Bayesian Network Modelling Assignment

Team Members:

1: Vignesh Srinivasan
2. Navneet Goswami
3. Shobhit Jaipurkar
4. Naitik Shukla

Table of Contents

1. Tools and Package Used

- GeNIe 2.1 Academic was used to Implement, Train and Test the Generated Bayesian Network
- R Studio was used for selecting and determining important variables to be used.

1.1 *Learning Algorithms Used*

- Tree Augmented Naive Bayes
- Augmented Naive Bayes
- Random Forest
- Linear regression

2. Pre-processing Data

2.1 *Missing Value Imputation*

We used the **average value** to fill in the gaps for the missing values. This was the predominant method used to impute variables for data which was found to be continuous in nature. Other variables which were found to be non-continuous (Eg. VE_GAD1) were imputed by checking the dependent relationships with other variables such as (VE_PDOF_TR)

2.2 *Variable Selection*

The overall performance of a Bayesian Network depends on the strength of the selected variables. For our Bayesian Network Model, we used two main techniques, for selecting the most important variables. Our target class, “OA_MAIS”, shows the different categories of fatality rates in Car-Crash Accidents. The two variable selection techniques we used are listed below:

2.2.1 *Using Pearson correlation*

Using the **Pearson Correlation Coefficient**, we were able to single out the most significant variable.
The main logic behind this approach is to find the set of variables with the highest correlation. We obtain this by looking at all the other variables within our target class, “OA_MAIS”.
Below is our sample R code and its corresponding result.

1. Correlation Coefficient

```
> csv<-read.csv('C:/Users/naiti/Downloads/Compressed/bayesian-model/vehicle_safety_NASS2
010_2000_2012.csv',sep=",")
> attach(csv)
> lm.fit<-lm(OA_MAIS~.,data=csv)
> summary(lm.fit)
```

```
Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     4.802e+01  6.176e+00   7.775 8.21e-15
GV_CURBWGT                     -4.828e-05  4.802e-05  -1.005 0.314739
GV_DVLAT                        4.967e-03  1.101e-03   4.510 6.54e-06
GV_DVLONG                      -9.424e-03  8.527e-04 -11.052  < 2e-16
GV_ENERGY                       6.052e-04  1.693e-05  35.750  < 2e-16
GV_LANES                       -9.007e-03  5.899e-03  -1.527 0.126839
GV_MODELYR                     -2.228e-02  2.969e-03  -7.504 6.69e-14
GV_OTVEHWGT                     1.489e-04  2.034e-05   7.318 2.70e-13
GV_SPLIMIT                      3.611e-03  7.658e-04   4.716 2.44e-06
GV_WGTCDTRTruck (<=10000 lbs.)  3.908e-03  4.644e-02   0.084 0.932934
GV_WGTCDTRTruck (<=6000 lbs.)  -3.299e-02  2.300e-02  -1.434 0.151582
OA_AGE                          7.967e-03  4.599e-04  17.324  < 2e-16
OA_BAGDEPLYNot Deployed        -3.159e-01  1.911e-02 -16.532  < 2e-16
OA_HEIGHT                      -6.111e-04  1.044e-03  -0.585 0.558333
OA_MANUSE                      -3.615e-01  2.557e-02 -14.137  < 2e-16
OA_SEXFemale                    1.946e-02  8.284e-02   0.235 0.814320
OA_SEXMale                     -2.164e-01  8.422e-02  -2.569 0.010203
OA_WEIGHT                       3.551e-03  4.730e-04   7.507 6.52e-14
VE_GAD1Front                    2.272e-01  8.228e-01   0.276 0.782488
VE_GAD1Left                     7.920e-01  8.234e-01   0.962 0.336104
VE_GAD1Rear                     4.130e-01  8.263e-01   0.500 0.617192
VE_GAD1Right                    4.601e-01  8.232e-01   0.559 0.576203
VE_ORIGAVTW                    -1.869e-02  9.965e-03  -1.876 0.060690
VE_WHEELBAS                    -1.009e-02  5.623e-03  -1.794 0.072778
VE_PDOF_TR                      1.316e-03  3.507e-04   3.753 0.000176
GV_FOOTPRINT                    4.836e-01  3.474e-01   1.392 0.163930
```

2.2.2 *Using Random Forest*

The other method we used to single out significant variables, was using Random Forests. We generated random forests from our data and subsequently filtered the results based on the value of the GINI Index Impurity. This allowed us to bottleneck the total variables and pick out the ones most suitable for our Network Model.

```
> crash1 <- read.csv("C:/Users/naiti/Downloads/Compressed/bayesian-model/vehicle_safety_
NASS2010_2000_2012-modified-fillNA.csv")
> attach(crash1)
> require(randomForest)
> fit= randomForest(OA_MAIS~., data = crash1)
> importance(fit)
> varImpPlot(fit,type = 2)
```

R code for the method of Random Forests

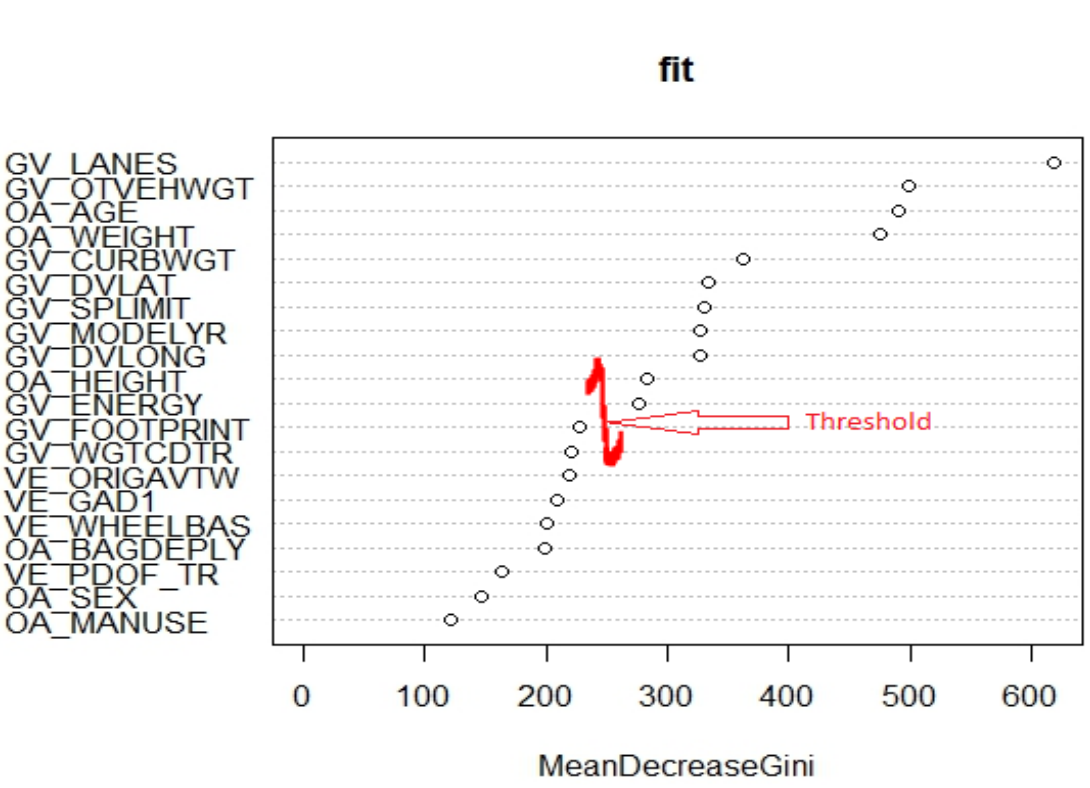


Fig 1: A plot between the Mean Gini Impurity and the Variables in OA_MAIS

2.2.3 *Outlier removal*

Some of the observations with outliers were removed. These are the descriptions of the removed data points.

1. **GV_ENERGY** > 4928 **45** RECORDS
2. **GV_CURBWGT** > 3300 **26** RECORDS
3. **GV_OTVEHWGT** > 3565 **22** RECORDS

- *Target Missing Value Removal*

All target variables within **OA_MAIS** containing missing values were removed from the observation. The total number of such null sets came to to **1132** records.

2.2.4 *Binning*

All the continuous variables are discretized using two binning methods:

- **Uniform width**
- **Uniform count**

3. Sampling Size

Sampling Size used for building the network was **50,** using the Augmented Naïve Bayes Algorithm. Sampling Size used for training and test phase was **14172**. This comes to around a 70:30 split between the number of observations to the test set.

4. Models Built

The two different learning algorithms implemented for our Bayesian network model were:

- Tree augmented Naive Bayes
- Augmented Naive Bayes

We implemented the aforementioned learning algorithms for three different test cases, to account for the differences in variable selection. These three test cases varied depending on the total number of variables used in the learning algorithm. Model efficiency was tested in three different ways, namely:

- Using all Variables
- Using variables by correlation
- Using variables by Random Forest.

In short, we tested the data with a total of 12 models, and compared their respective results. Here is a tabular depiction of the same:

Model No.	Learning Algorith	Variable Selection	Binning Used
1	Tree Augmented Naive Bayes	All Variables	Equal width
2	Tree Augmented Naive Bayes	All Variables	Equal count
3	Tree Augmented Naive Bayes	Correlation Variables	Equal width
4	Tree Augmented Naive Bayes	Correlation Variables	Equal count
5	Tree Augmented Naive Bayes	Random Forest Variable	Equal width
6	Tree Augmented Naive Bayes	Random Forest Variable	Equal count
7	Augmented Naive Bayes	All Variables	Equal width
8	Augmented Naive Bayes	All Variables	Equal count
9	Augmented Naive Bayes	Correlation Variables	Equal width
10	Augmented Naive Bayes	Correlation Variables	Equal count
11	Augmented Naive Bayes	Random Forest Variable	Equal width
12	Augmented Naive Bayes	Random Forest Variable	Equal count

Fig 2. Table of different Learning Algorithms used with corresponding Variable Selections

4. Model's Performance

Below is the comparison for all 12 models tested:

Model No.	Learning Algorith	Variable Selection	Binning Used
1	Tree Augmented Naive Bayes	All Variables	Equal width

Confusion Matrix:

	fat_none	fat_minor	fat_moderate	fat_serious	fat_severe	fat_critical	fat_maximum
fat_none	9	6	118	11	19	33	11
fat_minor	9	1	42	3	4	14	6
fat_moderate	20	5	6088	84	2437	136	21
fat_serious	8	5	1088	38	277	80	10
fat_severe	6	2	2998	24	4220	44	8
fat_critical	20	5	651	35	93	120	13
fat_maximum	5	3	181	14	42	40	8

Validation:

```
OA_MAIS = 0.54847  (10484/19115)
    fat_critical = 0.0434783  (9/207)
    fat_maximum = 0.0126582  (1/79)
    fat_minor = 0.692526  (6088/8791)
    fat_moderate = 0.0252324  (38/1506)
    fat_none = 0.577924  (4220/7302)
    fat_serious = 0.128068  (120/937)
    fat_severe = 0.0273038  (8/293)
```

Model No.	Learning Algorith	Variable Selection	Binning Used
2	Tree Augmented Naive Bayes	All Variables	Equal count

Confusion Matrix:

	fat_none	fat_minor	fat_moderate	fat_serious	fat_severe	fat_critical	fat_maximum
fat_none	4112	3102	18	54	5	7	4
fat_minor	2524	5967	57	164	37	32	10
fat_moderate	256	1122	17	76	20	9	6
fat_serious	85	657	23	137	21	10	4
fat_severe	31	197	6	30	18	11	0
fat_critical	12	144	5	26	10	7	3
fat_maximum	5	50	1	14	1	5	3

Validation:

```
OA_MAIS = 0.536804  (10261/19115)
    fat_none = 0.563133  (4112/7302)
    fat_minor = 0.678762  (5967/8791)
    fat_moderate = 0.0112882  (17/1506)
    fat_serious = 0.146211  (137/937)
    fat_severe = 0.0614334  (18/293)
    fat_critical = 0.0338164  (7/207)
    fat_maximum = 0.0379747  (3/79)
```

Model No.	Learning Algorith	Variable Selection	Binning Used
3	Tree Augmented Naive Bayes	Correlation Variables	Equal width

Confusion Matrix:

OA_MAIS	fat_none	fat_minor	fat_moderate	fat_serious	fat_severe	fat_critical	fat_maximum
fat_none	7	3	129	6	24	35	3
fat_minor	5	3	48	3	1	17	2
fat_moderate	15	0	6215	38	2377	134	12
fat_serious	6	4	1112	20	281	74	9
fat_severe	3	3	3050	10	4180	49	7
fat_critical	10	6	656	18	113	123	11
fat_maximum	6	1	183	10	47	41	5

Validation:

```
OA_MAIS = 0.55208  (10553/19115)
    fat_critical = 0.0338164  (7/207)
    fat_maximum = 0.0379747  (3/79)
    fat_minor = 0.706973  (6215/8791)
    fat_moderate = 0.0132802  (20/1506)
    fat_none = 0.572446  (4180/7302)
    fat_serious = 0.13127  (123/937)
    fat_severe = 0.0170648  (5/293)
```

Model No.	Learning Algorith	Variable Selection	Binning Used
4	Tree Augmented Naive Bayes	Correlation Variables	Equal count

Confusion Matrix:

OA_MAIS	fat_none	fat_minor	fat_moderate	fat_serious	fat_severe	fat_critical	fat_maximum
fat_none	4190	3070	0	32	7	2	1
fat_minor	2560	6028	17	165	12	3	6
fat_moderate	263	1163	6	67	3	4	0
fat_serious	69	739	10	101	9	8	1
fat_severe	30	221	3	31	5	2	1
fat_critical	15	160	1	27	2	1	1
fat_maximum	2	61	0	15	1	0	0

Validation:

```
        Correlated Variable
        -------------------
OA_MAIS = 0.540466  (10331/19115)
    fat_none = 0.573815  (4190/7302)
    fat_minor = 0.685701  (6028/8791)
    fat_moderate = 0.00398406  (6/1506)
    fat_serious = 0.107791  (101/937)
    fat_severe = 0.0170648  (5/293)
    fat_critical = 0.00483092  (1/207)
    fat_maximum = 0  (0/79)
```

Model No.	Learning Algorith	Variable Selection	Binning Used
5	Tree Augmented Naive Bayes	Random Forest Variable	Equal width

Confusion Matrix:

OA_MAIS	fat_none	fat_minor	fat_moderate	fat_serious	fat_severe	fat_critical	fat_maximum
fat_none	3	5	125	2	43	19	10
fat_minor	4	1	44	3	12	12	3
fat_moderate	9	2	6436	41	2233	58	12
fat_serious	2	5	1116	15	335	29	4
fat_severe	0	0	4036	15	3232	17	2
fat_critical	5	4	672	20	163	59	14
fat_maximum	3	1	186	7	66	24	6

Validation:

```
OA_MAIS = 0.510175  (9752/19115)
    fat_critical = 0.0144928  (3/207)
    fat_maximum = 0.0126582  (1/79)
    fat_minor = 0.732112  (6436/8791)
    fat_moderate = 0.00996016  (15/1506)
    fat_none = 0.442618  (3232/7302)
    fat_serious = 0.0629669  (59/937)
    fat_severe = 0.0204778  (6/293)
```

Model No.	Learning Algorith	Variable Selection	Binning Used
6	Tree Augmented Naive Bayes	Random Forest Variable	Equal count

Confusion Matrix:

OA_MAIS	fat_none	fat_minor	fat_moderate	fat_serious	fat_severe	fat_critical	fat_maximum
fat_none	2394	4900	0	6	1	1	0
fat_minor	1591	7174	0	20	3	1	2
fat_moderate	187	1305	1	11	1	1	0
fat_serious	76	826	0	32	3	0	0
fat_severe	29	249	1	8	5	1	0
fat_critical	16	181	0	8	1	1	0
fat_maximum	8	66	0	5	0	0	0

Validation:

```
OA_MAIS = 0.50259   (9607/19115)
    fat_none = 0.327855   (2394/7302)
    fat_minor = 0.816062   (7174/8791)
    fat_moderate = 0.000664011   (1/1506)
    fat_serious = 0.0341515   (32/937)
    fat_severe = 0.0170648   (5/293)
    fat_critical = 0.00483092   (1/207)
    fat_maximum = 0   (0/79)
```

Model No.	Learning Algorith	Variable Selection	Binning Used
7	Augmented Naive Bayes	All Variables	Equal width

Confusion Matrix:

Fatality	fat_none	fat_minor	fat_moderate	fat_serious	fat_severe	fat_critical	fat_maximum
fat_none	3998	3259	20	20	3	2	0
fat_minor	2557	6029	77	109	13	3	3
fat_moderate	265	1141	31	59	3	5	2
fat_serious	97	687	36	86	14	9	8
fat_severe	38	199	10	32	6	8	0
fat_critical	18	142	8	19	8	7	5
fat_maximum	6	42	5	15	5	3	3

Validation:

```
OA_MAIS = 0.53152   (10160/19115)
    fat_none = 0.547521   (3998/7302)
    fat_minor = 0.685815   (6029/8791)
    fat_moderate = 0.0205843   (31/1506)
    fat_serious = 0.0917823   (86/937)
    fat_severe = 0.0204778   (6/293)
    fat_critical = 0.0338164   (7/207)
    fat_maximum = 0.0379747   (3/79)
```

Model No.	Learning Algorith	Variable Selection	Binning Used
8	Augmented Naive Bayes	All Variables	Equal count

Confusion Matrix:

OA_MAIS	fat_none	fat_minor	fat_moderate	fat_serious	fat_severe	fat_critical	fat_maximum
fat_none	4076	3110	61	34	12	9	0
fat_minor	2539	5889	154	142	32	27	8
fat_moderate	285	1069	39	76	16	16	5
fat_serious	90	671	46	91	16	17	6
fat_severe	33	192	22	30	13	3	0
fat_critical	14	134	13	28	13	4	1
fat_maximum	4	49	4	13	7	2	0

Validation:

```
OA_MAIS = 0.529009  (10112/19115)
    fat_none = 0.558203  (4076/7302)
    fat_minor = 0.66989  (5889/8791)
    fat_moderate = 0.0258964  (39/1506)
    fat_serious = 0.0971185  (91/937)
    fat_severe = 0.0443686  (13/293)
    fat_critical = 0.0193237  (4/207)
    fat_maximum = 0  (0/79)
```

Model No.	Learning Algorith	Variable Selection	Binning Used
9	Augmented Naive Bayes	Correlation Variables	Equal width

Confusion Matrix:

Fatality	fat_none	fat_minor	fat_moderate	fat_serious	fat_severe	fat_critical	fat_maximum
fat_none	3991	3272	16	16	5	2	0
fat_minor	2513	6126	50	77	13	8	4
fat_moderate	277	1124	37	47	11	9	1
fat_serious	107	699	35	63	11	20	2
fat_severe	46	186	14	25	13	7	2
fat_critical	23	137	9	22	6	7	3
fat_maximum	6	47	5	11	3	6	1

Validation:

```
OA_MAIS = 0.5356  (10238/19115)
    fat_none = 0.546563  (3991/7302)
    fat_minor = 0.696849  (6126/8791)
    fat_moderate = 0.0245684  (37/1506)
    fat_serious = 0.0672359  (63/937)
    fat_severe = 0.0443686  (13/293)
    fat_critical = 0.0338164  (7/207)
    fat_maximum = 0.0126582  (1/79)
```

Model No.	Learning Algorith	Variable Selection	Binning Used
10	Augmented Naive Bayes	Correlation Variables	Equal count

Confusion Matrix:

OA_MAIS	fat_none	fat_minor	fat_moderate	fat_serious	fat_severe	fat_critical	fat_maximum
fat_none	4109	3145	17	23	3	1	4
fat_minor	2547	6056	48	109	19	9	3
fat_moderate	281	1155	15	49	3	2	1
fat_serious	96	739	23	66	4	7	2
fat_severe	46	219	4	17	4	1	2
fat_critical	17	159	3	21	3	4	0
fat_maximum	4	61	3	9	2	0	0

Validation:

```
OA_MAIS = 0.536437  (10254/19115)
    fat_none = 0.562723  (4109/7302)
    fat_minor = 0.688886  (6056/8791)
    fat_moderate = 0.00996016  (15/1506)
    fat_serious = 0.0704376  (66/937)
    fat_severe = 0.0136519  (4/293)
    fat_critical = 0.0193237  (4/207)
    fat_maximum = 0  (0/79)
```

Model No.	Learning Algorith	Variable Selection	Binning Used
11	Augmented Naive Bayes	Random Forest Variable	Equal width

Confusion Matrix:

	fat_none	fat_minor	fat_moderate	fat_serious	fat_severe	fat_critical	fat_maximum
fat_none	2463	4809	12	11	5	2	0
fat_minor	2253	6389	48	72	11	13	5
fat_moderate	327	1094	24	47	9	3	2
fat_serious	149	670	27	69	10	9	3
fat_severe	47	200	15	20	3	4	4
fat_critical	37	137	3	17	4	4	5
fat_maximum	8	52	4	9	2	3	1

Validation:

```
OA_MAIS = 0.468376  (8953/19115)
    fat_none = 0.337305  (2463/7302)
    fat_minor = 0.726766  (6389/8791)
    fat_moderate = 0.0159363  (24/1506)
    fat_serious = 0.0736393  (69/937)
    fat_severe = 0.0102389  (3/293)
    fat_critical = 0.0193237  (4/207)
    fat_maximum = 0.0126582  (1/79)
```

Model No.	Learning Algorith	Variable Selection	Binning Used
12	Augmented Naive Bayes	Random Forest Variable	Equal count

Confusion Matrix:

OA_MAIS	fat_none	fat_minor	fat_moderate	fat_serious	fat_severe	fat_critical	fat_maximum
fat_none	3264	3974	29	22	6	6	1
fat_minor	2817	5783	88	73	19	9	2
fat_moderate	362	1078	21	29	8	7	1
fat_serious	149	714	25	32	13	3	1
fat_severe	61	200	6	22	2	2	0
fat_critical	39	144	5	14	3	1	1
fat_maximum	10	59	0	8	2	0	0

Validation:

```
OA_MAIS = 0.476223  (9103/19115)
    fat_none = 0.447001  (3264/7302)
    fat_minor = 0.657832  (5783/8791)
    fat_moderate = 0.0139442  (21/1506)
    fat_serious = 0.0341515  (32/937)
    fat_severe = 0.00682594  (2/293)
    fat_critical = 0.00483092  (1/207)
    fat_maximum = 0  (0/79)
```

5. Important Findings

- Gender variable has a very little effect on the fatality rate

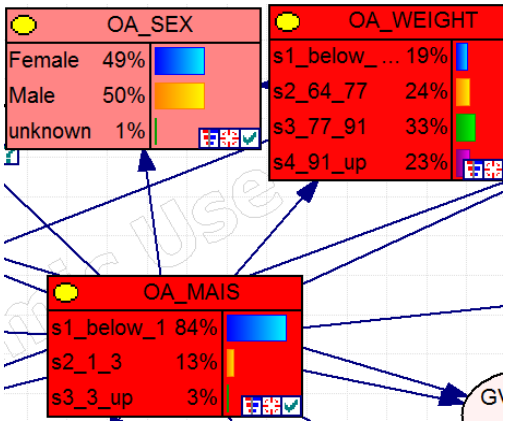


Fig1

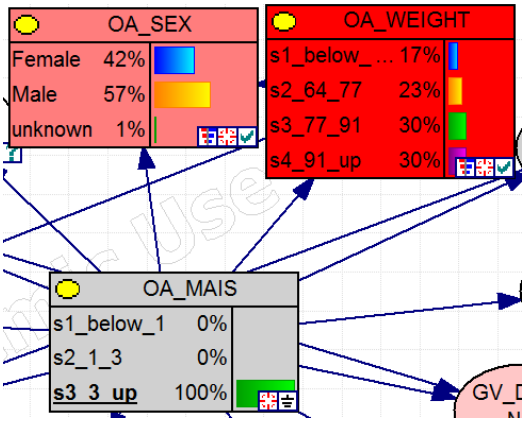
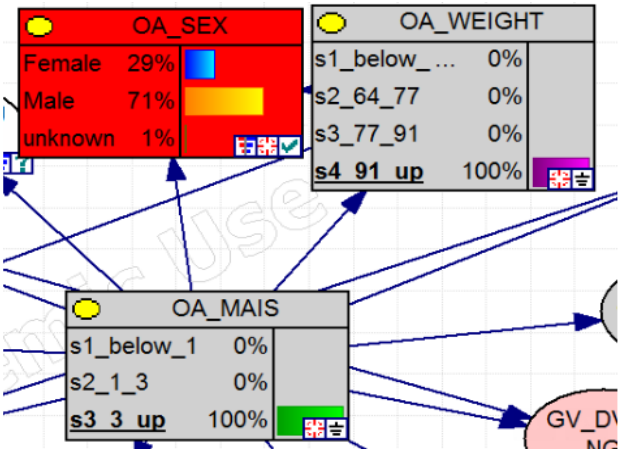


Fig2

- Men have a larger rate of fatality in serious accidents, as seen from the second figure above.

- We also infer that men have higher trend of fatal accidents as the weight of car increases. Hence men who have serious accidents are usually driving heavy-duty vehicles



- Other information we gather is that, vehicles with a weight between 1550-1840 has a slighter higher frequency in all types of accident. While the other kinds of vehicles have similar characteristics, we find that in the case of serious accidents, the fatality increases for the vehicles under the heavy-weight category

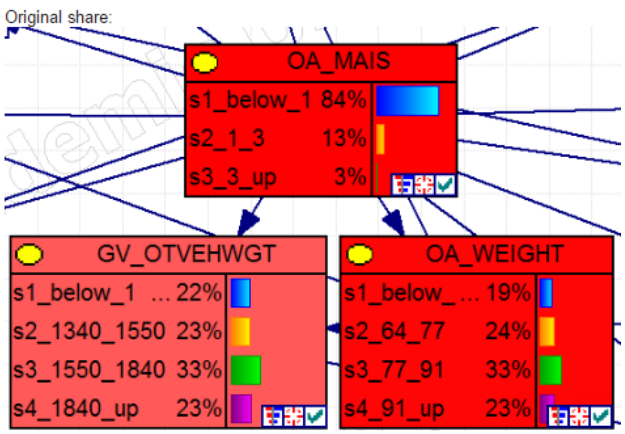


Fig3
(Original Share)

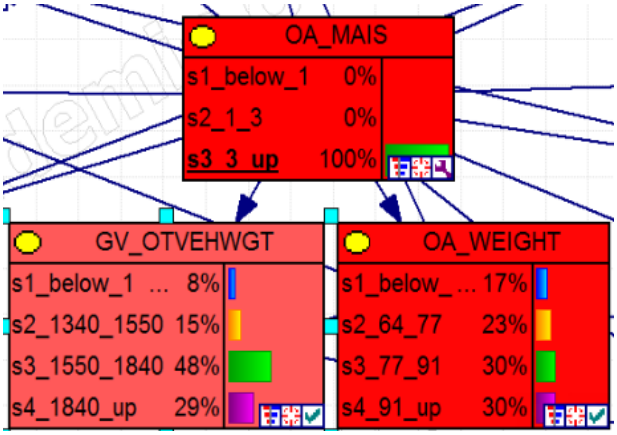


Fig4
(Share with Serious Accidents)

Note: *There is no major change between 1st vehicle weight and other vehicle weight after the accident happens.*

6. Conclusion

From above document we conclude that Model 4 with the Tree Augmented Naïve Bayes Learning algorithm and Correlated variables with equal width binning outperforms other Bayesian models, albeit, with a minimal little margin.

Model 4 Performance is: **55.20 %**