

**SIT743 Bayesian Learning and Graphical  
Models**  
**Assignment-2**

MUHAMMAD AHSAN ASIF  
218606833

Q:1)

1.1 The joint probability distribution is:

$$P(H, W, N, M, R, S) = P(H) P(M|H) P(W|H) P(N|W, H) \\ P(R|N, H) P(S|N)$$

Q:2 where

$$\begin{aligned} P(H) &= P(\text{Relative Humidity}) \\ P(M) &= P(\text{Maximum Temperature}) \\ P(N) &= P(\text{Minimum Temperature}) \\ P(R) &= P(\text{Precipitation}) \\ P(W) &= P(\text{Wind Speed}) \\ P(S) &= P(\text{Solar Radiation}) \end{aligned}$$

1.2 The minimum numbers of parameters required to fully specify is the distribution according to the network is :

- For  $H = (3-1) = 2$ ;  $P(H)$ : H has three potential values needing two parameters.
- For  $W = 3^*(3-1) = 6$ ;  $P(W)$ : W has one parent H which furthermore has three potential values.



- For  $N = 3^* 3^*(4-1) = 27$ ;  $P(N) : N$  which has two parents, H & W and both of them have three potential values whereas N has 4 potential values.

$P(M)$ :

- For  $M = 3^* (4-1) = 9$  M has one parent H which has three potential values. ~~M has 4 potential values~~
- For  $R = 3^* 4^*(2-1) = 12$ ; P(R) = R has two parents, H has three potential values and N has 4 potential values.
- For  $S = 4^*(3-1) = 8$ ; P(S) : S has one parent N, with 4 values and S has three potential values

→ Total Number of parameters required

$$2 + 6 + 27 + 9 + 12 + 8 = 64$$



1.3)

a) Joint probability function if there are no independence among the variables is assumed:

$$\rightarrow P(H, W, N, M, R, S) = P(H) P(W | H) P(N | H, W) \\ P(M | H, W, N, \cancel{S}) P(R | H, W, N, M) P(S | H, W, N, M, R)$$

b) The minimum number of parameters required if there are no independencies among the variable :

Using the 1.3.a) equation :

$$= (3-1) + (3 \times (3-1)) + (4 \times 3 \times (3-1)) + \\ (4 \times 4 \times 3 \times (3-1)) + (2 \times 4 \times 4 \times 3 \times (3-1)) + \\ (3 \times 2 \times 4 \times 4 \times 3 \times (3-1))$$

$$= 2 + 6 + 24 + 96 + 192 + 576$$

$$= 896 \text{ parameters}$$



c) Assuming that there is no independence among the variables results in a larger number of parameters being required as compared to the joint distribution. Conditional and marginal properties of the variables reduces the complexity of the model.

1.4)

a)  $M \perp S \mid \emptyset$  ( $M$  is marginally independent of  $S$ )

Paths :

- $M \rightarrow H \rightarrow N \rightarrow S$  : Unobserved Tail to Tail and head to tail nodes along this path which means that the path is unblocked.
- $M \rightarrow H \rightarrow R \rightarrow N \rightarrow S$  : Unobserved head to head node which means it is blocking. There is a blockage at  $R$  node.
- $M \rightarrow H \rightarrow W \rightarrow N \rightarrow S$  : Unobserved tail to tail and head to tail nodes which means that the path is unblocked.

Result :

Only one of the paths is blocked and the rest are non-blocked paths, therefore nodes  $M$  and  $S$  are not d-separated, hence not independent.

Therefore the result is FALSE.



b)  $W \perp R | \{N, H\}$  ( $W$  is conditionally independent of  $R$  given  $\{N, H\}$ )

Paths :

- $W \rightarrow N \rightarrow R$  :  $N$  is an observed head to tail node, therefore blocks the path at  $N$
- $W \rightarrow N \rightarrow H \rightarrow R$  :  $H$  is an observed tail to tail node, and  $N$  is an observed head to tail node and both block the path
- $W \rightarrow H \rightarrow N \rightarrow R$  :  $H$  is an observed tail to tail node, and  $N$  is an unobserved head to tail and both are blocking the path.
- $W \rightarrow H \rightarrow R$  :  $H$  is an observed tail to tail node and is blocking the path.

Results :

All paths are blocked therefore  $W$  is d-separated and conditionally independent of  $R$  given that  $N$  and  $H$  are observed.

Therefore it is TRUE

c)  $\{R, S\} \perp W | H$  ( $R$  and  $S$  conditionally independent of  $W$  given  $H$ )

Paths :

- $R \rightarrow N \rightarrow W$  :  $R$  is an unobserved head to head node therefore it is blocked at  $R$ .
- $R \rightarrow H \rightarrow W$  : Again  $R$  is an unobserved head to head node, and  $H$  is observed tail to tail node therefore both nodes are blocked
- $R \rightarrow N \rightarrow H \rightarrow W$  :  $R$  is an unobserved head to head node, and  $H$  is an observed tail to tail node and both are blocked
- $R \rightarrow H \rightarrow N \rightarrow W$  :  $R, N$  both are unobserved head-to-head nodes and  $H$  is an observed tail-to-tail node and all three nodes are blocked
- $S \rightarrow N \rightarrow W$  :  $N$  is an tail to head node therefore it is not blocking
- $S \rightarrow N \rightarrow H \rightarrow W$  :  $H$  is an observed tail to tail node therefore it is blocked.
- $S \rightarrow N \rightarrow R \rightarrow H \rightarrow W$  :  $R$  is an unobserved head to head node, and  $H$  is an observed tail to tail node and both are blocked.

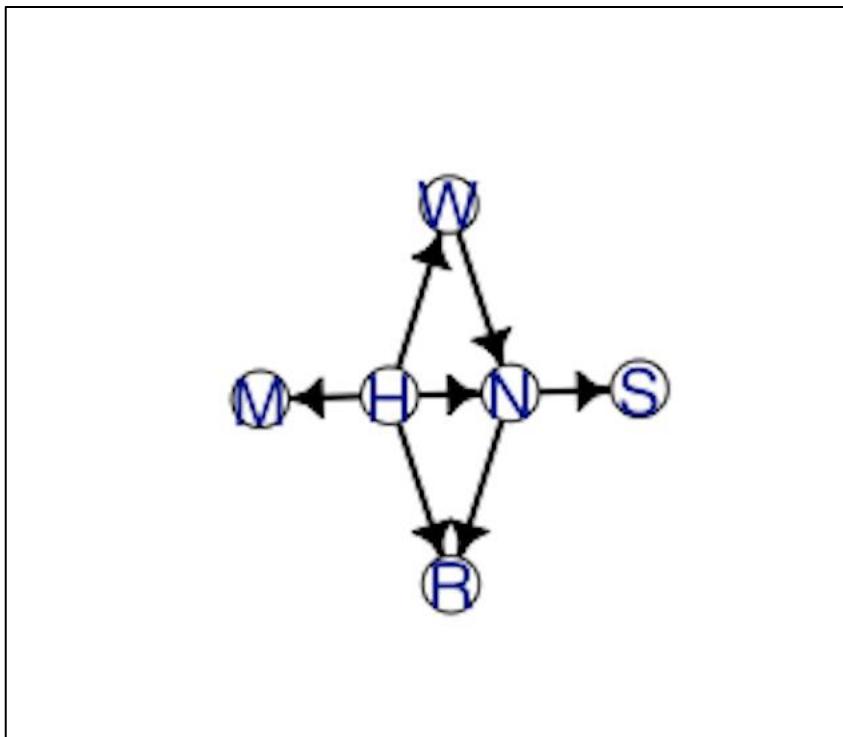
Results :

There is one path between S and W that is not blocked and not ~~d~~ d-separated given H, even being R d-separated from W. R and S are not conditionally independent of W given H.

Therefore the result is FALSE

1.5)

Plot of the network :



a)  $M \perp S \mid \emptyset$  ( $M$  is marginally independent of  $S$ )

```
> dSep(dag, first="M", second="S", cond=NULL)
[1] FALSE
```

b)  $W \perp R \mid \{N, H\}$  ( $W$  is conditionally independent of  $R$  given  $\{N, H\}$ )

```
> dSep(dag, first="W", second="R", cond=c("N", "H"))
[1] TRUE
```

c)  $\{R, S\} \perp W \mid H$

```
> dSep(dag, first=c("R", "S"), second="W", cond="H")
[1] FALSE
```

1.6)

a)  $P(W|S = \text{low}, R = \text{low})$

$$P(W, S, R) = \sum_{M, H, N} P(H, M, W, N, R, S)$$

Factors for conditional probability :

$$\begin{aligned} \cdot P(W, S, R) &= \sum_{M, H, N} P(H) P(W|H) P(M|H) \\ &\quad P(N|W, H) P(R|N, H) \\ &\quad P(S|N) \end{aligned}$$

$$\begin{aligned} \cdot P(W, S, R) &= \sum_{M, H, N} f_0(H) f_1(W, H) \\ &\quad f_2(M, H) f_3(N, W, H) \\ &\quad f_6(N, H) f_7(N) \end{aligned}$$

Decomposing with the order  $N, H, M$  :

$$\begin{aligned} \cdot P(W, S = \text{low}, R = \text{low}) &= \sum_{M, H} f_0(H) f_1(W, H) \\ &\quad f_2(M, H) \sum_N f_3(N, W, H) f_6(N, H) f_7(N) \end{aligned}$$

$$\cdot P(W, S = \text{low}, R = \text{low}) = \sum_{M, H} f_0(H) f_1(W, H) f_2(M, H) f_8(W, H)$$



- $\sum_M \sum_H f_0(H) f_1(W, H) f_2(M, H) f_8(W, H)$
- $\sum_M f_9(W) f_2(M)$
- $f_9(W) \sum_M f_2(M)$
- $f P(W, S=\text{Low}, R=\text{Low}) = f_{10}(W)$

Normalising to get the posterior distribution

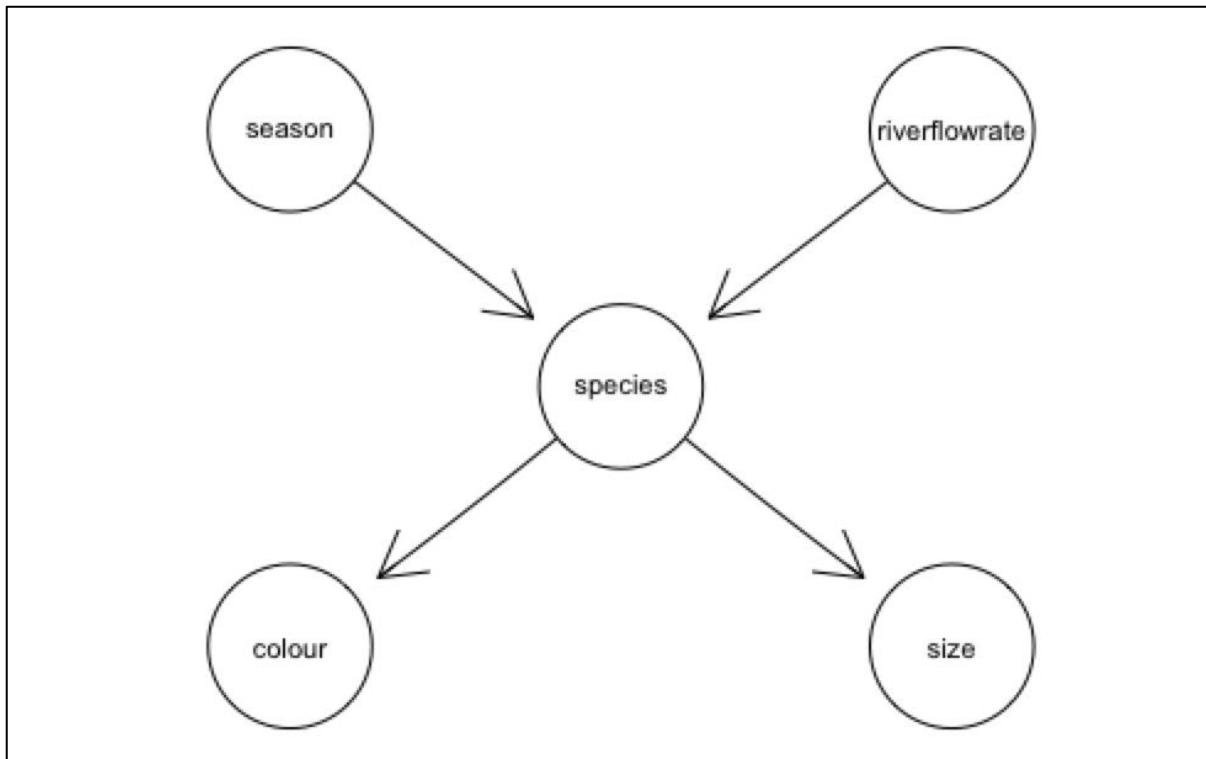
$$P(W, S=\text{Low}, R=\text{Low}) = \frac{f_{10}(W)}{\sum_W f_{10}(W)}$$

b) Given the above elimination ordering  
the treewidth is 2.

## Question 2)

2.1 :

a) Belief Network :



b) Probability Tables :

1) Season Table :

```
> plist$season
season
wet dry
0.3 0.7
```

2) River Flow Rate Table

```
> plist$riverflowrate
riverflowrate
low high
0.8 0.2
```

### 3) Species Table

```
> plist$species
, , riverflowrate = low

      season
species wet dry
bass   0.6 0.3
cod    0.4 0.7

, , riverflowrate = high

      season
species wet dry
bass   0.4 0.5
cod    0.6 0.5
```

### 4) Colour Table

```
> plist$colour
      species
colour   bass cod
light    0.2 0.5
medium   0.4 0.3
dark     0.4 0.2
```

### 5) Size Table

```
> plist$size
      species
size    bass cod
wide   0.6 0.4
thin   0.4 0.6
```

Marginal Probabilities across all nodes :

```
> querygrain(net1)
$season
season
wet dry
0.3 0.7

$riverflowrate
riverflowrate
low high
0.8 0.2

$species
species
bass cod
0.406 0.594

$colour
colour
light medium dark
0.3782 0.3406 0.2812

$size
size
wide thin
0.4812 0.5188
```

## 2.2 :

- a) Given that the river flow rate is low, the probability that size is thin is 0.522.

```
> # a) P(riverflowrate = low, size = thin)
> net1WithEvidence_A <- setEvidence(net1, evidence=list(riverflowrate="low"))
> querygrain(net1WithEvidence_A, nodes=c("size"))
$size
size
wide thin
0.478 0.522
```

- b) Given that the colour is dark, and the season is dry, the probability that the fish species is cod is 0.492.

```
> # b) P(species=cod|season=dry, color=dark)
> net1WithEvidence_B <- setEvidence(net1, evidence=list(season="dry", colour="dark"))
> querygrain(net1WithEvidence_B, nodes=c("species"))
$species
species
  bass      cod
0.5074627 0.4925373
```

- c) The joint distribution of color and fish species is :

```
> querygrain(net1, nodes=c("colour", "species"), type="joint")
  colour
species light medium   dark
  bass 0.0812 0.1624 0.1624
  cod  0.2970 0.1782 0.1188
```

- d) The marginal distribution of fish species is :

```
> querygrain(net1, nodes=c("species"), type="marginal")
$species
species
  bass      cod
0.406 0.594
```

Question 3)

Q.3)

31)

$$P(A, B, C, D) = P(A) P(B) P(C|A, B) P(D|B, C)$$

$$P(A, B) = \sum_{CD} P(A) P(B) P(C|A, B) P(D|B, C)$$

$$= P(A) P(B) \sum_C P(C|A, B) \sum_D P(D|B, C)$$

$$= P(A) P(B)$$

Since  $\sum_C P(C|A, B) = 1$  and  $\sum_D P(D|B, C) = 1$

Therefore  $A \perp B | \emptyset$

3.2)

$$P(D=1 | A=1, B=1) = \frac{P(D=1, A=1, B=1)}{\sum_{C} P(D, A=1, B=1)}$$

$$P(D, A, B) = \sum_{C} P(D, A, B, C)$$

$$P(D, A, B) = \sum_{C} P(A) \times P(B) \times P(D | C_{10}, B) \times P(C_{10} | A, B)$$

$$= [P(A) \times P(B) \times P(D | \neg C_{10}, B) \times P(\neg C_{10} | A, B)] \\ + [P(A) \times P(B) \times P(D | C_{10}, B) \times P(C_{10} | A, B)]$$

→ Replacing the values using CPT

$$[(1-\alpha) \times (1-\beta) \times 0.8 \times 0.8] + [(1-\alpha) \times (1-\beta) \times (1-\beta) \times 0.2]$$

$$= (1-\alpha)(1-\beta)(0.64 + 0.20 - 0.2\beta)$$

$$(1-\alpha)(1-\beta)(0.84 - 0.2\beta) \rightarrow \text{equation (i)}$$

$$\rightarrow P(A, B) = P(A) \times P(B) \rightarrow \text{equation(ii)}$$

$$P(A=1, B=1) = (1-\alpha)(1-\alpha)$$

$$P(A, B) = (1-\alpha)(1-\alpha)$$

$$\therefore \text{Since } P(D|A, B) = \frac{P(D, A, B)}{P(A, B)}$$

$$P(D|A, B) = \frac{(1-\alpha)(1-\alpha)[0.84 - 0.2\beta]}{(1-\alpha)(1-\alpha)}$$

$$P(D=1 | A=1, B=1) = 0.84 - 0.2\beta$$

$$3.3) d = P(A=0)$$
$$= \frac{\#(A=0)}{N}$$

$$= \frac{7}{20} \Rightarrow 0.35$$

$$\underline{d = 0.35}$$

$$Q = P(B=0)$$

$$= \frac{\#(B=0)}{N}$$

$$= \frac{3}{20} \Rightarrow 0.15$$

$$\rightarrow Q = 0.15$$

$$Y = P(C=1 | A=1, B=0)$$

$$= \frac{\#(C=1, A=1, B=0)}{\#(A=1, B=0)}$$

$$= \frac{1}{3} \Rightarrow 0.33$$

$$\rightarrow Y = 0.33$$

$$B = P(D=0 | B=1, C=1)$$

$$= \frac{\#(D=0, B=1, C=1)}{\#(B=1, C=1)}$$

$$= \frac{1}{3} \Rightarrow 0.33$$

3.4) From 3.3  $\beta = 0.33$

$$P(D=1 | A=1, B=1) = 0.84 - 0.2\beta$$

$$P(D=1 | A=1, B=1) = 0.84 - 0.2(0.3)$$

$$\Rightarrow P(D=1 | A=1, B=1) = 0.774$$

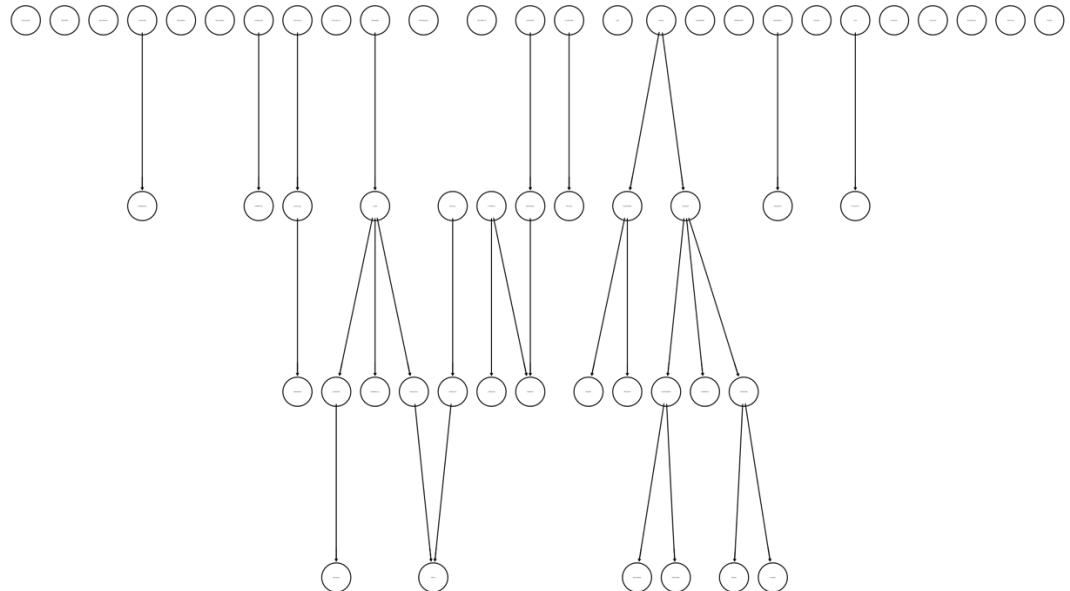
**Q.4 )**

**4.1**

a) 100 (first 100 data)

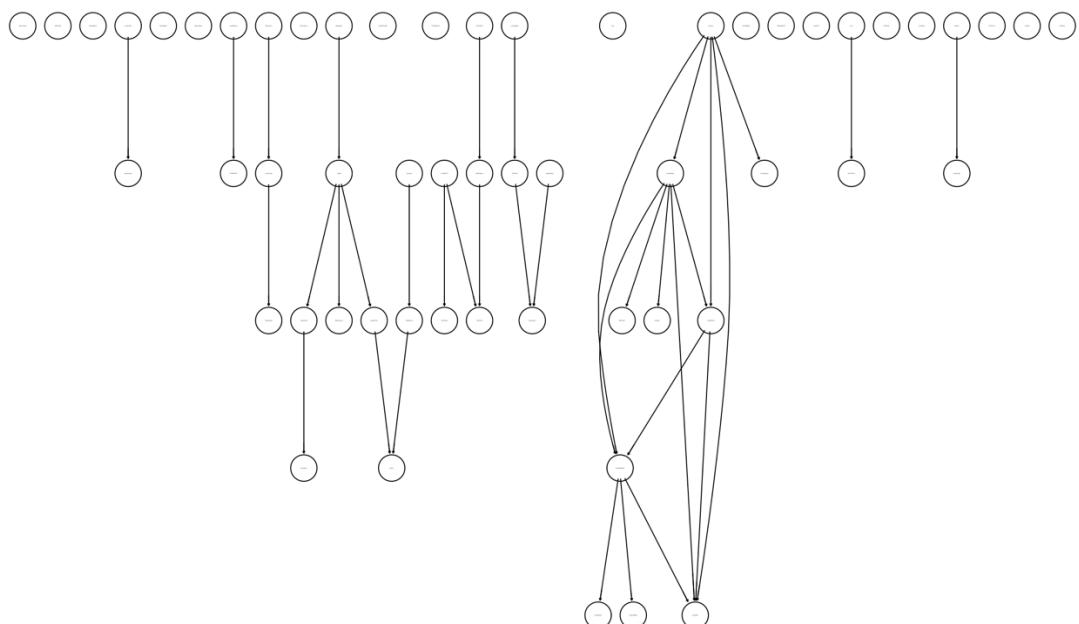
BIC Score : -6429.161

Network Structure obtained for BIC :



BDe Score : -5980.807

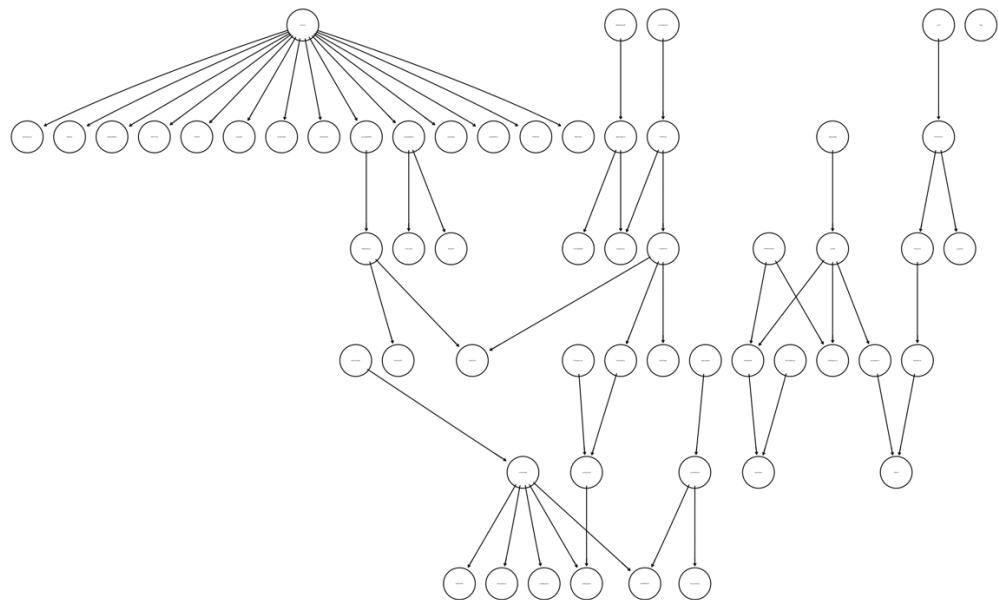
Network Structure obtained for BDe :



b) 1000 (first 1000 data)

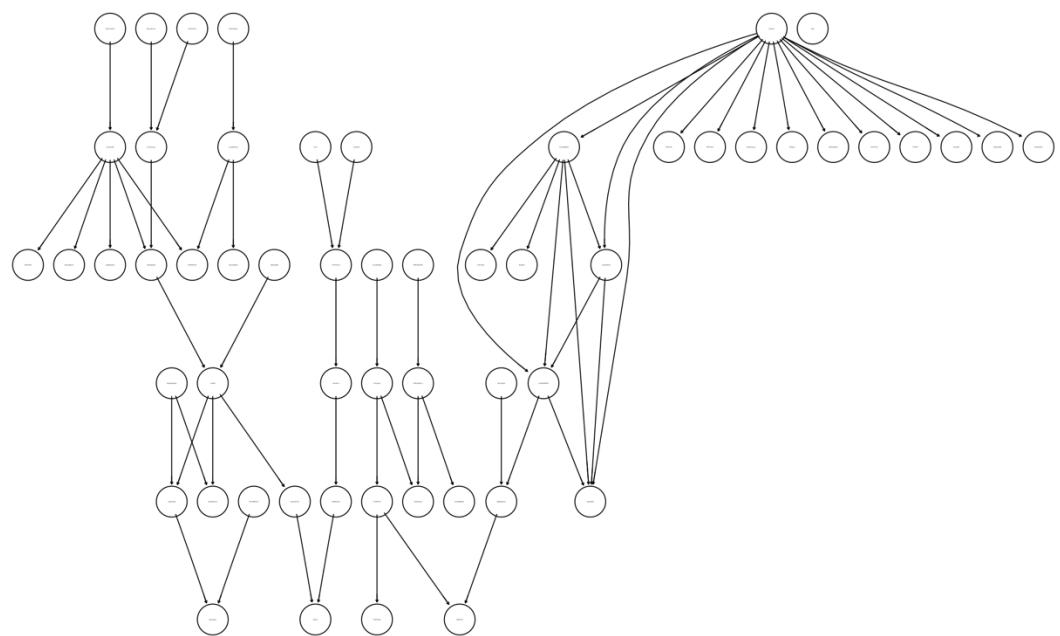
BIC Score : -52906.94

Network Structure obtained for BIC :



BDe Score : -52256.44

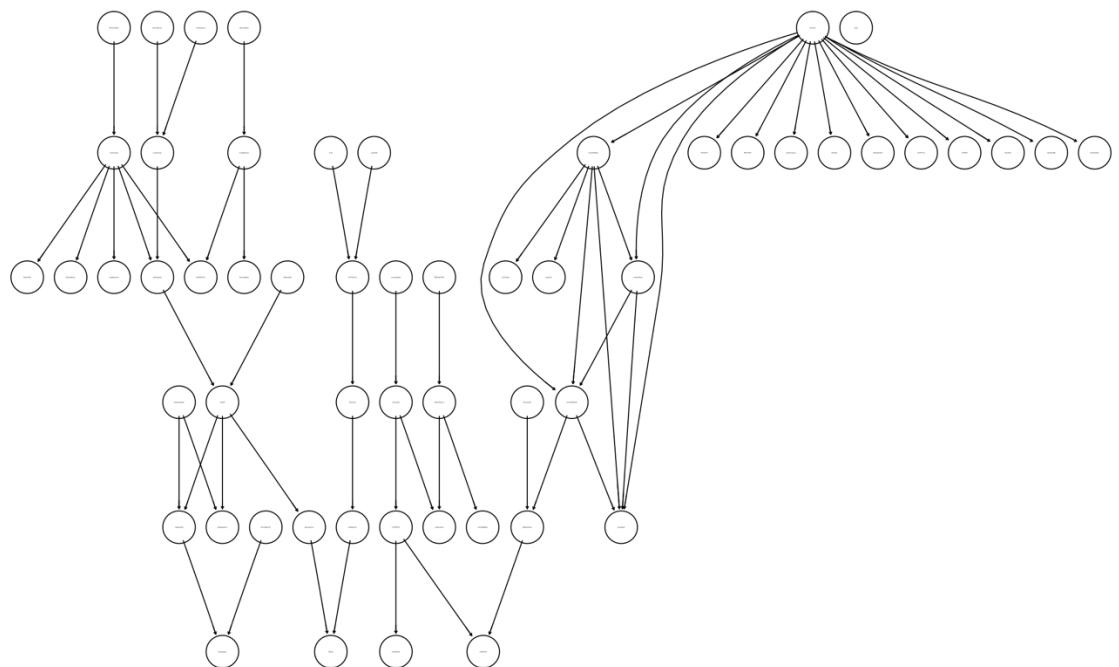
Network Structure obtained for BDe :



c) 10000 (first 10000 data)

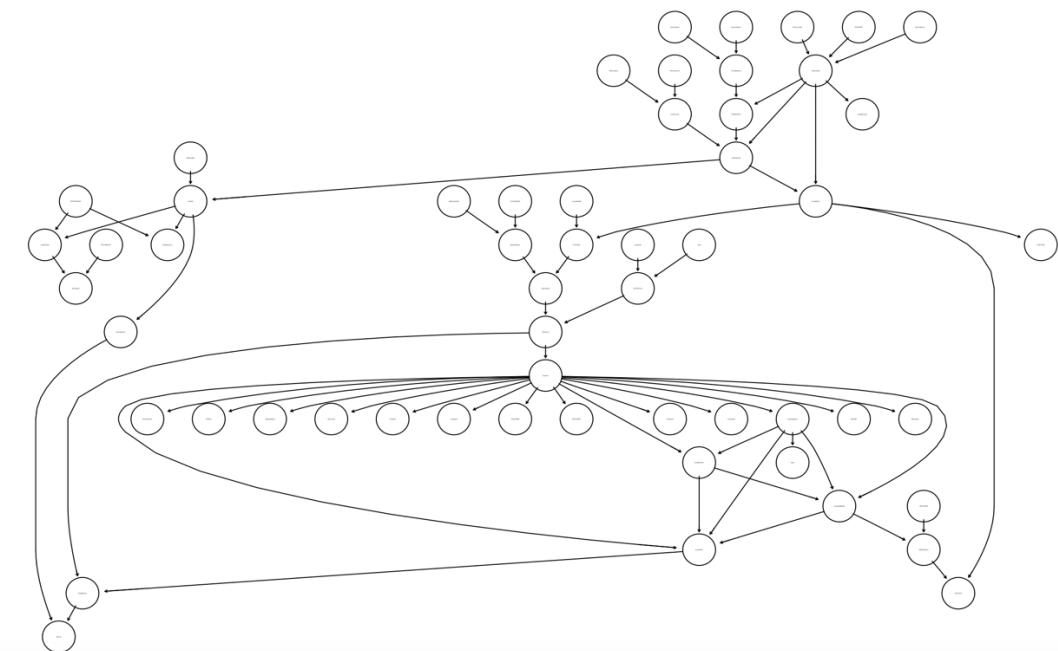
BIC Score : -498383.8

Network Structure obtained for BIC :



BDe Score : -497403.6

Network Structure obtained for BDe :



4.2)

Scoring function helps to measure how well the model fits. As seen in the above scores, the BIC score is seemed to be less than the Bde score consistently, and the scores gap increases as the sample size increases. Moreover, with increasing sample size the complexity of the model also increases. BDe is a scoring function using Dirichlet prior. They tend to get more complex with a large number of sample sizes. Bayesian scores are biased towards simpler structure and as sample size increases, they get less complex.

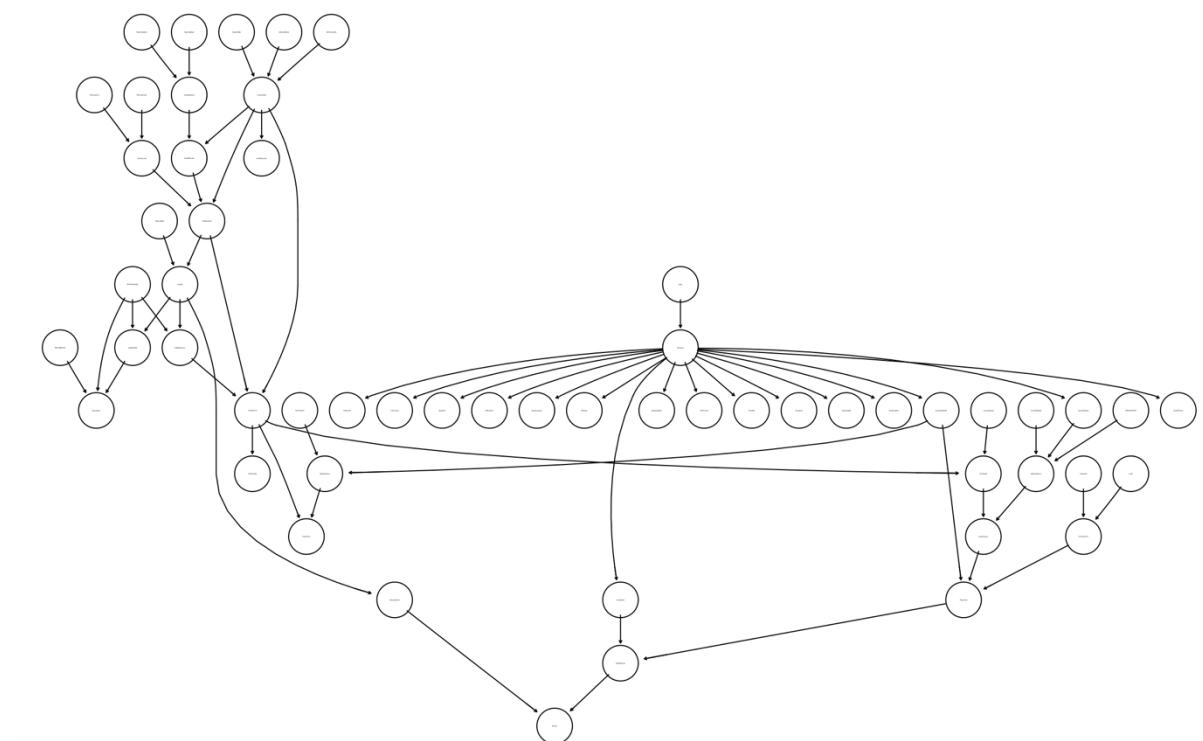
BIC is penalized version of the likelihood score. The score is directly proportional to the dependence of a variable on its parent. The more complex the network the lower BIC score it will result in. When the sample sizes increases, the score function would be preferring a structure that fits the dependencies in the data. With increasing number of sample size the similarity between BIC and Bde network structures tends to vanish away

4.3)

a)

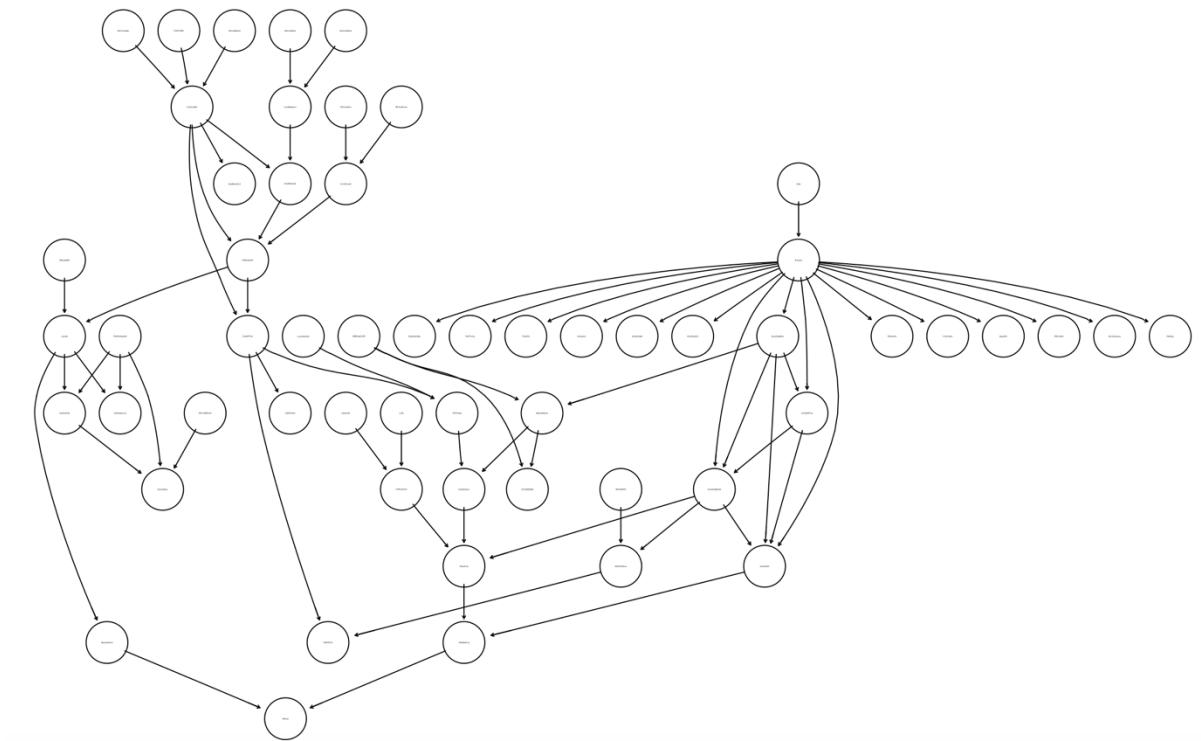
BIC score of the whole dataset : -990474.8

### Network structure using BIC for the whole dataset :



BIC score of the whole dataset : -989470.3

Network structure using BIC for the whole dataset :



b)

Comparing the true network with network obtained through bic score :

```
> compare(dag,bnet_bicfull)
```

```
$tp
```

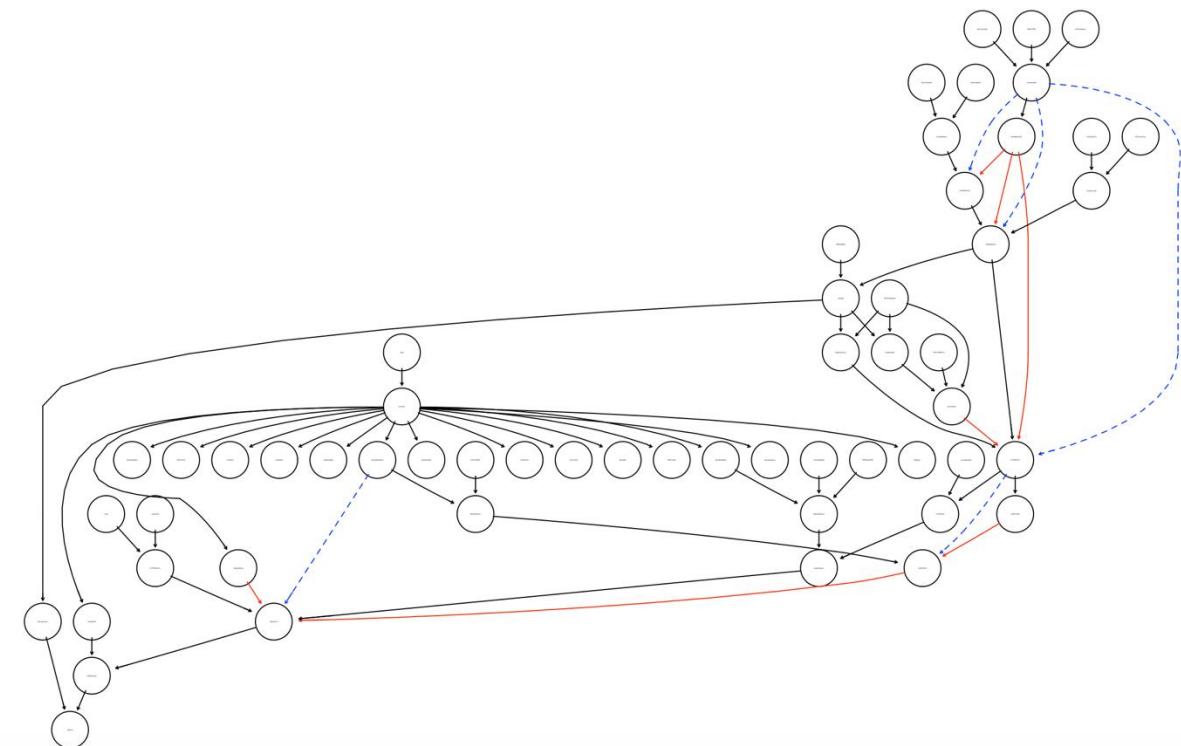
```
[1] 59
```

```
$fp
```

```
[1] 5
```

```
$fn
```

```
[1] 7
```



After seeing the graph we can see that the red line indicates the false positives which is 5 and the blue line denotes the false negatives which are 7 out of the 59 archs.

```
> compare(dag,bnet_bdefull)
```

```
$tp
```

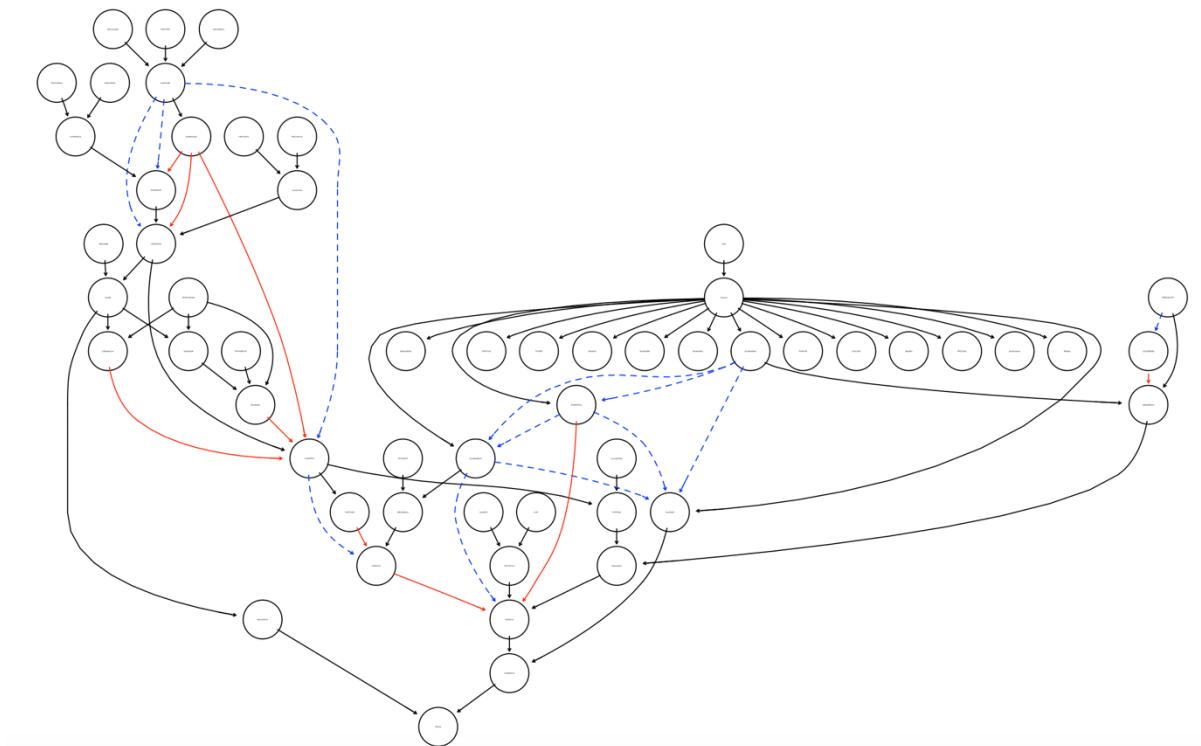
```
[1] 57
```

```
$fp
```

```
[1] 13
```

```
$fn
```

```
[1] 9
```



After seeing the graph we can see that the red line indicates the false positives which are 13 and the blue line denotes the false negatives which are 9 out of the 57 archs. The archs are more inconsistent in Bde scoring function if compared to the BIC scoring function.

c)

```
> fitted_params$CombClouds
```

Parameters of node CombClouds (multinomial distribution)

Conditional probability table:

```
, , IRCloudCover = Clear
```

| CombClouds | Clear       | Cloudy      | PC          |
|------------|-------------|-------------|-------------|
| Clear      | 0.980757483 | 0.073929961 | 0.496460825 |
| Cloudy     | 0.000000000 | 0.801556420 | 0.044910910 |
| PC         | 0.019242517 | 0.124513619 | 0.458628265 |

```
, , IRCloudCover = Cloudy
```

| CombClouds | Clear       | Cloudy      | PC          |
|------------|-------------|-------------|-------------|
| Clear      | 0.500905797 | 0.018248175 | 0.027590848 |
| Cloudy     | 0.087862319 | 0.941605839 | 0.442799462 |
| PC         | 0.411231884 | 0.040145985 | 0.529609690 |

```
, , IRCloudCover = PC
```

| CombClouds | Clear       | Cloudy      | PC          |
|------------|-------------|-------------|-------------|
| Clear      | 0.693900392 | 0.008705114 | 0.104643413 |
| Cloudy     | 0.019585898 | 0.871599565 | 0.103532548 |
| PC         | 0.286513710 | 0.119695321 | 0.791824039 |

- d) The probability that CombClouds is equal to Cloudy, given that the MeanRH is Very Moist and the IRCloudCover is Cloudy is equal to 0.379.

```
> cpquery(fitted_params, event = (CombClouds=="Cloudy"), evidence = ((MeanRH== "VeryMoist") & (IRCloudCover== "Cloudy")))
[1] 0.3793911
```

## Q.5)

a)

- i) The dataset used in the research paper is the road accident data set gathered from an Iranian highway, the Hamedan Qazvin highway. In a 56 month period there were about 65,535 road accidents recorded. The variables are the following :

- Sex
- Education
- Vehicle Type
- Age
- License Type
- Seatbelt status
- Injury Type

All the variable present in the dataset were discrete categorical variables.

- ii) The name of the algorithm which was used in this research paper is “**PC**” Algorithm in order to learn the network structures of the road accident dataset. There were two steps involved in the dataset ; the first one was to use the conditional independence test to structure the network and then it uses the d-separation method to determine the direction of nodes.

- iii) The name of the software/tool that was used to visualize the Bayesian network in the paper was called **Netica**. The link is provided as followed :  
<https://www.norsys.com/netica.html>

iv)

1) The probability of being injured while wearing seat belt and driving a car, knowing that the driver has a diploma degree and a type 2 driving license is 1.66%.

2) The probability of death while wearing seat belt and driving a car, knowing that the driver has a diploma degree and a type 2 driving license is 0.12%.

3) The probability of being injured while not wearing the seatbelt, knowing that the driver has a diploma degree and a type 2 driving license is 11.4%.

4) The probability of death while not wearing the seatbelt, knowing that the driver has a diploma degree and a type 2 driving license is 1%.

5) It could be concluded that if we use the seatbelt than the probability of being injured or resulting in a fatal crash could decrease.

b)

I)

In this article machine learning tools are used in order to show the reader the risk estimation in medical science known as Bayesian Networks. (Arora et al., 2019). The application is about how the Bayesian networks tend to be advantageous in regression based approaches. There are certain things which would be covered in this article such as generation of the network structures, their ability to apply the Bayes theorem for risk estimation and finally the transformation of into decision models (Arora et al., 2019). Moreover, the application is about how the Bayesian networks would play their role in order to analyse health economics outcomes in this era of medicine. (Arora et al., 2019).

II)

In this paper (Arora et al., 2019) the methods used were Bayesian networks in order to predict the risks. There were three techniques which were introduced in the (Arora et al., 2019) which were purely expert elicited BN, machine learned BN or a combined approach. BN present directed acyclic graphs which then comprise to be a network of nodes connected by edges. The characteristics of the patients were stored and the likelihood of the disease was to be predicted through the passing on to the nodes through the network. Conditional dependence was computed and linked to various tests in order to see the characteristics of the patients leading to diseases like tuberculosis and lung cancer. Moreover, then the algorithm of Expected Maximization (EM) was used in order to compute the parameter of the Bayesian Network with some missing data dependent on some other variables. This algorithm is comprised of two steps which are; the first predicts a potential value for the missing data and then in the second step the imputed dataset is used to calculate the maximum likelihood estimate in the model (Arora et al., 2019). With this technique the learning and efficiency of the model would not be affected in the case of missing data present in the data set.

**References :**

Arora, P., Boyne, D., Slater, J., Gupta, A., Brenner, D. and Druzdzel, M., 2019. Bayesian Networks for Risk Prediction Using Real-World Data: A Tool for Precision Medicine. *Value in Health*, 22(4), pp.439-445.