DATA-ANALYTICS ECS784P
Coursework-2

Q1.)

I have picked a dataset from Kaggle [1] and its termed as 'Personal Key Indicators of Heart Disease'. This dataset holds 18 variables, but I have decided to pick only 12 important variables out of it, and they are the factors which leads to heart disease. For example: AlcholDrinking, smoking, stroke,etc. In addition, some factors can have a direct impact. Factors such as GenHealth, BMI, Stroke and Smoking have a significant impact on others, connected to each other. A sedentary lifestyle can lead to heart disease, even I people without other risk factors. It can also increase the chance of developing other risk factors for heart disease like high blood pressure, obesity, and cholesterol. Based on the dataset I determined that it will be the best fit for learning structure.
In addition, we will investigate how the structure learning algorithm learns.

HeartDisease	BMI	Smoking	AlcoholDrinking	Stroke	Sex	AgeCategory	Diabetic	PhysicalActivity	GenHealth
No	overweight	No	No	No	Female	60-64	No	Yes	Excellent
Yes	overweight	No	No	No	Female	65-69	No	Yes	Very good
No	overweight	No	No	No	Male	50-54	No	No	Good
No	overweight	No	No	No	Female	70-74	No	Yes	Very good
No	overweight	No	No	No	Female	25-29	No	Yes	Very good
Yes	overweight	No	No	No	Male	70-74	No	No	Good
No	overweight	No	No	No	Male	18-24	No	Yes	Excellent
No	overweight	Yes	No	No	Male	50-54	No	Yes	Good
Yes	overweight	No	No	No	Male	50-54	Yes	Yes	Very good
No	overweight	Yes	No	No	Female	35-39	No	Yes	Fair
Yes	overweight	Yes	No	No	Female	75-79	Yes	Yes	Fair
Yes	overweight	Yes	No	Yes	Male	50-54	No	No	Good
Yes	overweight	No	No	No	Female	65-69	No	Yes	Good
Yes	overweight	Yes	No	No	Female	70-74	Yes	Yes	Fair
Yes	obese	Yes	No	No	Male	80 or older	No	Yes	Good
Yes	obese	Yes	No	No	Male	65-69	No	No	Good
No	healthy weight	No	No	No	Female	35-39	No	Yes	Very good

Figure 1. Personal Key Indicators of Heart Disease

Q2.)

Based on some references and own knowledge about Heart Disease. I have created a graph using 12 variables. From the graph shown below it can be seen that GenHealth is one of the major reasons for Heart Disease.

According to NHS-UK [2] causes of Heart Disease include Smoking, PhysicalActivity, Diabetic and SleepTime.
It has been proven that overweight and obesity which affects BMI can cause many serious health issues and can also increase the risk of Heart Disease and stroke, the following information is extracted from the given reference [3].

It is reported by British Heart Foundation [4] that AlcholDrinking is one of the reasons which increases the risk of Stroke and Diabetic and it affects the BMI as well.

Smoking influences/triggers Heart Disease through Asthma. However, Smoking and Heart Disease are independent given Asthma. Smoking along with PhysicalActivity, Obesity (BMI) and Diabetic, tops the list as a primary risk factor of Heart Disease shown in the reference [5]

Based on my own knowledge Sex(male) are more likely than woman(female) to develop Daibetic. In general, it can also be seen that SleepTime is the common cause of BMI and GenHealth which causes the Heart Disease.

From the given reference (National Institute of Aging)[6] it has been proven that Diabetes is a serious disease, and it affects many older adults, AgeFactor affects diabetes.

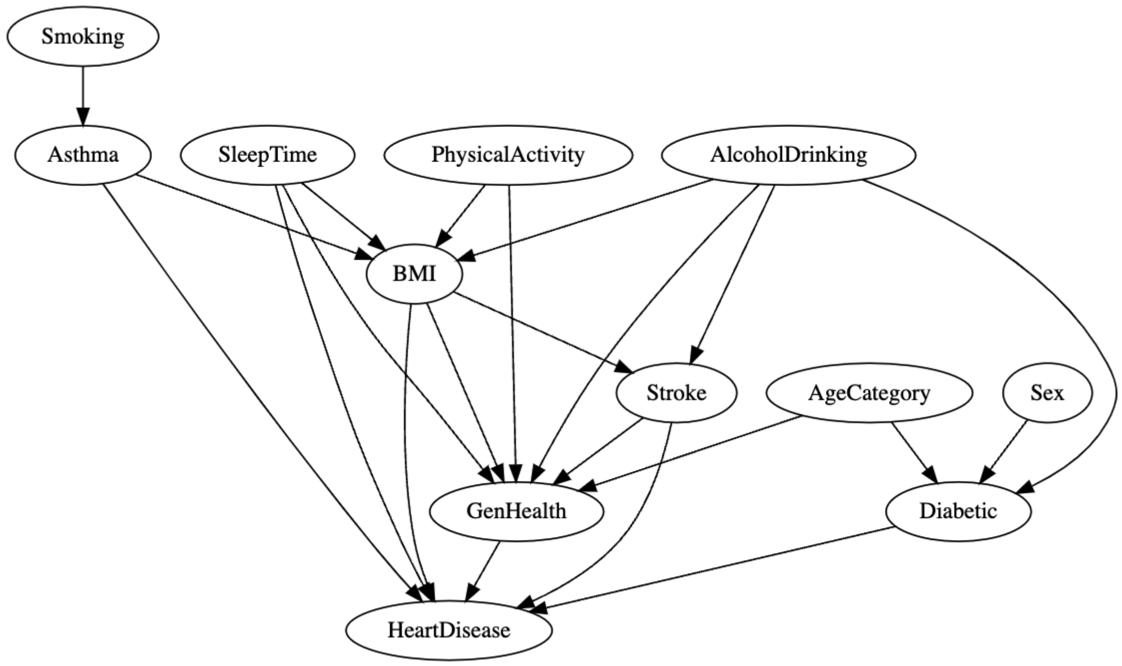

Figure 2. Knowledge-based graph

Q3)

Algorithm	CPDAG scores			Log-Likelihood (LL) score	BIC score	#free parameters	Structure learning elapsed time
	BSF	SHD	F1				
HC_CPDAG	-0.205	35.500	0.184	-685335.144	-687680.986	299	1 Second
HC_DAG	-0.159	34.500	0.224	-685245.568	-687889.544	337	1 Second
TABU_CPDAG	-0.205	35.500	0.184	-685335.144	-687680.986	299	1 Second
TABU_DAG	-0.068	32.000	0.280	-685259.611	-687597.607	298	4 Seconds
SaiyanH	-0.091	30.000	0.195	-688674.739	-689843.738	149	5 Seconds
MAHC	-0.205	35.000	0.170	-685628.568	-687699.813	264	6 Seconds

Table Q3. The scores of the six algorithms

From the table above, the BSF scores for all the algorithms are lower than the average given in the manual. It implies that learned graph of my model matches the true graph worse.

My F1 scores are almost low which implies that general accuracy is lower than the average accuracy shown in Bayesys manual. Moreover, my runtime results are lower than the average runtime shown in the manual for all the algorithms, and this is due to differences in sample size, hardware, and node count.

From the given SHD scores, it can be observed that my results are lower for this as well when compared with the Bayeys manual, the reason is that my SHD results are based on a single dataset of 12 nodes, while the results shown in the manual are from different datasets with up to 109 nodes, I think it's because of that.

This is my predicted result due to the fact I anticipate that there may be the inconsistency, it could be because of my knowledge graph, and it may also be due to my dataset which contains a lot of data noise due to real word data, subsequently it affects the overall performance of the algorithms.

Q4)

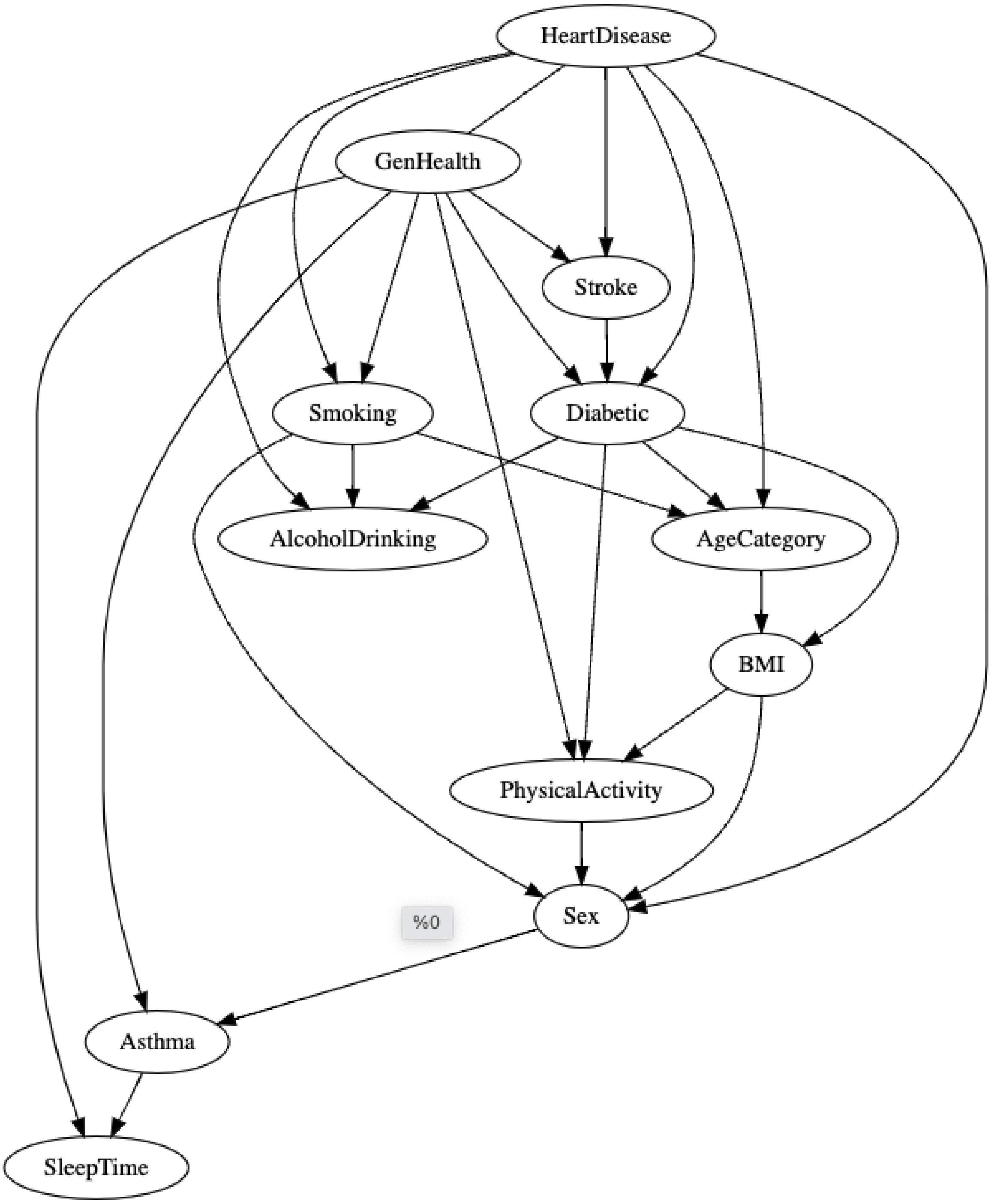

Figure 4. CPDAG generated by HC_CPDAG

From the Figure 4 shown above it can be seen that CPDAG has one edge less than the actual graph. And we can also say that GenHealth and Diabetics are the most common cause of other factors.

The first causal class, the causal chain, is represented by three nodes: HeartDisease, Stroke and Diabetics. In particular, HeartDisease affects Diabetics through stroke.

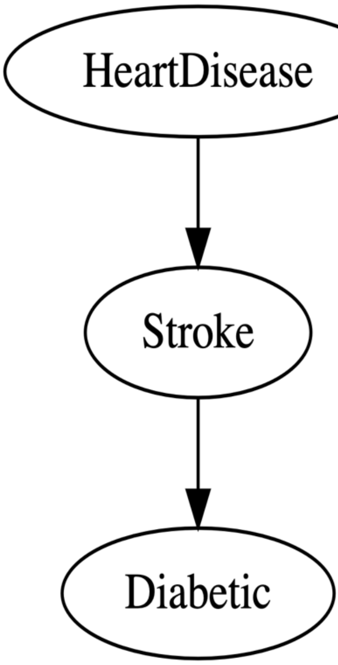

Figure 4.1 Causal Chain

The second causal class, common causal chain, is represented by three nodes: GenHealth, Stroke and PhysicalActivity. And we can say that GenHealth is the common cause of Stroke and PhysicalActivity.

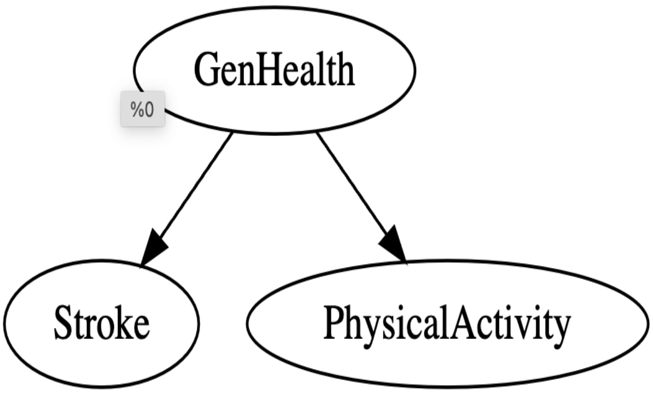

Figure 4.2 Common Causal

The third causal class, common effect chain, is represented by three nodes: Stroke and GeneralHealth cause Diabetic.

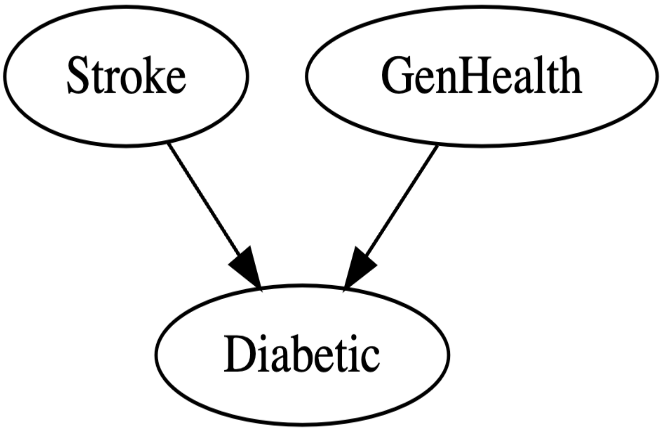

Figure 4.3 Common Effect

Q5)

Rank	Your ranking			Ranking according to the Bayesys Manual		
	BSF [single score]	SHD [single score]	F1 [single score]	BSF [average score]	SHD [av. normalised score]	F1 [average score]
1	TABU_DAG [-0.068]	HC_CPDAG [35.500]	TABU_DAG [0.280]	TABU_CPDAG [0.533]	MAHC [0.481]	SaiyanH [0.576]
2	Saiyant [-0.091]	TABU_CPDAG [35.500]	HC_DAG [0.224]	SaiyanH [0.515]	TABU_CPDAG [0.44]	TABU_CPDAG [0.564]
3	HC_DAG [-0.159]	MAHC [35.000]	Saiyant [0.195]	HC_CPDAG [0.506]	SaiyanH [0.438]	MAHC [0.562]
4	HC CPDAG [-0.205]	HC DAG [34.500]	HC_CPDAG [0.184]	MAHC [0.499]	HC_CPDAG [0.402]	HC_CPDAG [0.537]

5	TABU_CPDAG [-0.205]	TABU DAG [32.000]	TABU_CPDAG [0.184]	TABU_DAG [0.484]	TABU_DAG [0.397]	TABU_DAG [0.53]
6	MAHC [-0.205]	Saiyant [30.000]	MAHC [0.170]	HC_DAG [0.438]	HC_DAG [0.314]	HC_DAG [0.479]

Table Q5. Rankings of the algorithms on my dataset

As shown in the table above, my ranking does not match the ranking significantly given in the Bayesys manual. As the observation show the TABU_CPDAG performs the best BSF score for the given datasets in Bayesys manual but in my case, it produced the worst score for my dataset. This is my expected result; I believe that six different techniques will have different impact on different datasets.
The scores in Bayesys manual are evaluated from six different synthetic datasets, whereas my dataset is real data from medicine.
The size also affects the result of the algorithm. The number of nodes and samples of dataset also affect the algorithms results. Therefore, this contradiction is understandable.

Q6)

From the observation my structural learning time does not match the results shown in the Bayesys manual.
It is because number of nodes in my dataset has 12 variables, and this is different from the number of variables in the record in the Bayesys manual. And another reason is that the number of data samples in my dataset is around 52,000 which is very different from the dataset given in the manual.

Thus, it strongly affects the execution time of structural learning.
Furthermore, data noise, probably inconsistent data also has a significant effect on the execution time of algorithm. additionally, differences in hardware also affect the runtime.

Q7)

Algorithm	Your task 4 results				Your Task 5 Results		
	BIC score	Log Likelihood	Free parameters	Algorithm	BIC Score	Log likelihood	Free parameters
Your knowledge-based graph	- 703680.986	- 714335.144	203	HC_CPDAG	- 687680.986	- 685335.144	299
				HC_DAG	- 687889.544	- 685245.568	337
				TABU_CPDAG	- 687680.986	- 685335.144	299
				TABU_DAG	- 687597.607	- 685259.611	298
				SAIYANH	- 689843.738	- 688674.739	149
				MAHC	- 687699.813	- 685628.568	264

Table Q7. BIC scores, Log-Likelihood (LL) scores and number of free parameters

For task 4, I used the DAGleanred file which is same as DAGtrue file. Basically, the Log-Likelihood and BIC scores exhibit how effectively the parameter of the knowledge-based graph matches the training data, and the free parameters show the complication of the knowledge-based graph.

It can be seen from the above table that Log-Likelihood scores in task-5 are high (less negative) in values compared to the Task-4 and the number of free parameters is less in Task 4 as compared to Task 5. This means that by using the 6 algorithms the graphs learned are mostly fitting the training data and are more complex than the presented knowledge graph.

These are my predicted results as my dataset is from the medical dataset and it has been collected by analysis and inquiry and my dataset may contain unpredictability. If taken an example the if we see the learned graph in Q4 we can observe that GenHealth causes smoking and Diabetic causes AgeCategory which is at odds with the knowledge-based graph and its senseless as far as I know.

Q8)

Knowledge Approach	CPDAG Scores			LL	BIC	Free parameters	Number of edges	Runtime
	BSF	SHD	F1					
Without knowledge	-0.205	35.500	0.184	-685335.144	-687680.986	299	22	1
With knowledge (List 1st knowledge approach)	0.114	28.500	0.434	-685770.774	-689560.212	483	22	1
With knowledge (List 2nd)	0.023	29.000	0.320	-685398.175	-688057.842	339	22	1

Table Q8. HC_CPDAG scores

The first knowledge approach I used is Directed, it adds some constraints based on the knowledge graph that shows the direct relationship between the nodes.

As shown in Table Q8, the BSF, SHD and F1 scores that result from the HC_CPDAG algorithm integrated with the indicated approach are better than those without knowledge.

This implies that this knowledge approach helps the trained graph match the actual graphs better. Moreover, the value of the free parameters is greater than without knowledge, but the number of edges and execution time are same as without the knowledge. This means that the graph of learned knowledge usage is less complex due to the reduced BIC. Also, the generated LL value is lower when knowledgeable(instructed) than when not knowledgeable and fits the data when the trained graph is knowledgeless.

The second knowledge approach I used is undirected. This approach defines a relationship between 2 nodes but does not know the direction of the relationship. Similarly, the results obtained with this method for BSF, SHD and F1 scores outperform those obtained without knowledge.

This meets my expectations as my intention to apply knowledge base constraints is to help the algorithm bring the trained graph closer to the actual graph, improve accuracy and reduce execution time, my dataset can be noisy, so omitting the knowledge technique will help the trained graph fits the data rather than force it.

REFERENCES:

[1] Personal **Key Indicators of Heart Disease**

[2] coronary heart disease

[3] **Obesity**

[4] **Effects of alcohol on your heart**

[5] Smoking and Cardiovascular Disease

[6] **Diabetes in Older People**