Instructor: ***Dr. H.T. Luong***

FINAL EXAMINATION
Course: AT72.9014 – Applied Data Analytics
(Time Duration: 3 hours)

Q1. (20 points)

A researcher is interested in finding the effects of Age and Weight on Systolic Blood Pressure. The available data are as follows:

Patient	Systolic Blood Pressure	Age (years)	Weight (pounds)
1	132	52	173
2	143	59	184
3	153	67	194
4	162	73	211
5	154	64	196
6	168	74	220
7	137	54	188
8	149	61	188
9	159	65	207
10	128	46	167
11	166	72	217

(*Data in this table is provided in the attached file Data.xlsx*)

Let help the researcher to select the best multiple regression model among

$$Systolic\ Blood\ Pressure = \beta_0 + \beta_1 * Age + \beta_2 * Weight + \varepsilon$$
$$Systolic\ Bood\ Pressure = \beta_0 + \beta_1 * Age + \beta_2 * Weight + \beta_3 * Age^2 + \varepsilon$$
$$Systolic\ Blood\ Pressure = \beta_0 + \beta_1 * Age + \beta_2 * Weight + \beta_3 * Weight^2 + \varepsilon$$
$$Systolic\ Blood\ Pressure = \beta_0 + \beta_1 * Age + \beta_2 * Weight + \beta_3 * Age * Weight + \varepsilon$$

For each model, show the detailed results (no need to show the graphs even though you should analyze the graphs) and explain why the model you select is the best model.

Q2. (20 points)

PART A Information:

Among 2000 persons who took Covid test for an intended international flight in a recent survey, there are 3 groups:

- Group A: 1000 persons who took only RT-PCR test.
- Group B: 700 persons who took only ART test.
- Group C: 300 persons who took both tests.

PART B information:

Among the persons in group A, 5% got positive result.
Among the persons in group B, 10% got positive result.
Among the persons in group C, 7% got positive result in at least one test.
.

Questions:

1. A new person is planning for an international trip by airplane. What will be the anticipated group classification of that person?
2. A new person is planning for an international trip by airplane. What will be the anticipated test result classification of that person?
3. Assuming that the test result of the person is positive, which group does the person most likely belong to?

Q3. (20 points)

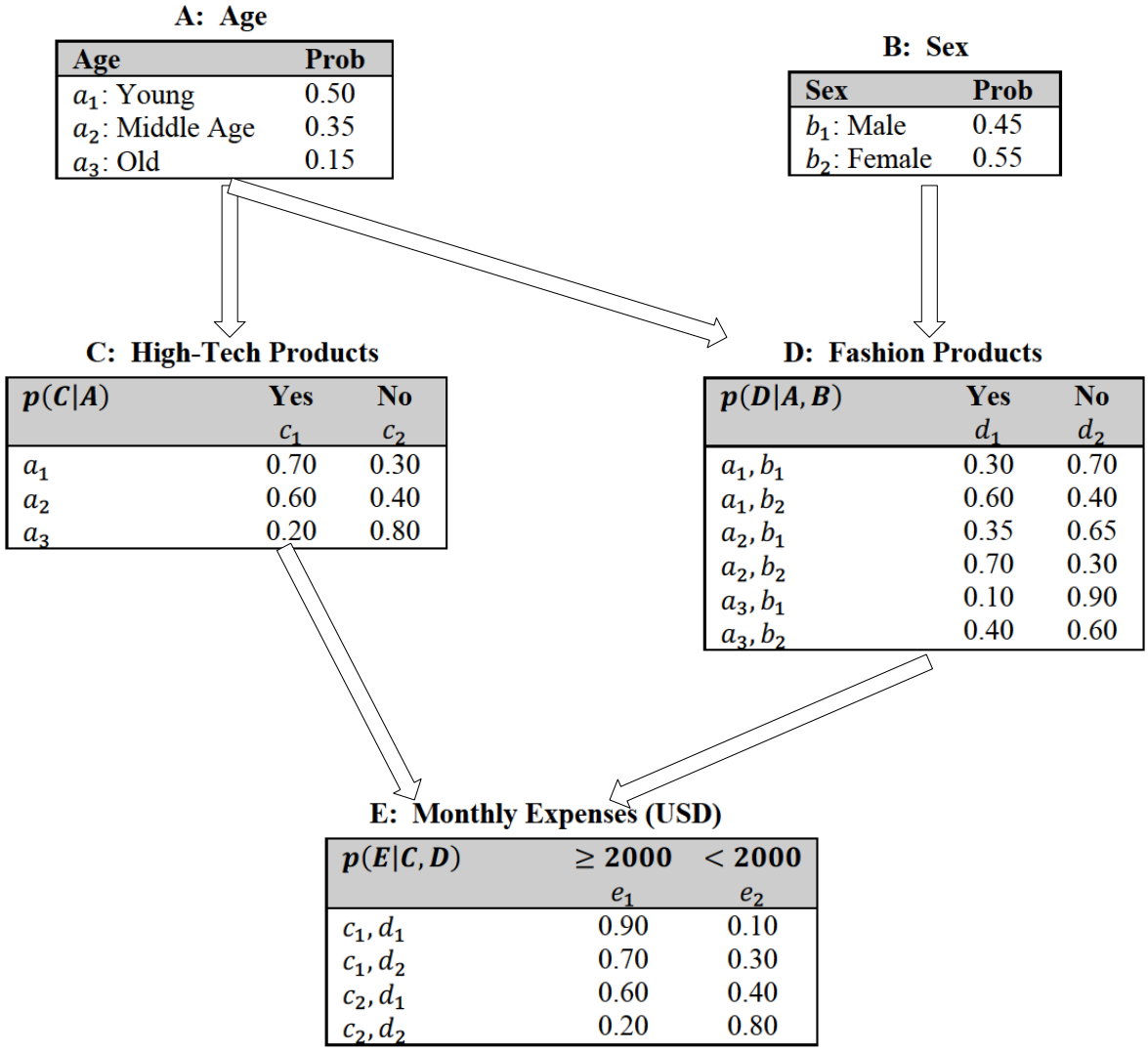
A customer survey has been conducted to examine the effects on monthly expense of customer (dependent variable E) from various factors including

Factor A:	Age of customer
Factor B:	Sex of customer
Factor C:	Preference in buying high-tech products
Factor D:	Preference in buying fashion products

The survey results have been summarized in a Bayesian Belief Network (see below figure).

Let determine

1. The probability that a customer prefers to buy high-tech products and spends less than 2000USD per month.
2. Given that a customer prefers to buy high-tech products, what is the probability that the customer is an old female?



Q4. (20 points)

Consider the following data set

a	b	c	d	e	f	g	h
(1,3)	(3,3)	(4,3)	(5,3)	(1,2)	(4,2)	(1,1)	(2,1)

Let apply the k-means algorithm to divide the data set into two clusters with the initial cluster centers selected to be $m_1 = (3,3)$ - Point b, and $m_2 = (2,1)$ - Point h

Q5. (20 points)

Records on transaction at a retail shop related to seven types of product (i.e., A,B,C,D,E,F,G) are given in the following table

Transaction	A	B	C	D	E	F	G
1	0	1	0	0	1	0	1
2	0	1	1	1	0	0	0
3	1	1	1	0	0	1	0
4	1	1	0	0	0	1	1
5	1	0	0	1	1	0	0
6	1	0	1	1	0	1	0
7	0	1	0	0	0	1	0
8	0	0	0	0	1	1	1
9	1	0	1	1	0	0	0
10	1	1	0	0	0	0	0
11	1	1	1	0	1	0	1
12	1	0	1	1	0	0	0
13	1	1	1	1	0	0	0
14	1	1	0	1	1	1	1
Total	**10**	**9**	**7**	**7**	**5**	**6**	**5**

Let derive the frequent itemsets that satisfy the requirement: itemset frequency ≥ 5. Then derive all association rules and their support & confidence.

If the minimum required support level is 40% and the minimum required confidence level is 80%, which rules are remained?