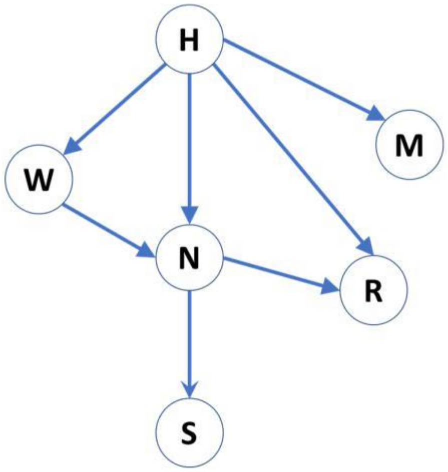
Q1) [31]

Weather conditions influence the production of good quality coffee in a region. A list of factors that influence the coffee cultivation, along with their possible values, and a Bayesian network that represents the relationship between these factors (variables) are given below.

M (Maximum Temperature) ∈ { < 20, 20-30, 30-40, > 40 }
N (Minimum Temperature) ∈ { < 0, 0-10, 10-20, > 20}
W (Wind speed) ∈ {Low, Medium, High}
H (Relative humidity) ∈ {< 50, 50-60, > 60}
R (Precipitation) ∈ {Low, High}
S (Solar radiation) ∈ {Low, Medium, High}



1.1) Write down the joint distribution

for the above network.

1.2) Find the minimum number of parameters required to fully specify the distribution according to the above network.

1.3)

a) Write down a joint probability density function if there are **no independence among the variables is assumed.**
b) How many parameters are required, at a minimum, if there are **no independencies among the variables is assumed?**
c) Compare with the result of the above question (Q1.2) and comment.

1.4) ***d-separation*** method can be used to find two sets of independent or conditionally independent variables in a Bayesian network. For **each of the statements** given below from (a) to (c), perform the following:

□ List **all** the possible paths from the first (set of) node/s to the second (set of) node/s.
□ State if each of those paths is *blocking* or *non-blocking* **with reasons**.
□ Hence, mention if the statement is **true** or **false**.

a) $\perp S | \varnothing$ (M is marginally independent of S)

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

c) ,
$\perp W | H$

1.5) Write a R-Program to produce the above Bayesian network, and perform the d-separation tests for all of the above cases mentioned in Q1.4 (a) to (c). Show the **plot of the network** you obtained and the **output (of d-separation test)** from your program.

1.6)

a) Show the step by step process to perform **variable elimination** to compute P(W|S=Low,R=Low). Use the following variable ordering for variable elimination process N,H,M.

. Use the following variable ordering for the elimination process:
N, H, M.

b) What is the treewidth of the network, given the above elimination ordering?

[Marks 2+4+5+10+3+7 = 31]

Q2) [16] **Implementing a Bayesian network in R and performing inference**

A belief network models the relation between the variables ***A, B, C, D and E,*** which represents the *season, river flow rate, fish species, color* and *size* respectively. Each variable takes different states as given below.

!" #$

∈ %!&, '()

* (+ ,!(−.#% ("&!

∈ .#%, + 0

1 −+ *h* 2!3+ !

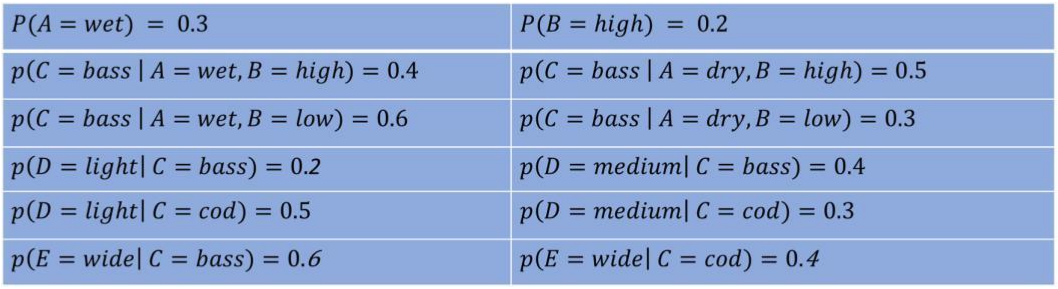
∈ 4" , 5#'

6 3#.#7(

∈ .+ 0 &, 8!'+ 78, '"(9 : + ;!

∈ %+ '!, & + $

The belief network that models these variables has (probability) tables as shown below.



$P(A = wet) = 0.3$	$P(B = high) = 0.2$
$p(C = bass \mid A = wet, B = high) = 0.4$	$p(C = bass \mid A = dry, B = high) = 0.5$
$p(C = bass \mid A = wet, B = low) = 0.6$	$p(C = bass \mid A = dry, B = low) = 0.3$
$p(D = light \mid C = bass) = 0.2$	$p(D = medium \mid C = bass) = 0.4$
$p(D = light \mid C = cod) = 0.5$	$p(D = medium \mid C = cod) = 0.3$
$p(E = wide \mid C = bass) = 0.6$	$p(E = wide \mid C = cod) = 0.4$

2.1) Use the below libraries in R to create this belief network in R along with the probability values, as shown in the above table.

You may use the following **libraries** for this:

```
#https://www.bioconductor.org/install/

#BiocManager::install(c("gRain", "RBGL", "gRbase"))

#BiocManager::install(c("Rgraphviz"))

library("Rgraphviz")

library(RBGL)

library(gRbase)

library(gRain)

#define the appropriate network and use the "compileCPT()"function to Compile list of conditional probability tables, and create the network.
```

a) Show the obtained **belief network** for this distribution

b) Show the probability tables **obtained from the R output,** (and verify with the above table).

2.2) Use R program to compute the following probabilities:

a) Given that the **river flow rate** is *low*, what is the probability that **size** is *thin*?

b) Given that the **colour** is *dark* and the **season** is *dry*, what is the probability that the **fish species** is **Cod**?

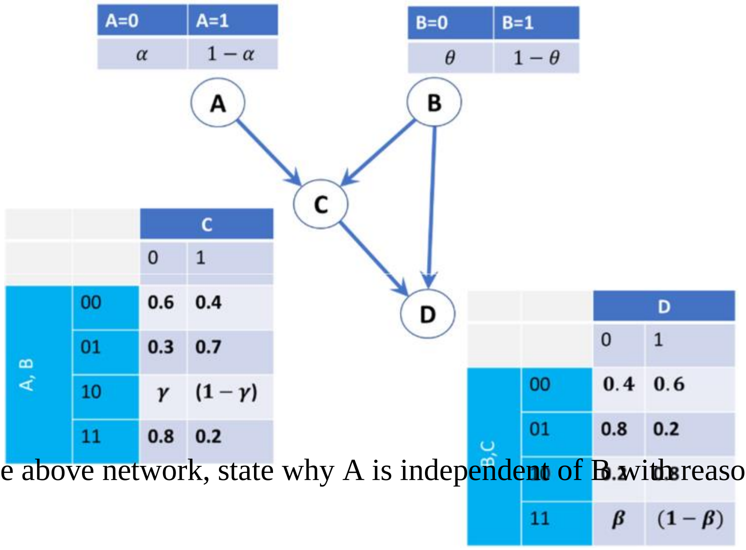
c) Find the joint distribution of **colour** and **fish species.**

d) Find the marginal distribution of ***fish species.***

[Marks: (3+5) + (2+2+2+2) = 16]

Q3) [15]

Consider four **binary** variables A, B, C, D. The Directed Acyclic Graph (DAG) shown below describes the relationship between these variables along with their conditional probability tables (CPT).



A=0	A=1
α	$1-\alpha$

B=0	B=1
θ	$1-\theta$

		C	
		0	1
A, B	00	0.6	0.4
	01	0.3	0.7
	10	γ	$(1-\gamma)$
	11	0.8	0.2

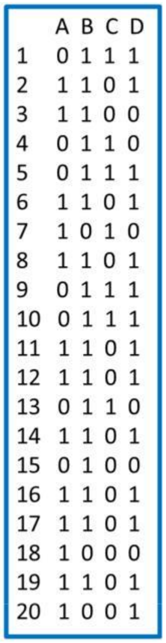
		D	
		0	1
B,C	00	0.4	0.6
	01	0.8	0.2
	10	0.2	0.8
	11	β	$(1-\beta)$

3.1) In the above network, state why A is independent of B with reasons, i.e., A⊥B.

3.2) Hence, obtain an expression (in a simplified form) for 6 >| >, * >

in terms of ? only.

3.3) The table shown below provides 20 simulated data obtained for the above Bayesian network. Use this data to find the maximum likelihood estimates of @, ?, A and B.



	A	B	C	D
1	0	1	1	1
2	1	1	0	1
3	1	1	0	0
4	0	1	1	0
5	0	1	1	1
6	1	1	0	1
7	1	0	1	0
8	1	1	0	1
9	0	1	1	1
10	0	1	1	1
11	1	1	0	1
12	1	1	0	1
13	0	1	1	0
14	1	1	0	1
15	0	1	0	0
16	1	1	0	1
17	1	1	0	1
18	1	0	0	0
19	1	1	0	1
20	1	0	0	1

3.4) Find the value of 6 >| >, * >

using the values obtained for ? from the above question Q3.3.

[Marks 3+ 7 + 4 + 1 = 15]

Q4) Bayesian Structure Learning [27]

For this question, you will be using a dataset, called **“*hailfinder”*** available from the ‘bnlearn’ R package. which contains 56 variables. This has meteorological data.

Use the following R code to load the ***hailfinder*** dataset:

```
library (bnlearn)
# load the data.
data(hailfinder)
summary(hailfinder)
```

The ***true network structure*** of this dataset can be viewed (plot) using the following R code.

```
library(bnlearn)
# create and plot the network structure.
modelstring          =          paste0("[N07muVerMo][SubjVertMo][QGVertMotion][SatContMoist][RaoContMoist]",
             "[VISCloudCov][IRCloudCover][AMInstabMt][WndHodograph][MorningBound][LoLevMoistAd][Date]",
             "[MorningCIN][LIfr12ZDENSd][AMDewptCalPl][LatestCIN][LLIW]",
             "[CombVerMo|N07muVerMo:SubjVertMo:QGVertMotion][CombMoisture|SatContMoist:RaoContMoist]",
             "[CombClouds|VISCloudCov:IRCloudCover][Scenario|Date][CurPropConv|LatestCIN:LLIW]",
             "[AreaMesoALS|CombVerMo][ScenRelAMCIN|Scenario][ScenRelAMIns|Scenario][ScenRel34|Scenario]",
             "[ScnRelPlFcst|Scenario][Dewpoints|Scenario][LowLLapse|Scenario][MeanRH|Scenario]",
             "[MidLLapse|Scenario][MvmtFeatures|Scenario][RHRatio|Scenario][SfcWndShfDis|Scenario]",
             "[SynForcng|Scenario][TempDis|Scenario][WindAloft|Scenario][WindFieldMt|Scenario]",
             "[WindFieldPln|Scenario][AreaMoDryAir|AreaMesoALS:CombMoisture]",
             "[AMCINInScen|ScenRelAMCIN:MorningCIN][AMInsWliScen|ScenRelAMIns:LIfr12ZDENSd:AMDewptCalPl]",
             "[CldShadeOth|AreaMesoALS:AreaMoDryAir:CombClouds][InsInMt|CldShadeOth:AMInstabMt]",
             "[OutflowFrMt|InsInMt:WndHodograph][CldShadeConv|InsInMt:WndHodograph][MountainFcst|InsInMt]",
             "[Boundaries|WndHodograph:OutflowFrMt:MorningBound][N34StarFcst|ScenRel34:PlainsFcst]",
             "[CompPlFcst|AreaMesoALS:CldShadeOth:Boundaries:CldShadeConv][CapChange|CompPlFcst]",
             "[InsChange|CompPlFcst:LoLevMoistAd][CapInScen|CapChange:AMCINInScen]",
             "[InsSclInScen|InsChange:AMInsWliScen][R5Fcst|MountainFcst:N34StarFcst]",
             "[PlainsFcst|CapInScen:InsSclInScen:CurPropConv:ScnRelPlFcst]")

dag = model2network(modelstring)
par(mfrow = c(1,1))
#BiocManager::install(c("Rgraphviz"))
graphviz.plot(dag)
```

Use R programming, as appropriate, to answers the following questions.

4.1) Use the *hailfinder* dataset to learn Bayesian network structures using **hill-climbing (hc) algorithm**, utilizing two different scoring methods, namely **Bayesian Information Criterion score (BIC score)** and the **Bayesian Dirichlet equivalent (Bde score), for each of the** following **sample sizes** of the data**:**

a) **100 (first 100 data)**
b) **1000 (first 1000 data)**
c) **10000 (first 10000 data)**

For each of the above cases,

- □ provide the scores obtained for BIC and BDe,
- □ Plot the network structure obtained for the BIC and BDe scores.

4.2) Based on the results obtained for the above question (Q 4.1), discuss how the BIC score compare with BDe score for different sample sizes in terms of **structure** and **score** of the learned network.

4.3)

a) Find the Bayesian network structures utilising the **full dataset, and using both BIC and Bde scores.** Show the scores and the obtained networks.

b) **C**ompare the networks obtained above (in Q4.3.a) for each BIC and Bde scoring methods with the ***true network structure*** and **comment**. Use the “compare()” function and “graphviz.compare()” function available in the “bnlearn” R package to perform these comparisons and comment.

c) Fit the data to the network obtained using the **BIC score** in the above question (Q4.3.a) in order to compute the conditional probability distribution table entries (CPD table values). Show the obtained CPD table entries for the variable “**CombClouds**”.

d) Use the above learned network obtained (in Q4.3.c) to find the probability of :
P(CombClouds ="Cloudy" | MeanRH = "VeryMoist", IRCloudCover ="Cloudy")

[Marks (3*4) + 3 + (4+3+3+2) = 27]

Q5) Research based questions (Practical applications in real world) [11]

a) Download the following article from the link provided below. Read that article and answer the following questions. This article provides a real life case study on creating and using a Bayesian network for road accident data analysis.

Ali Karimnezhad & Fahimeh Moradi (2017), **Road accident data analysis using Bayesian networks, Transportation Letters**, 9:1, 12-19,
DOI: 10.1080/19427867.2015.1131960
Web: https://www.tandfonline.com/doi/full/10.1080/19427867.2015.1131960

Note that you will be able to download this paper ***via Deakin library using your Deakin credentials (username and password). (https://www.deakin.edu.au/library/help/add-browser-bookmarklet)***

i) Describe the dataset used for their analysis. What are the variables used? Are the variables numerical or categorical or mixed? How many records of data have been used?

ii) What is the name of the algorithm used for learning the Bayesian network structure?

iii) What software tool have been used to build and visualize the Bayesian network? Provide a web link to that software.

iv) Read the section titled "*Parameter learning in the road accident network*" in that paper and extract the following probability values that they have computed, and mention them:

I. 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.

II. 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

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

IV. The probability of death while not wearing the seatbelt, knowing that the driver has a diploma degree and a type 2 driving license

V. Based on the probability values obtained above, what conclusions are made?

b) Do a research (using journal or conference papers/publications) and describe **ONE other real-world application** of any Bayesian methods/Bayesian networks. Your description should include the following:

i) Briefly describe on your own words ***what the application is about***.

ii) ***The details of the techniques used***.

Provide references for the applications/papers used. **Description for this question Q5(b) should NOT exceed 400 words (including references).**

NOTE: Your answers for all of the above questions must be written in your own words. Copying directly from the paper/reference text will constitute to Plagiarism and zero.

[Marks 7 + 4 =11]