Data Analytics
ECS 784P

Coursework 2

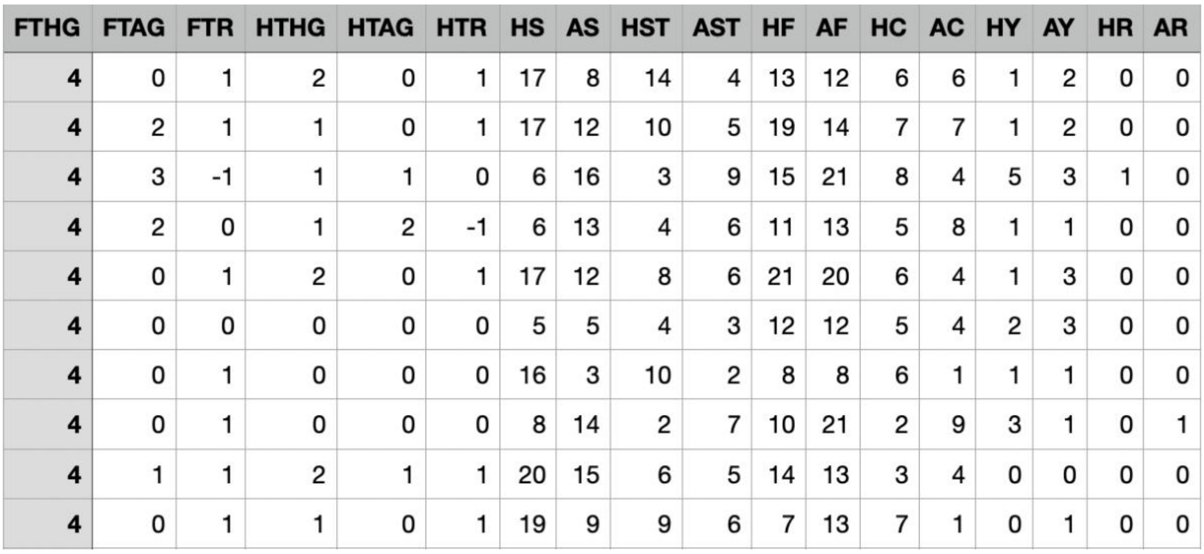
- Semil Halani
(210147012)

Glossary:

FTHG	Full Time Home Team Goals
FTAG	Full Time Away Team Goals
FTR	Full Time Result (1=Home Win, 0=Draw, -1=Away Win)
HTHG	Half Time Home Team Goals
HTAG	Half Time Away Team Goals
HTR	Half Time Result (1=Home Win, 0=Draw, -1=Away Win)
HS	Home Team Shots
AS	Away Team Shots
HST	Home Team Shots on Target
AST	Away Team Shots on Target
HF	Home Team Fouls Committed
AF	Away Team Fouls Committed
HC	Home Team Corners
AC	Home Team Corners
HY	Home Team Yellow Cards
AY	Away Team Yellow Cards
HR	Home Team Red Cards
AR	Away Team Red Cards

Question 1:

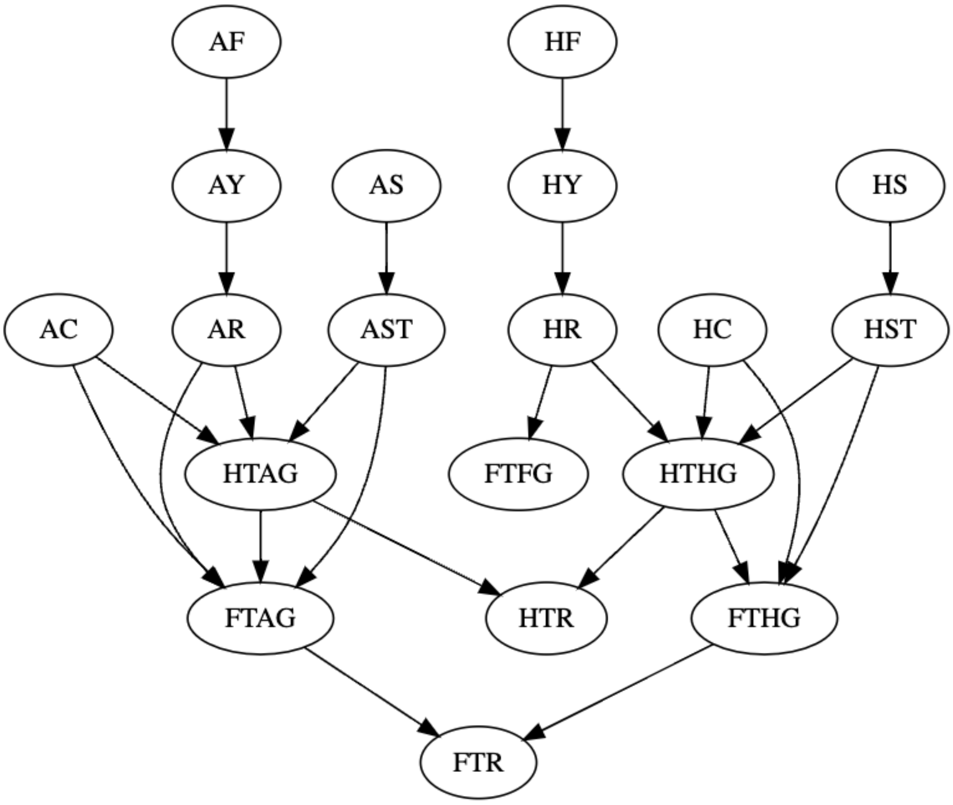
This selected research area discusses the ability of a football team in the England Premier League (EPL), to win in their home stadium and in the way which the home and away teams' basic stats are affected. In the EPL, the team in their home stadium, has a lot of home fans in the stands cheering the home team and booing the away team, which builds a vibe much more favourable for the home team as their confidence strengthens with the cheers whereas it might weaken for the away team. This further helps in getting more shots on target leading to more goals for the home team. These home ground advantages have historically proved to highly influence the results for these games.



FTHG	FTAG	FTR	HTHG	HTAG	HTR	HS	AS	HST	AST	HF	AF	HC	AC	HY	AY	HR	AR
4	0	1	2	0	1	17	8	14	4	13	12	6	6	1	2	0	0
4	2	1	1	0	1	17	12	10	5	19	14	7	7	1	2	0	0
4	3	-1	1	1	0	6	16	3	9	15	21	8	4	5	3	1	0
4	2	0	1	2	-1	6	13	4	6	11	13	5	8	1	1	0	0
4	0	1	2	0	1	17	12	8	6	21	20	6	4	1	3	0	0
4	0	0	0	0	0	5	5	4	3	12	12	5	4	2	3	0	0
4	0	1	0	0	0	16	3	10	2	8	8	6	1	1	1	0	0
4	0	1	0	0	0	8	14	2	7	10	21	2	9	3	1	0	1
4	1	1	2	1	1	20	15	6	5	14	13	3	4	0	0	0	0
4	0	1	1	0	1	19	9	9	6	7	13	7	1	0	1	0	0

The chosen dataset from 2000-18 seasons is suitable for structure learning because most of the variables are causally related. The structure learning algorithms reflect the correlation between home team statistics and their winning rate in their home grounds.

Question 2:



Usually, the entire season of EPL is played with the objective of winning with as many goals as possible with the home ground advantage so when they play an away game (game at the opposition's home ground), they have an upper hand[1]. History shows that teams have a higher winning rate at home when compared to away games. Knowledge of and practice at the ground results in an increased field goal percentage, corners, and penalties which in turn results in a win with a better score for the home team. At the home ground, it is easier for home fans to occupy more seats in the stadium and thus, cheer their team in a better way, which usually boosts the home team's confidence drastically, helping to convert shots to shots on target and shots on target, corners, and penalties into goals. Moreover, statistics such as attacks, shots and shots on targets are correlated between the two teams as more attacking play by the confident team results the other team to play defensive. On the basis of sufficient knowledge, the correlations between a few nodes can be easily discerned.

[1] "Home Advantage - England Premier League". *Footystats*, https://footystats.org/england/premier-league/home-advantage-table.

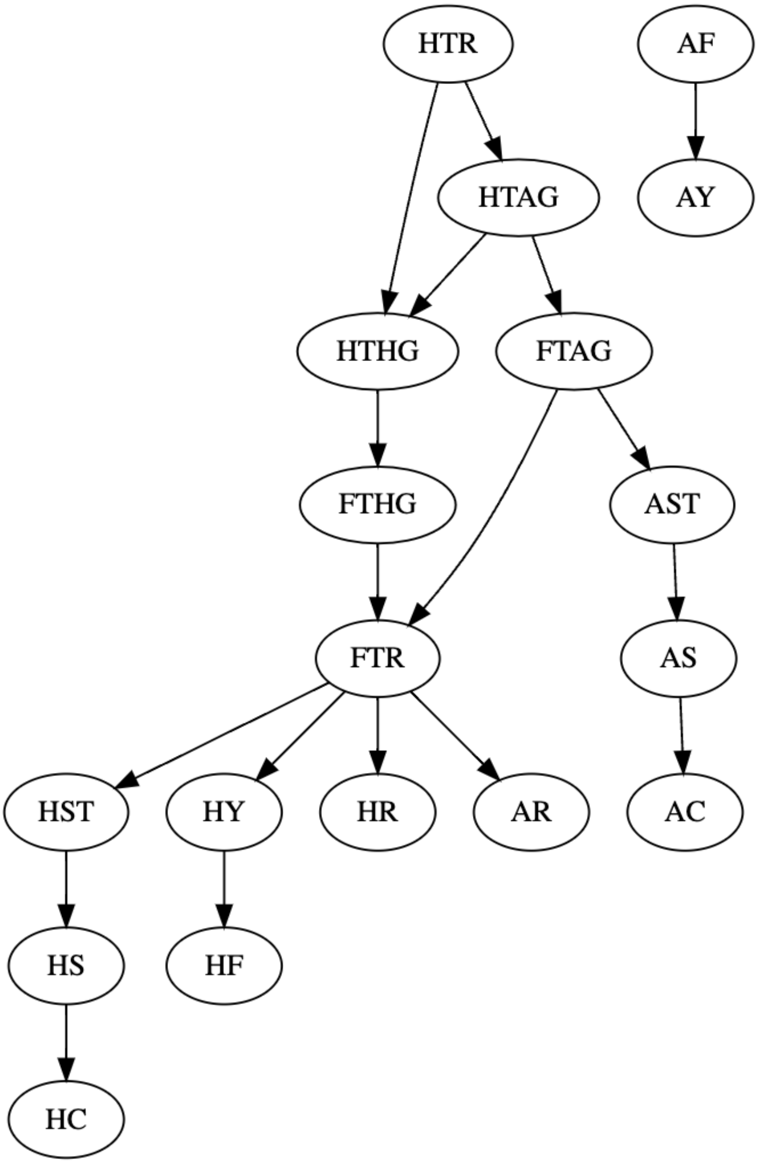
Question 3:

Algorithm	CPDAG scores			Log-Likelihood (LL) score	BIC score	# free parameters	Structure Learning elapsed time
	BSF	SHD	F1				
HC_CPDAG	0.081	27.5	0.199	-5084.177	-5502.077	106	0 seconds
HC_DAG	0.081	27.5	0.199	-5084.177	-5502.077	106	0 seconds
TABU_CPDAG	0.081	27.5	0.199	-5084.177	-5502.077	106	0 seconds
TABU_DAG	0.081	27.5	0.199	-5084.177	-5502.077	106	0 seconds
SaiyanH	0.055	32.5	0.138	-4831.907	-5600.047	194	0 seconds
MAHC	0.081	27.5	0.199	-5084.177	-5502.077	106	0 seconds

Table 1.2

As shown in the above table, these are the results obtained on running the required six algorithms. On comparison of the above values with those given in table 2.1 of the Bayesys manual, it is seen that the average values available are more than the values generated for all the algorithms. The produced values are as per expectations due to the difference between the visible and generated graphs. For all the algorithms, the BSF scores are quite close to 0, which indicates that the graph is fully connected. The SHD scores generated here are a bit lower than average, which shows that it is much more easier and simpler for the learned graph to generate the true graph. F1 scores indicate the precision for the learned graphs. The observation of the poor F1 scores obtained, shows that most of the linked nodes are different than that of the DAGtrue file. Moreover, the SHD score gives the number of steps that would be required from the learned model to reach the true graph. Thus, the lower the SHD score, the lower the number of recalls required to reach the true graph from the learned model.

Question 4:



The three causal classes in the CPDAG are highlighted below:

Causal chain (X → Y → Z):

FTHG → FTR → HST

The causal chain illustrated above indicates that a home team’s goals in the full time game directly influences the full time result which further shows that it is influencing home teams shots on target.

Common cause (X ⟵ Y → Z):

FTR ⟵ FTAG → AST

The common cause illustrated above shows that the number of goals by the home team influences the full time result and the away team’s shots on target as they would be losing confidence when the home team keeps scoring.

Common effect (X → Y ⟵ Z):

There is no common effect form in given knowledge.

This is because, there is no feasible situation where one factor is a result of 2 other factors simultaneously.

Answer 5:

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 [avera
1	HC_CPDAG [0.071]	SaiyanH [32.500]	HC_CPDAG [0.199]	TABU_CPDAG [0.533]	MAHC [0.481]	Saiyan
2	HC_DAG [0.071]	HC_CPDAG [27.500]	HC_DAG [0.199]	SaiyanH [0.515]	TABU_CPDAG [0.44]	TABU [0.564
3	TABU_CPDAG [0.071]	HC_DAG [27.500]	TABU_CPDAG [0.199]	HC_CPDAG [0.506]	SaiyanH [0.438]	MAHC
4	TABU_DAG [0.071]	TABU_CPDAG [27.500]	TABU_DAG [0.199]	MAHC [0.499]	HC_CPDAG [0.402]	HC_C [0.537
5	MAHC [0.071]	TABU_DAG [27.500]	SaiyanH [0.138]	TABU_DAG [0.484]	TABU_DAG [0.397]	TABU [0.53]
6	SaiyanH [0.055]	MAHC [27.500]	SaiyanH [0.138]	HC_DAG [0.438]	HC_DAG [0.314]	HC_D [0.479

Table 1.3

On running the SaiyanH algorithm, the values generated are distinct and they seem to have a range of values which were unexpected. The BSF value for SaiyanH in the Bayesys manual is ranked higher while the generated BSF value lies at the lowest rank. In the manual, the SHD value for SaiyanH indicates a mid-level value of 0.438 while on the other hand, the value generated which is much higher which was unexpected. Finally, the F1 score visible in the Bayesys manual is the highest average value for the SaiyanH algorithm. Nevertheless, the F1 value which was generated in the lowest value visible. Overall, the ranks displayed above are unexpected due to the position of the BSF, SHD and F1 values for the SaiyanH algorithm.

Question 6:

The chosen dataset post pre-processing consists of 6461 rows and 18 nodes. According to the Bayesys manual for this size of a dataset, we can say that the average structure learning run time should be 0 seconds. Table 2.1 in the manual shows that the datasets of the approximate size of 10^4 should have an avg. runtime of 0 seconds. On running the HC_DAG, HC_CPDAG, TABU_DAG, TABU_CPDAG, MAHC and SaiyanH algorithms, it is seen that the run time for all the algorithms is 0 seconds. Thus, the results obtained are consistent with the ones of the table 2.1 of the Bayesys manual.

Question 7:

Algorithm	Your Task 4 results			Algorithm	Your Task 5 results		
	BIC score	Log-likelihood	Free parameters		BIC score	Log-likelihood	Free parameters
Your knowledge-based graph	-1461703.926	-44518.545	251241	HC_CPDAG	-5502.077	-5084.177	106
				HC_DAG	-5502.077	-5084.177	106
				TABU_CPDAG	-5502.077	-5084.177	106
				TABU_DAG	-5502.077	-5084.177	106
				SaiyanH	-4831.907	-5600.047	194
				MAHC	-5502.077	-5084.177	106

Table 1.4

Following the execution of task 4, the results which are obtained are very different to those generated after the task 5 as it is seen above. As observed, the Log Likelihood and the BIC score both decrease drastically after the implementation of the six algorithms present under task 5. The Log Likelihood and BIC score are similar for all the algorithms executed under task 5. It is visible that when compared to the original after the execution of the required algorithms, the number of free parameters has also decreased significantly. The complexity of the structural model is usually indicated by the BIC score. A higher BIC score is seen in Task 4 when compared to Task 5, which is unexpected as the complexity of the structural model should increases after the execution of the algorithms. On the basis of the selected dataset, the MAHC algorithm produces the lowest BIC score and thus, appears to be the most efficient. The Log Likelihood illustrates how good the learned model fits the true model. Based on the table above, it is visible that the Log Likelihood highly reduces after running the algorithms. This is not expected as the learned graph usually fits the true model better after running the required algorithms. Nonetheless, the MAHC algorithm finished with the lowest number of free parameters.

Question 8:

	CPDAG scores							
Knowledge Approach	BSF	SHD	F1	LL	BIC	Free parameters	Number of edges	Runtime
Without knowledge	0.081	27.500	0.199	-5084.177	-5502.077	106	18	0 seconds
With knowledge (Directed)	0.207	25.500	0.380	-43522.298	-47369.532	470	19	0 seconds
With knowledge (Forbidden)	0.227	24.500	0.388	-44037.245	-47163.348	480	18	0 seconds

Table 1.5

Following the analysis of the chosen dataset for football games, the two knowledge-based approaches selected to ahead with were Directed and Forbidden. The directed knowledge approach was selected as statistics such as shots, shots on targets and the half time home team goals have a direct and significantly critical effect on the ability of a team to win a home game. On the other hand, the forbidden knowledge approach was chosen because statistics such as an away team’s shots and yellow cards have no relation to the home team’s goals. Furthermore, more instances were also considered when running the structure learning model with the knowledge-based approach. The SHD score shows the number of steps required to reach the true graph from the learned model. It is seen that the SHD score reduces after implementing the Directed approach and it further keeps decreasing on the implementation of the forbidden knowledge approach. Through these values we can deduce that the selected approaches denote a lesser number of recalls required to reach the true graph. The BIC score illustrates the overall complexity of the model as BIC score is directly proportional to the complexity of the model; the lower the BIC score, the less complex is the model. It can be seen in the table 1.5 above that the BIC score slightly increases on the implementation of the knowledge-based approaches, which is expected, as the addition of these approaches to the learning model further increases its complexity. Moreover, F1 denotes the correctness of the table, the closer is the value to 1, the better it is. In table 1.5, we can note that values of F1 get nearer to 1 after the implementation of the knowledge-based approaches, which is also expected as the application of these approaches increases the correctness of the learning model further.