Systems) 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.

The rest of this section covers Data Cleaning, Imputation, and Binning.

**Data Cleaning**

The data set contained 20,247 records. OA-MAIS which measures the injury is our target variable.

It is observed form the dataset that there are 1,044 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.

|  |  |
| --- | --- |
| #Missing Values for Each of the columns before cleaning | |
|  |  |
|  |  |

From the above charts, GV\_DVLAT, GV\_DVLONG and GV\_ENERGY have a maximum amount of missing data. We will impute to balance the missing data

**Data Imputation**

We analyzed all continuous variables’ missing data and its pattern in the data set. Below is the plot depicting the same. Assuming the data is missing at Random, we decide to impute the data using PMM using ‘MICE’ package in R.

|  |
| --- |
| Breadth of Missing data |
|  |

|  |  |  |
| --- | --- | --- |
| Top 3 Variables – Where is the missing data?  Overlay the values of predictor when the values are missing and when they are not. | | |
|  |  |  |

**IMPUTATION USING PMM (PREDICTIVE MEAN MATCHING)**

Using [MICE](https://cran.r-project.org/web/packages/mice/index.html) package, we create multiple imputations instead of single imputation techniques 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.

**DATA BINNING: INPUT VARIABLES**

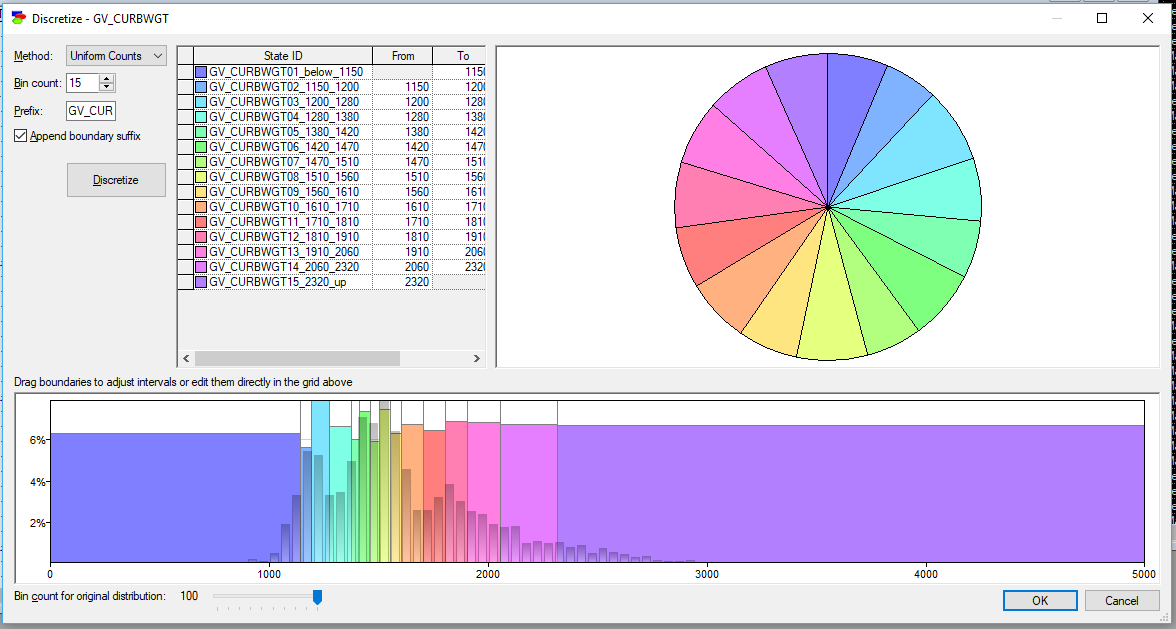
The naive and tree augmented Bayes network that we are building does not accept continuous variables as input, therefore we must group the continuous variables present in our dataset into distinct ranges called bins. We have thirteen continuous variables that we must discretize to bins. We can choose 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 have used a mixture of hierarchical, uniform width and based on range based on the distribution of the variable.

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 below.

GV\_CURBWGT is prefixed for all the bins created

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.



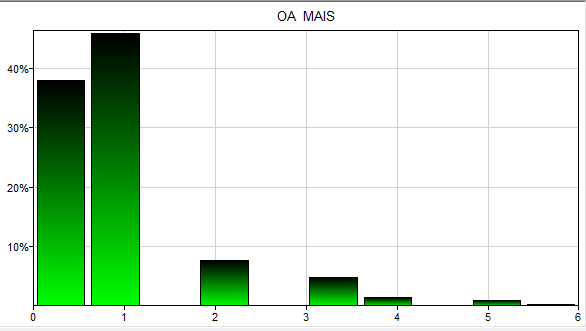
Input Bins are mostly in the range of 5-10 because the predictor Y-Variable has almost about 5 classes or bins.

**BINNING for all Variables**

|  |  |
| --- | --- |
| GV\_CURBWGT | GC\_DVLAT |
| GV\_DVLONG | GV\_ENERGY |
| GV\_MODELYR | GV\_Lanes |
| GV\_OTVEHWGT | GV\_SPLIMIT |
| OA\_AGE | OA\_HEIGHT |
| OA\_WEIGHT | VE\_PDOF\_TR |
| VE\_ORIGAVTW | VE\_WHEELBAS |
| GV\_FOOTPRINT | Data Set Snapshot |

**DATA BINNING: OUTPUT VARIABLE**

Distribution of injury levels



The target variable OA\_MAIS, rates injury by body zone and according to relative importance. Scored on a 6-point scale, 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%

Based on the distribution, we have categorized the bins into 5 levels as shown in the fig

# Our Network Models

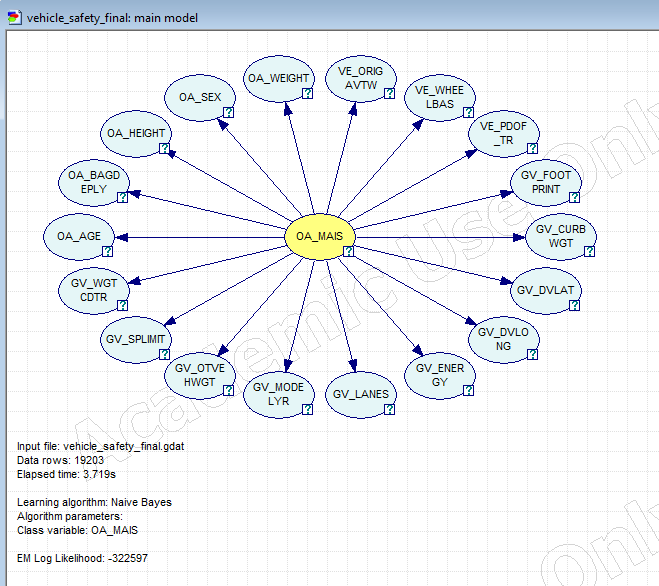
We used GeNIe to build two types of Bayesian network models.

1) Naïve Bayes Network

2) Tree Augmented Naïve Bayes (TAN) Network

**NAIVE BAYES**

NETWORK DIAGRAM



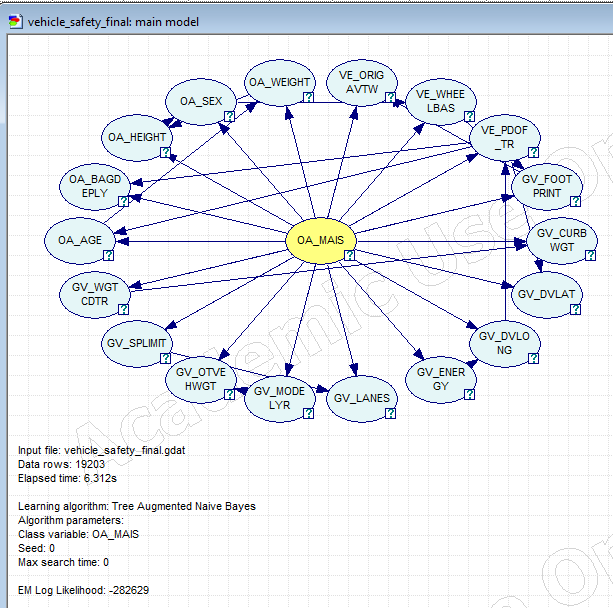
ACCURACY

|  |  |
| --- | --- |
| Confusion Matrix | Accuracy |
|  | OA\_MAIS = 0.839192 (16115/19203)  s1\_below\_1 = 0.971239 (15669/16133)  s2\_1\_3 = 0.165788 (409/2467)  s3\_3\_4 = 0.0199336 (6/301)  s4\_4\_5 = 0.0967742 (21/217)  s5\_5\_up = 0.117647 (10/85) |

|  |  |
| --- | --- |
| Area under the Curve (AUC) | |
|  |  |
|  |  |
|  |  |

**TREE AUGMENTED NAÏVE BAYES**

NETWORK DIAGRAM

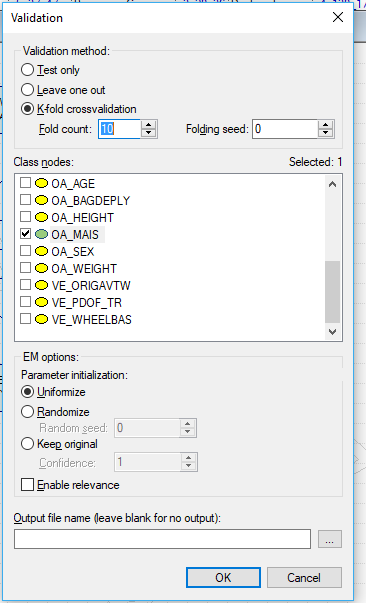


**ACCURACY**

|  |  |
| --- | --- |
| Confusion Matrix | Accuracy |
|  | OA\_MAIS = 0.847628 (16277/19203)  s1\_below\_1 = 0.977933 (15777/16133)  s2\_1\_3 = 0.160519 (396/2467)  s3\_3\_4 = 0.116279 (35/301)  s4\_4\_5 = 0.211982 (46/217)  s5\_5\_up = 0.270588 (23/85) |

|  |  |
| --- | --- |
| **Area under the Curve (AUC)** | |
|  |  |
|  |  |
|  |  |

**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 subsamples 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.

# How Good Are Our Network Models?

Both networks had similar accuracy rates (about 0.8). They do very well in predicting class 0; however, they do shockingly worse for the other classes. Only TAN had a slight edge in accuracy in other classes.

We suspect it is the dataset distribution – most of the observations were skewed towards class 1. For the rest of this section, we will use the TAN Network.

Intuitively, there are a number of questions that we use to interrogate the network.

1. Would deployed airbags result in lesser injury?
2. Would greater vehicle footprint result in greater injury?
3. Which gender is likely to face greater injury?
4. For each injury state, what is likely to cause it?
5. **Would deployed airbags result in lesser injury?**

|  |  |
| --- | --- |
|  | Oddly enough it doesn’t. In fact, it increases the likelihood of severe injuries (See red boxes). Drivers seem to be better off without air bags. |

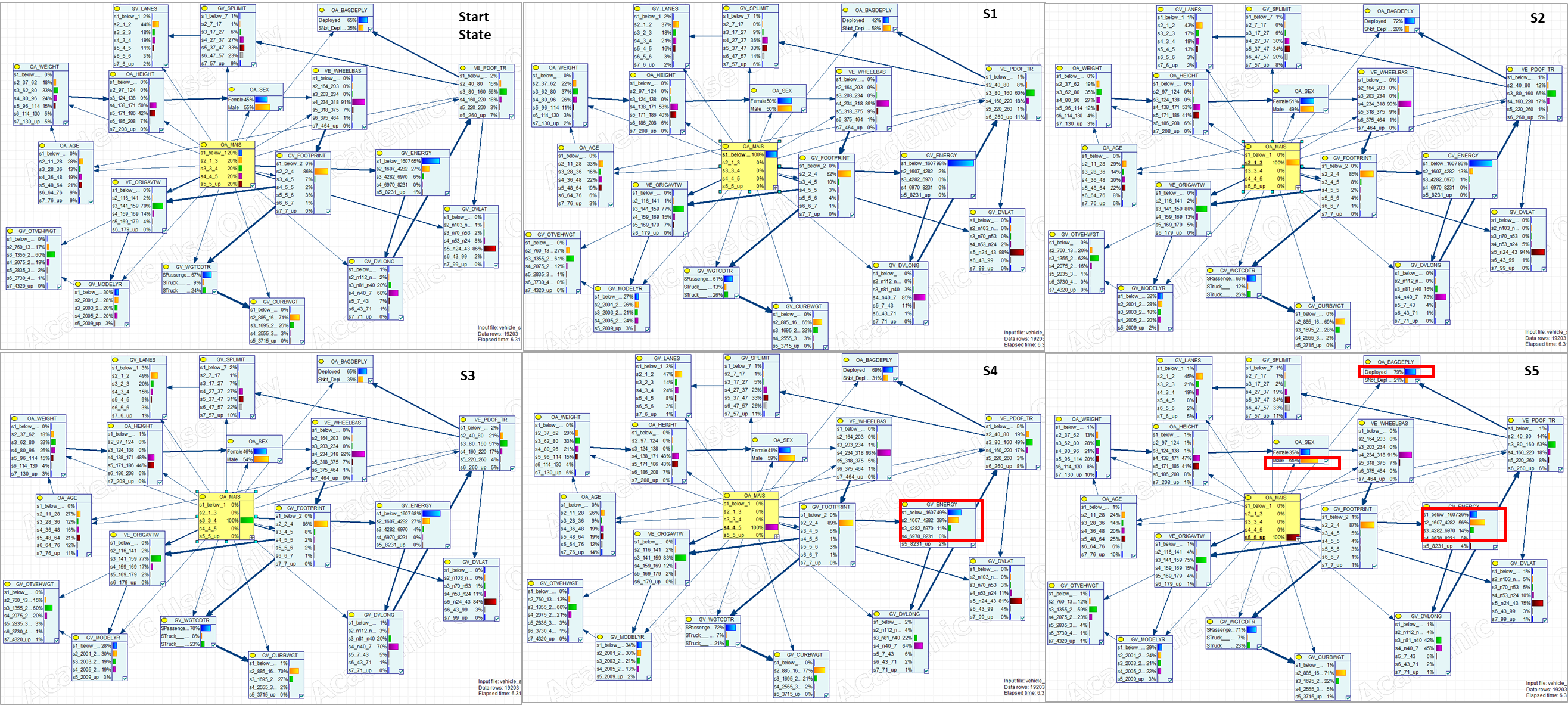
1. **Would greater vehicle footprint result in greater injury?**

|  |  |
| --- | --- |
|  | Oddly enough, the smallest and largest vehicles have the highest rates of serious injuries.  It could simply be due to the protective coverage and amount of energy transfer respectively. |

1. **Which gender is likely to face greater injury?**

|  |  |
| --- | --- |
|  | There is a bit of difference but not by much. It’s likely due to weight and height difference of the gender. |

1. **For each injury state, what is likely to cause it?**



The differentiation is quite minute until you get to more serious injury states. In this case, we believe that the accident is more likely to have males, with deployed air bags, and have energy levels of s2\_1607\_4282.

|  |  |
| --- | --- |
|  | For accidents without injury, we can say that the vehicle was traveling between 27 to 47 MPH, likely to have a small vehicle size, likely to be a passenger vehicle, and somehow the energy was below 1607J. It is a toss up between male and female drivers. |

**Conclusion**

We have compared two kinds of Bayesian networks – Naïve Bayes & Tree Augmented Network (TAN). The Naïve Bayes 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.

**REFERENCES**

1. Bayesian Lab
2. [Naniar](https://cran.r-project.org/web/packages/naniar/vignettes/naniar-visualisation.html)
3. [Bind\_Shadow](https://cran.r-project.org/web/packages/naniar/vignettes/getting-started-w-naniar.html)
4. [Binning](https://www.statisticshowto.datasciencecentral.com/choose-bin-sizes-statistics/)
5. [PMM Imputation](https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/)
6. C. Gower. A general coefficient of similarity and some of its properties. Biometrics, 27:857--874, 1971.