COURSEWORK – 2: CAUSAL STRUCTURE LEARNING

ECS784U/ECS784P - Thaarik Ahamed Ahemed Ali -220207760

QUESTION 1:

A stroke is a medical disorder that harms the brain by rupturing blood vessels and cutting off the flow of blood there. According to the World Stroke Organisation, 13 million people get strokes each year. But as a result, almost 5.5 million patients pass away [1]. To predict this health issue, a dataset of 11 variables, and sample size of 4981 including behavioural and lifestyle risk factors, was created. The dataset was obtained from Kaggle - Brain Stroke Dataset. The appropriate columns in the dataset include discretized categorical values from the discretized dataset. Figure 1.1 shows the dataset after cleaning. Since the variables in this dataset might influence the likelihood that a stroke will occur, establishing a causal relationship between these factors is necessary. The probabilistic causal relationship between hypertension, BMI, and smoking behaviours, as well as providing insight on how residence type and married status lead to stroke attacks, are learned by the structure learning algorithm.

gender	age	hypertension	bmi	avg_glucose_level	heart_disease	ever_married	work_type	Residence_type	smoking_status	stroke
Male	60-80	0	obesity	high	1	Yes	Private	Urban	formerly smoked	1
Male	60-80	0	obesity	normal	1	Yes	Private	Rural	never smoked	1
Female	40-60	0	obesity	high	0	Yes	Private	Urban	smokes	1
Female	60-80	1	normal	high	0	Yes	Self-employed	Rural	never smoked	1
Male	80+	0	overwieght	high	0	Yes	Private	Urban	formerly smoked	1
Male	60-80	1	overwieght	normal	1	Yes	Private	Rural	never smoked	1
Female	60-80	0	normal	normal	0	No	Private	Urban	never smoked	1
Female	60-80	0	normal	low	0	Yes	Private	Urban	Unknown	1
Female	80+	1	overwieght	normal	0	Yes	Private	Rural	never smoked	1
Female	60-80	0	obesity	normal	1	Yes	Govt_job	Rural	smokes	1
Female	40-60	0	overwieght	normal	0	Yes	Private	Urban	smokes	1
Female	60-80	0	overwieght	high	1	Yes	Private	Urban	never smoked	1
Female	40-60	1	obesity	high	0	Yes	Self-employed	Rural	never smoked	1
Male	60-80	0	obesity	high	1	Yes	Private	Urban	smokes	1
Male	60-80	1	overwieght	high	0	Yes	Private	Urban	smokes	1
Female	40-60	0	obesity	normal	0	No	Private	Urban	never smoked	1
Female	60-80	0	normal	high	0	Yes	Govt_job	Rural	smokes	1
Female	40-60	1	obesity	high	0	Yes	Self-employed	Urban	never smoked	1
Female	60-80	0	overwieght	high	0	Yes	Self-employed	Urban	never smoked	1

Figure 1.1

QUESTION 2:

The information provided in this knowledge-based DAG(Figure 2.1 – DAGTrue.csv, Figure 2.2: DAGTrue graph) is based on a lot of literature in this medical field and on a personal knowledge-based approach. Both age and gender influence hypertension [2] [3](high blood pressure) and abnormal glucose levels [4] [5](avg_glucose_level) that result in diabetes. Depending on their age and gender, a person might have different heart diseases probabilities [6] [7](heart_disease). The age of the individual influences his or her marital status (ever_married) and type of work (work_type). The marital status and the type of works in shifts influence the BMI [8] [9]value, which has an association with a sedentary lifestyle. In recent studies [10], it was found that the rural residence type (urban or rural) has a higher number of smokers, which shows the association between residence type (Residence_type) and smoking status (smoking_status). With a high BMI and smoking history, the study shows that the blood vessels get damaged or thicken, resulting in an increase in blood pressure, and paving the way to hypertension [11]. A high level of abnormal glucose (diabetes) and the signs of hypertension damage the heart, block the blood vessels and result in heart disease [12]. If the person has heart disease, then there is an increased risk of a stroke attack [13].

ID	Variable 1	Dependency	Variable 2
1	gender	->	heart_disease
2	gender	->	avg_glucose_level
3	age	->	hypertension
4	age	->	heart_disease
5	age	->	work_type
6	age	->	ever_married
7	age	->	avg_glucose_level
8	work_type	->	bmi
9	ever_married	->	bmi
10	Residence_type	->	smoking_status
11	bmi	->	hypertension
12	avg_glucose_level	->	heart_disease
13	hypertension	->	heart_disease
14	smoking_status	->	hypertension
15	heart_disease	->	stroke

Figure 2.1 - DAGTrue.csv

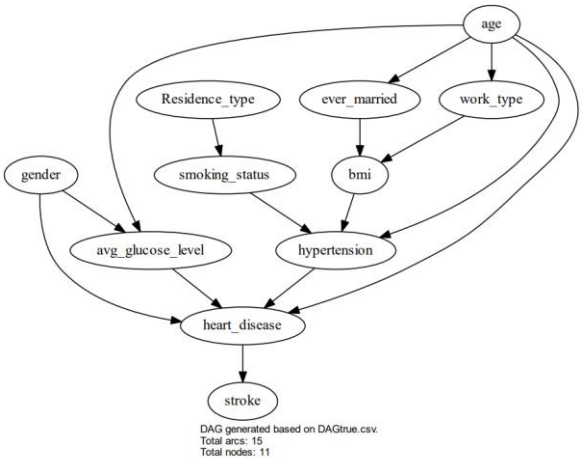


Figure 2.2 -DAGTrue Graph (Knowledge based graph)

QUESTION 3:

Table Q3 - The scores of the five algorithms when applied to the data set.

Algorithm	CPDAG scores			Log-Likelihood (LL) score	BIC score	# of free parameters	Structure learning elapsed time (sec)
	BSF	SHD	F1				
HC	0.325	12.000	0.462	-55492.201	-56069.465	94	0
TABU	0.325	12.000	0.462	-55492.201	-56069.465	94	0
SaiyanH	0.367	12.000	0.500	-55522.585	-56105.991	95	0
MAHC	0.325	12.000	0.462	-55492.201	-56069.465	94	0
GES	0.325	12.000	0.462	-55492.201	-56069.465	94	0

All CPDAG scores are the same in the HC, TABU, MAHC, and GES algorithms, and the scores are higher in the SaiyanH algorithm except for the SHD score.

The CPDAG scores from the HC, TABU algorithm obtained from the model are slightly higher than the CPDAG scores under the 'Sports' network of Bayesys manual table, which contains 9 nodes, 15 edges, and a sample size of 1000. But the other three algorithms have a lower score than the one under the 'Sports' network. In the case of an SHD value in all algorithms, the penalty score is higher than the one present in the 'Sports'

Network row. When compared to the average scores in the manual table, the CPDAG scores of all algorithms in the learned model, including runtime, are low.

This variation in scores is expected because in our model there are 13 nodes, 15 edges, and a 4981 dataset sample size. By looking at the runtime and SHD score of the average scores in the table, we can assume that the average scores are calculated from low to highly complex models used in Bayesys manual, and comparatively, our model is less complex. The DAGtrue graph is constructed from the knowledge gained from the literature. So, there might be some independencies present between the nodes that are not associated with each other as per our dataset. This shows that the dataset we use might have different causal relationships between variables, and this must be sorted by applying some constraints.

QUESTION 4:

There are three causal classes present in the learned CPDAG generated from the HC algorithm (Figure 4.1,4.2). CPDAG contains both directed and undirected edges.

Causal Chain: This class is represented by the chaining of arrows between variables following in the same direction i.e., the arrows in each variable are pointed in the same direction. Example: Even though, there is an undirected edge between 'age' and 'work_type', 'age' influences 'work_type' (Equivalence class) and 'work_type' influences 'bmi'. So, 'age' -> 'work_type'->'bmi'.

Common Cause: This class occurs when two or more variables are influenced by a common cause. i.e., a common node pointing towards two or more nodes. Example: 'work_type' and 'avg_glucose_level' have undirected edges towards the 'age' node. But they are dependent and caused by the 'age' node and 'work_type' is independent to 'avg_glucose_level'. So, the common cause is as follows: 'avg_glucose_level'<-'age'->'work_type'.

Common Effect: This class occurs when two or more variables cause a single effect node, i.e., the arrows of two or more (cause)nodes pointing towards a common node(effect). Example: 'bmi' is dependent on 'work_type' and 'ever_married' variables. So, 'work_type'->'bmi'<-'ever_married'.

'Residence_type' node does not have edges linking to other nodes because of no direct or indirect causal relationship between the variables in the learned CPDAG.

ID	Variable 1	Dependency	Variable 2
1	age	-	hypertension
2	age	-	avg_glucose_level
3	age	-	heart_disease
4	age	-	work_type
5	age	-	stroke
6	heart_disease	->	gender
7	ever_married	-	age
8	ever_married	->	bmi
9	work_type	->	bmi
10	work_type	-	smoking_status
11	smoking_status	->	gender

Figure 4.1 - CPDAGlearned.csv

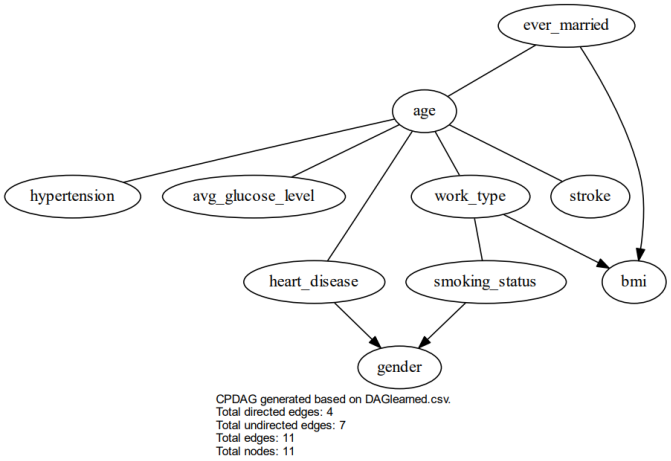


Figure 4.2 - CPDAGlearned Graph - learned graph

QUESTION 5:

Table Q5 - Rankings of the algorithms based on the data set, versus ranking of the algorithms based on the results shown in Table 3.1 in the Bayesys manual.

	Your rankings			**Rankings according to the Bayesys manual**		
Rank	**BSF [single score]**	**SHD [single score]**	**F1 [single score]**	**BSF [average score]**	**SHD [average score]**	**F1 [average score]**
1	HC [0.325], TABU [0.325], MAHC [0.325], GES [0.325]	HC [12.000], TABU [12.000], MAHC [12.000], GES [12.000], SaiyanH [12.000]	HC [0.462], TABU [0.462], MAHC [0.462], GES [0.462]	SaiyanH [0.559]	MAHC [50.96]	SaiyanH [0.628]
2	SaiyanH [0.367]	-	SaiyanH [0.500]	GES [0.506]	Saiyan [57.98]	MAHC [0.579]
3	-	-	-	MAHC [0.503]	HC [62.36]	GES [0.552]
4	-	-	-	TABU [0.499]	TABU [62.63]	TABU [0.549]
5	-	-	-	HC [0.498]	GES [63.3]	HC [0.548]

Under the BSF and F1 score columns, the HC, TABU, MAHC, and GES algorithms have the same scores of 0.325 and 0.462, respectively, ranking them in first place. The SaiyanH algorithm is ranked second under the BSF and F1 score columns with scores of 0.367 and 0.500, respectively. In the SHD column, all five algorithmic scores have the same value of 12.000 and hold first place.

The learned model's scores are lower than the average scores of the Bayesys manual. The observation shows that the ranking of the learned model is not consistent when compared to the ranking of BSF, SHD, and F1 average scores present in the Bayesys manual. This outcome is expected in my learning algorithm. The reason would be that the performance of the algorithm is the same, giving the same scores, but the performance is low in terms of the average scores in the Bayesys manual. The number of free parameters used, the dataset used in the learned model might contain noise, and imbalanced categorical values might exist in some

variables. Sometimes, there might be latent confounders that result in incorrect causal relationships in the learned model, decreasing the scores below the average Bayesys manual score.

QUESTION 6:

The elapsed structure learning time is around 0 seconds for all five algorithms, which is inconsistent with the average elapsed structure learning time of around 8 to 90 seconds in Bayesys module. This large elapsed timing depends upon the complexity of the dataset, variables, and features used in the calculation of the average score. In my case, the sample size is 4921 samples with 11 variables. This has less complexity when compared to the one used for the calculation of average scores affecting the elapsed structure learning time. It also depends on the hardware used, which influences the computational complexity.

QUESTION 7:

Table Q7. The BIC scores, Log-Likelihood (LL) scores, and number of free parameters generated by each of the five algorithms during Task 3 and Task 4.

Algorithm	Task 3 results			Algorithm	Task 4 results		
	BIC Score	LL Score	Free parameters		BIC Score	LL Score	Free parameters
Knowledge-based graph	-57810.634	-56471.872	218	**HC**	-56069.465	-55492.201	94
				TABU	-56069.465	-55492.201	94
				SaiyanH	-56105.991	-56105.991	95
				MAHC	-56069.465	-55492.201	94
				GES	-56069.465	-55492.201	94

As we know, the LL score is the likelihood score of the data in the model, and the BIC score shows the trade-off between the complexity of the model and its fitness.

In comparison (Table Q7), the BIC score and LL score of the learned graph in task 4 (application of five algorithms) have a slightly lower value than the BIC score and LL score of the knowledge-based graph as per task 3 (without any application of any algorithms). The number of free parameters in task 4 is lower than the one in task 3. This shows that the task 3 knowledge-based graph might result in poor performance due to overfitting with a high number of free parameters and higher complexity. After the application of five algorithms in task 4, the number of free parameters reduced, resulting in a balance in model performance and a good fit that helped us find the underlying causal relationship between the variables.

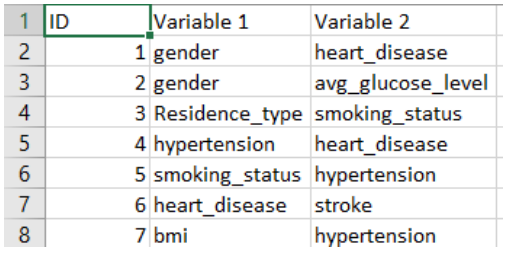
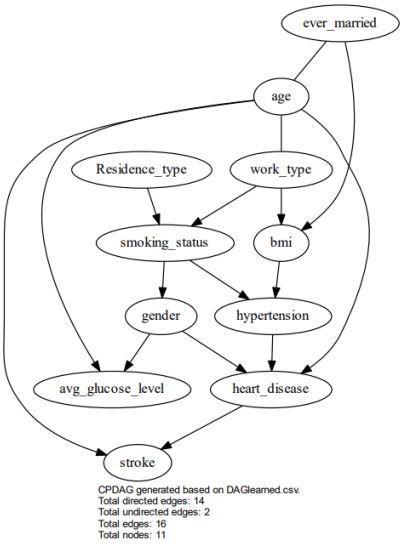
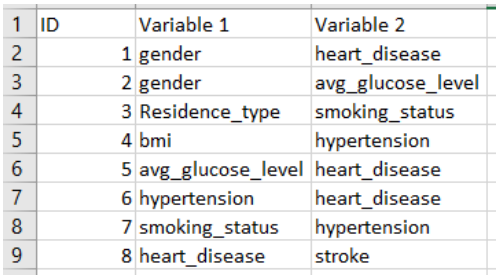
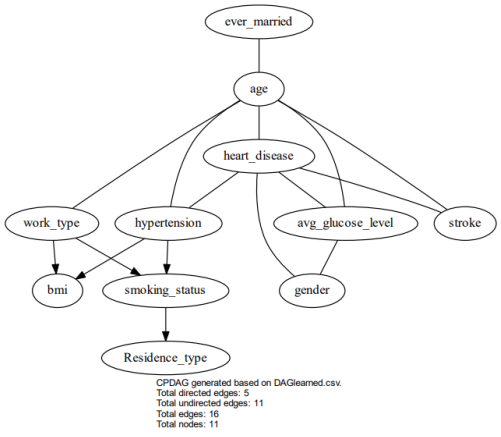
The results obtained are consistent, as expected, and could produce a better model fit that captures the dependencies between the variables. But further optimisations by application of constraints methods to generate the learned graph more accurately as per the given knowledge conditions.

QUESTION 8:

Table Q8.1. The scores of HC applied to your data, with and without knowledge.

Knowledge approach	CPDAG Scores			LL	BIC	Free parameters	Number of edges	Runtime (seconds)
	BSF	SHD	F1					
Without knowledge	0.325	12.000	0.462	-55492.201	-56069.465	94	11	0
With knowledge – Directed	0.792	5.000	0.839	-55592.988	-56471.167	143	16	0
With Knowledge - Undirected	0.683	6.000	0.733	-55851.041	-56544.986	113	15	0

Table Q8.2 The scores of HC applied to your data, with and without knowledge.

Knowledge approaches used	Constraints in csv	CPDAG learned after knowledge approach with constraints
Directed constraint		
Undirected constraint		

	ID	Variable 1	Variable 2
1			
2	1	gender	heart_disease
3	2	gender	avg_glucose_level
4	3	Residence_type	smoking_status
5	4	hypertension	heart_disease
6	5	smoking_status	hypertension
7	6	heart_disease	stroke
8	7	bmi	hypertension

	ID	Variable 1	Variable 2
1			
2	1	gender	heart_disease
3	2	gender	avg_glucose_level
4	3	Residence_type	smoking_status
5	4	bmi	hypertension
6	5	avg_glucose_level	heart_disease
7	6	hypertension	heart_disease
8	7	smoking_status	hypertension
9	8	heart_disease	stroke

We have applied two knowledge approaches: directed (a) and undirected (b). The constraints of these two knowledge approaches are presented above (Table Q8.1, Table Q8.2).

The reason for choosing the directed and undirected knowledge approaches with the above constraints is based on the literature knowledge which plays a crucial role in determining the stroke.

With the application of directed constraints, the BSF score and F1 score increased from 0.325 to 0.792 and 0.462 to 0.839, respectively. This shows that the accuracy and performance of the model are improved with a better fit. The SHD value decreased from 12.0000 to 5.000, showing that there is a decrease in penalty when comparing this learned graph with the true graph. The number of free parameters increased from 94 to 143, with an increase in edges from 11 to 16, resulting in an increase in the complexity of the model graph.

With the application of undirected constraints, the BSF score and F1 score increased from 0.325 to 0.683 and 0.462 to 0.733, respectively. This shows that the accuracy and performance of the learned model are improved with a better fit. The SHD value decreased from 12.0000 to 6.000, showing that there is a decrease in penalty when comparing this learned graph with the true graph. The number of free parameters increased from 94 to 113, with an increase in edges from 11 to 15, resulting in an increase in the complexity of the model graph.

These results were expected since adding knowledge in the form of constraints gives the structure learning algorithm new information, enhancing the model's accuracy, with more F1 score and usefulness. In exchange for the model's increased accuracy, there are more free parameters and edges, but the runtime is unaffected since the knowledge-based restrictions have no impact on the algorithm's computational complexity.

References

[1] A. I. M. L. I. Mohammed Guhdar, "Optimizing Accuracy of Stroke Prediction Using Logistic Regression," *Journal of Technology and Informatics,* vol. 4, no. 2, 2023.

[2] I. B. W. A. P. A. Carmel M McEniery, "AGE, HYPERTENSION AND ARTERIAL FUNCTION," *Wiley Online Library,* vol. 34, no. 7, pp. 665 - 671, 2007.

[3] J. F. Reckelhoff, "Gender differences in hypertension," *Current Opinion in Nephrology and Hypertension,* vol. 27, no. 3, pp. 176-181, 2018.

[4] K. P. Markku Laakso, "Age of Onset and Type of Diabetes," *Diabetes Care,* vol. 8, no. 2, pp. 114 - 117, 1985.

[5] K. M. G. E. A.M. Gale, "Diabetes and gender," *Diabetologia ,* vol. 44, pp. 3-15, 2001.

[6] S. B. S. G. Shlomo Stern, "Aging and Diseases of the Heart," *Circulation,* vol. 108, pp. e99-e101, 2003.

[7] A. A. Y. Maas, "Gender differences in coronary heart disease," *Neth Heart J,* vol. 18, pp. 598-603, 2010.

[8] J. R. B. S. F. E. A. Sobal, "Marital status, fatness and obesity," *Social Science & Medicine,* vol. 35, no. 7, pp. 915-923, 1992.

[9] H. N. K. M. Y. S. M. I. T. K. Y. N. Y. S. a. K. N. Yuko Morikawa, "Effect of shift work on body mass index and metabolic parameters," *Scandinavian Journal of Work, Environment & Health,* vol. 33, no. 1, pp. 45-50, 2007.

[10] W. A. E. E. P. E. V. A. Parker MA, "Trends in Rural and Urban Cigarette Smoking Quit Ratios in the US From 2010 to 2020," *JAMA Netw Open,* vol. 5, no. 8, 2022.

[11] E. Z. I. Z. M. P. E. P. C. T. D. G. E. George Papathanasiou, "Association of High Blood Pressure with Body Mass Index, Smoking and Physical Activity in Healthy Young Adults," *The Open Cardiovascular Medicine Journal,* vol. 9, pp. 5-17, 2015.

[12] C. J. Cardiol, "Diabetes, Hypertension, and Cardiovascular Disease: Clinical Insights and Vascular Mechanisms," *The Canadian Journal of Cardiology,* vol. 34, no. 5, pp. 575-584, 2018.

[13] M. Heidi Moawad, "How Heart Disease Can Lead to Stroke," 2023. [Online]. Available: https://www.verywellhealth.com/heart-disease-that-leads-to-stroke-4083060.