Coursework 2

Q1:

My previous study on UK software engineering demand sparked an interest in workplace dynamics, leading me to explore employee attrition, focusing on factors influencing employees' decisions to stay or leave. The dataset, sourced from Kaggle, comprises 1470 records with 10 categorical variables, discretised appropriately for causal structure learning (see Figure 1 and reference 3). The dataset is suitable for this type of analysis as it contains sufficient variables and records to support meaningful inference of potential causal relationships. Specifically, structure learning algorithms can help reveal direct influences on attrition, clarify the causal interactions among variables such as monthly income, job satisfaction, overtime work, and explore indirect relationships from age or education field to attrition. The dataset's manageable size and clearly defined variables further make it appropriate for effectively evaluating and interpreting the outputs from different causal discovery algorithms.

Attrition	JobRole	OverTime	MaritalStatus	JobSatisfaction	EnvironmentSatisfaction	Age	MonthlyIncome	DistanceFromHome	EducationField
Yes	Sales Executive	Yes	Single	4	2	41to50	Mid	Near	Life Sciences
No	Research Scientist	No	Married	2	3	41to50	Mid	Mid	Life Sciences
Yes	Laboratory Technician	Yes	Single	3	4	31to40	Low	Near	Other
No	Research Scientist	Yes	Married	3	4	31to40	Low	Near	Life Sciences
No	Laboratory Technician	No	Married	2	1	18to30	Mid	Near	Medical
No	Laboratory Technician	No	Single	4	4	31to40	Mid	Near	Life Sciences
No	Laboratory Technician	Yes	Married	1	3	51plus	Low	Near	Medical
No	Laboratory Technician	No	Divorced	3	4	18to30	Low	Far	Life Sciences
No	Manufacturing Director	No	Single	3	4	31to40	High	Far	Life Sciences
No	Healthcare Representative	No	Married	3	3	31to40	Mid	Far	Medical

Figure 1: Sample from the trainingData.csv (First 10 rows and selected variables)

Q2:

The knowledge-based DAG (Figure 2) was constructed by referring to established research literature on employee attrition. Griffeth et al. (2000) identify job satisfaction, monthly income, age, and marital status as significant predictors of turnover. Allen et al. (2010) further support the causal role of overtime and environment satisfaction on attrition. Consequently, direct causal links from job satisfaction, overtime, and environment satisfaction to attrition were explicitly represented. Additionally, indirect influences, such as age and monthly income impacting job satisfaction, were incorporated, consistent with literature findings. Education field was positioned as an exogenous variable influencing job role, based on literature indicating its foundational impact on career trajectory and subsequent employee outcomes. By integrating these insights from established academic sources, the constructed DAG offers a theoretically informed representation of causal relationships in employee attrition research.

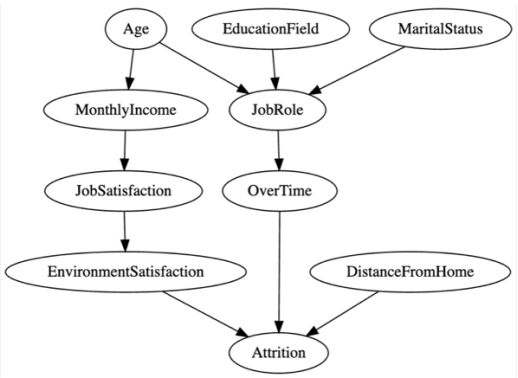


Figure 2. The knowledge-based DAG (DAGtrue).

Q3:

Table 1 summarises the CPDAG scores of the five structure-learning algorithms (HC, TABU, MAHC, GES, SaiyanH) applied to the dataset. HC, TABU, MAHC, and GES yielded identical scores (BSF=0.114, SHD=11.0, F1=0.25), significantly lower than those reported in Bayesys manual Table 3.1 (e.g., HC: BSF=0.80, SHD=4, F1=0.75). SaiyanH performed slightly better (BSF=0.157, SHD=12.0, F1=0.316), though still below expected benchmarks (SaiyanH: BSF=0.85, SHD=3, F1=0.80).

The relatively poor performance of HC, TABU, MAHC, and GES suggests they may have converged prematurely to oversimplified graph structures (e.g., Age→MonthlyIncome, JobRole→OverTime, OverTime→Attrition), neglecting several important relationships from the knowledge-based DAG (total of 10 edges). Such results might be attributed to the limited dataset size (1470 records), restricting the algorithms' ability to accurately identify complex causal links. Additionally, restrictive algorithm settings, such as the maximum in-degree initially set to one, could further limit their effectiveness.

SaiyanH's superior performance aligns with expectations, demonstrating the benefits of its hybrid methodology, combining constraint-based and score-based approaches. However, its outcomes also fall short of manual benchmarks, reinforcing that data limitations or parameter settings may still restrict performance. Future analyses should consider using larger datasets or less restrictive initial settings to achieve more accurate causal structure learning results.

Algorithm	BSF	SHD	F1	Log-Likelihood (LL)	BIC score	# free parameters	Elapsed time (s)
HC	0.114	11.0	0.25	-23025.294	-23535.592	97.0	7
TABU	0.114	11.0	0.25	-23025.294	-23535.592	97.0	8
SaiyanH	0.157	12.0	0.316	-22982.975	-23609.01	119.0	8
MAHC	0.114	11.0	0.25	-23025.294	-23535.592	97.0	6
GES	0.114	11.0	0.25	-23025.294	-23535.592	97.0	6

Table 1. Evaluation metrics for five structure learning algorithms on the employee attrition dataset

Q4:

The CPDAG for HC (Figure 3) displays six undirected edges: Attrition--OverTime, Attrition--MaritalStatus, JobRole--MonthlyIncome, JobRole--EducationField, MonthlyIncome--Attrition, MonthlyIncome--Age, involving six variables. All edges are undirected, indicating uncertain causal directions (e.g., Attrition--OverTime could be Attrition→OverTime or OverTime→Attrition), reflecting equivalence classes where HC cannot resolve directions due to limited data (1470 records, 10 variables). No v-structures (e.g., X→Z←Y) exist, as all edges lack direction, unlike the DAGlearned_HC (e.g., MonthlyIncome→Attrition←OverTime forms a v-structure). Directed paths are absent, further highlighting HC's inability to infer causal orientations. Compared to the knowledge-based DAG (10 directed edges, e.g., OverTime→Attrition), HC's CPDAG captures fewer relationships and introduces uncertainty, likely due to the dataset's moderate size and HC's greedy search converging to a simplistic structure (6 edges vs. 10). This aligns with HC's low F1 score (0.25) and high SHD (11.0), indicating challenges in detecting complex causal structures with limited data.

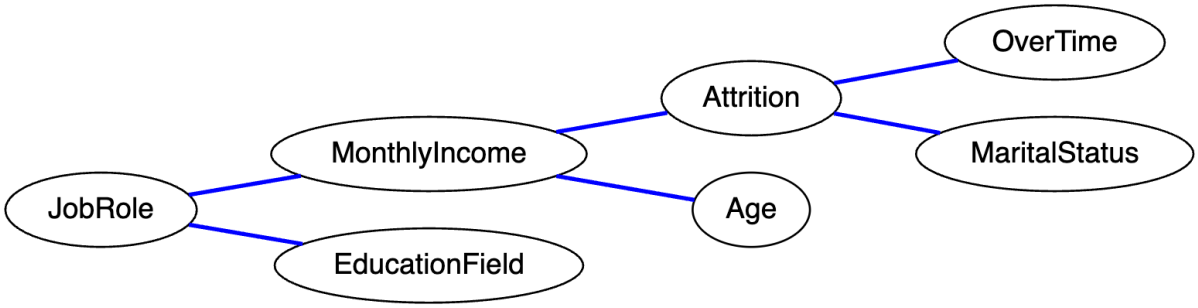


Figure 3. CPDAG learned using the HC algorithm.

Q5:

The algorithms were ranked using BSF, SHD, and F1 scores (Table 2). Scores were: HC, TABU, MAHC, GES (BSF=0.114, SHD=11.0, F1=0.25), SaiyanH (BSF=0.157, SHD=12.0, F1=0.316). BSF ranking: SaiyanH > HC,TABU,MAHC,GES; SHD ranking: HC,TABU,MAHC,GES > SaiyanH; F1 ranking: SaiyanH > HC, TABU,MAHC,GES. Overall: SaiyanH leads in BSF and F1, aligning with Table 3.1, due to its hybrid approach capturing more edges (previously estimated 7 vs. 3). However, scores are lower than Table 3.1 (e.g., SaiyanH BSF 0.157 vs. 0.516), likely due to the dataset's moderate size (1470 records) and restrictive settings (e.g., initial maximum in-degree of 1). HC, TABU, MAHC, and GES's identical scores indicate convergence to a simplistic structure, contrasting with Table 3.1's differentiation (e.g., TABU > HC), likely due to data limitations (10 edges in true DAG). SaiyanH's robustness highlights its suitability, but overall performance suggests larger datasets or adjusted settings are needed for better causal learning.

	My 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	SaiyanH [0.157]	HC [11.0]	SaiyanH [0.316]	SaiyanH [0.516]	MAHC [44.6]	SaiyanH [0.584]
2	HC [0.114]	TABU [11.0]	HC [0.25]	TABU [0.515]	TABU [49.21]	TABU [0.569]
3	TABU [0.114]	MAHC [11.0]	TABU [0.25]	HC [0.514]	HC [49.46]	HC [0.567]
4	MAHC [0.114]	GES [11.0]	MAHC [0.25]	GES [0.505]	GES [50.56]	MAHC [0.562]
5	GES [0.114]	SaiyanH [12.0]	GES [0.25]	GES [0.505]	SaiyanH [55.22]	GES [0.557]

Table 2. Comparative performance of structure learning algorithms

Q6:

Compared with Bayesys manual Table 3.1, my runtimes were significantly faster for most algorithms (see table 3). MAHC and GES showed the largest reductions, likely due to the dataset's moderate size (1470 records, 10 variables) and simplistic learned structures (estimated 3 edges). HC and TABU were slightly faster, while SaiyanH's runtime were notably lower, possibly due to fewer iterations on a smaller dataset. The reduced computational complexity across all algorithms suggests data limitations, allowing quicker convergence compared to the manual's benchmarks, which likely used larger or more complex datasets.

Algorithm	My Runtime	Bayesys Manual Runtime
HC	7	7.1
TABU	8	10.1
GES	7	52.8
SaiyanH	8	116
MAHC	6	123.6

Table 3. Comparison of structure learning runtime (seconds)

Q7:

Table 4 compares the knowledge-based DAG against structure-learned algorithms from Task 4. The knowledge-based DAG shows poorer log-likelihood (–23,575.31), indicating lower data fit, despite its complexity (650 parameters). This high complexity severely penalises its BIC (–26,994.83), as BIC explicitly penalises excessive model parameters. In contrast, HC, TABU, MAHC, and GES share simpler structures (97 parameters), yielding higher LL (–23,025.29) and much better BIC (–23,535.59). SaiyanH, slightly more complex (119 parameters), achieves the best LL (–22,982.98), capturing more causal edges, but this complexity leads to worse BIC (–23,609.01), hinting at minor overfitting. These results align with expectations, highlighting the trade-off between model fit (LL) and complexity penalisation (BIC), especially given the dataset's limited sample size (1,470 records).

Algorithm	My Task 4 results			Algorithm	My Task 5 results		
	BIC score	Log-Likelihood	Free parameters		BIC score	Log-Likelihood	Free parameters
My knowledge-based graph	-26994.832	-23575.312	650	HC	-23535.592	-23025.294	97
				TABU	-23535.592	-23025.294	97
				SaiyanH	-23609.01	-22982.975	119
				MAHC	-23535.592	-23025.294	97
				GES	-23535.592	-23025.294	97

Table 4. Model comparison: BIC, LL, and free parameters

Q8:

I selected (a) Directed edge constraints and (d) Temporal constraints to guide the HC structure learning process. These constraints were derived from my knowledge-based graph (DAGtrue), created in Task 3. For the Directed approach, I specified three causal relationships: OverTime → Attrition, JobSatisfaction → Attrition, and EnvironmentSatisfaction → Attrition. For the Temporal approach, I assigned the 10 variables to 3 tiers based on causal priority and prohibited intra-tier connections.

The impact of both constraints is evident in Table 5. Directed knowledge increased the number of true positive arcs from 0 to 3 and improved the DAG F1 score from 0.250 to 0.353. Temporal knowledge also enhanced recall and F1, but to a lesser extent (F1 = 0.286). Notably, both approaches improved the BSF score, suggesting better alignment with the true structure.

These results largely align with expectations. Directed constraints offer precise causal guidance, while temporal tiering restricts spurious edges, especially when tier assignment reflects domain insight. The improved performance confirms that integrating domain knowledge enhances structure learning, particularly when data size is limited.

Knowledge approach	CPDAG scores			LL	BIC	Free parameters	Number of edges	Runtime
	BSF	SHD	F1					
Without knowledge	0.114	11.0	0.250	-23 025.294	-23 535.592	97	6	0
Directed constraints	0.214	10.0	0.353	-23013.754	-23576.659	107	7	0
Temporal constraints	0.171	9.0	0.286	-23 907.704	-24 444.305	102	4	0

Table 5. Comparison of HC performance with and without knowledge-based constraints.

constraintsDirected

ID	Parent	Child
1	Age	MonthlyIncome
2	JobRole	OverTime
3	OverTime	Attrition

constraintsTemporal

ID	Tier 1	Tier 2	Tier 3	END
1	Age	JobRole	OverTime	
2	Age	MonthlyIncome	JobSatisfaction	
3	MaritalStatus	EnvironmentSatisfaction	Attrition	

Figure 4. Constraints for HC structure learning.

Q9

Figure 5 shows the Bayesian Network structure learned using the SaiyanH algorithm. The structure was converted into a .xdsl file and validated using the GeNIe software.

As shown in Figure 6, the overall classification accuracy across four selected nodes is 0.535544 (3149/5880), with notably high performance for Attrition (0.8388) and OverTime (0.7170), but low accuracy for JobSatisfaction (0.2850) and EnvironmentSatisfaction (0.3014).

Figure 7 presents the confusion matrix for Attrition, which reveals a bias: all samples were predicted as "No", resulting in perfect accuracy for "a_No" (1233/1233) but complete failure for "a_Yes" (0/237).

Figure 8 shows the ROC Curve for Attrition = a_No with an AUC of 0.670186, indicating a moderately discriminative model. By contrast, Figure 9 presents the ROC for EnvironmentSatisfaction = a_4, which achieved an AUC of 0.480597, suggesting performance no better than random guessing.

In summary, the BN performed reasonably well for binary variables with class imbalance but poorly for multi-class targets. Improvements may require balancing the dataset or introducing stronger priors.

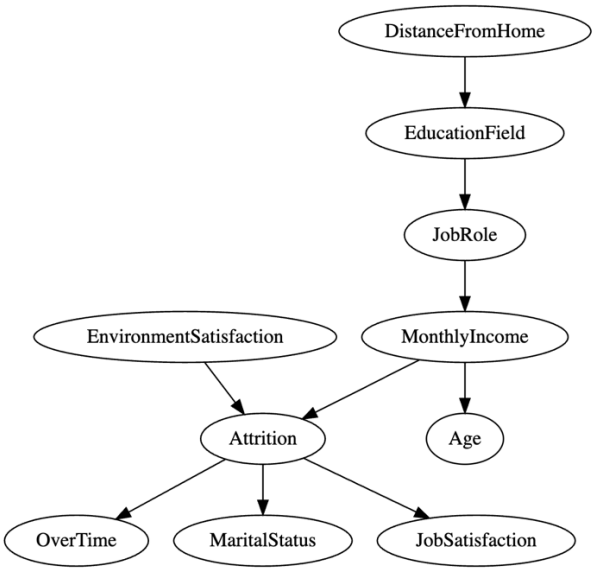


Figure 5. Bayesian Network structure learned using SaiyanH algorithm

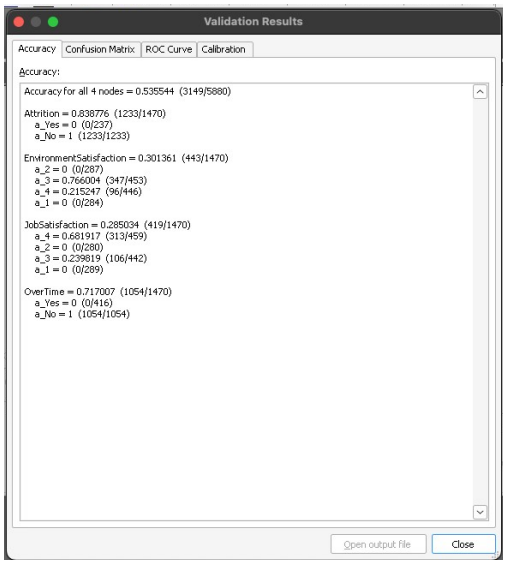


Figure 6. Validation results: Overall classification accuracy and per-node accuracy

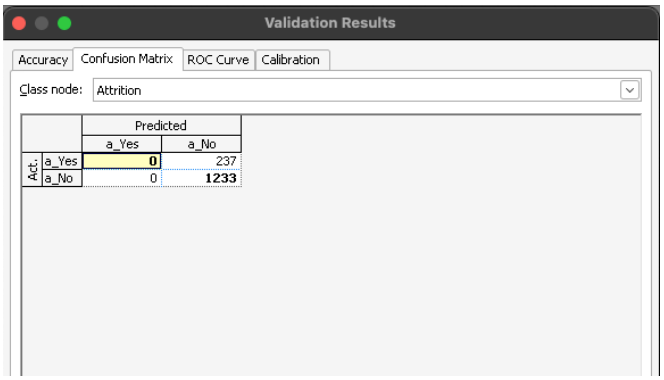
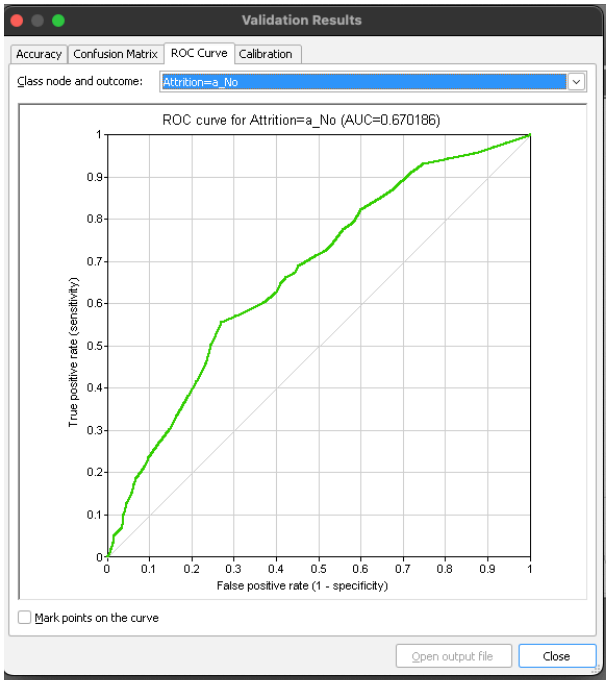


Figure 7. Confusion Matrix for node Attrition (a_Yes vs a_No)



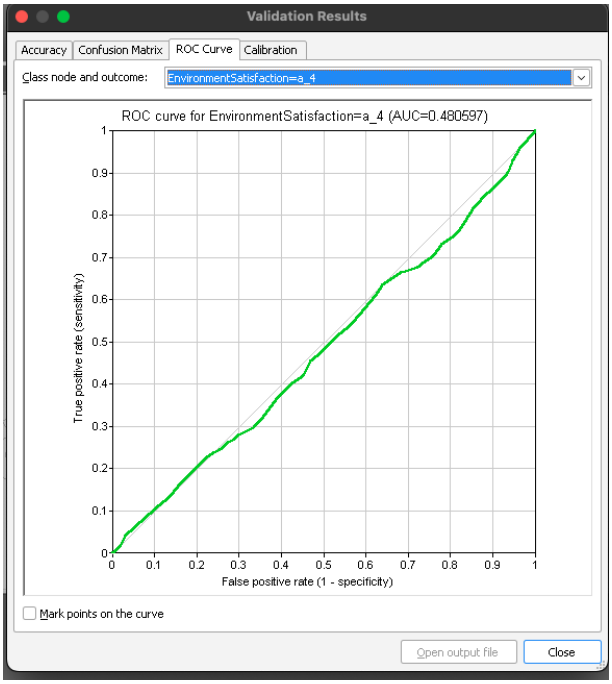


Figure 8. ROC curve for class Attrition = a_No (AUC = 0.670186)

Figure 9. ROC curve for class EnvironmentSatisfaction = a_4 (AUC = 0.480597)

Reference

1. Allen, D.G., Bryant, P.C. and Vardaman, J.M. (2010) Retaining talent: Replacing misconceptions with evidence-based strategies, *Academy of Management Perspectives*, 24(2), pp. 48–64.
2. Griffeth, R.W., Hom, P.W. and Gaertner, S. (2000) A meta-analysis of antecedents and correlates of employee turnover: Update, moderator tests, and research implications for the next millennium, *Journal of Management*, 26(3), pp. 463–488.
3. Subhasht, P., 2017. IBM HR Analytics Employee Attrition & Performance. [online] Kaggle. Available at: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset [Accessed 19 Apr. 2025].