UECM3993 PREDICTIVE MODELLING May 2024 Marking Guide

PART A: Answer **ALL** questions.

Q1. (a) State the phases involved in the CRISP-DM (Cross Industry Standard Process for Data Mining). (3 marks)

Ans. The phases in CRISP-DM are:

- Business understanding ..... [0.5 mark]
- Data understanding ..... [0.5 mark]
- Data preparation ..... [0.5 mark]
- Modelling ..... [0.5 mark]
- Evaluation ..... [0.5 mark]
- Deployment ..... [0.5 mark]

(b) Given the confusion matrix of a trained predictive model in Table 1.1 with 0 as positive.

Table 1.1: Confusion matrix on the training data. 0 is positive.

	Actual	
Prediction	0	1
0	579	127
1	71	143

(i) Find the balanced accuracy, i.e. the average of the recalls. (2 marks)

Ans.

$$\text{Balanced accuracy} = \frac{\frac{579}{650} + \frac{143}{270}}{2} = 0.7102 \quad \text{[2 marks]}$$

(ii) Find the accuracy and kappa statistic

$$\text{Kappa} = \frac{\text{Accuracy} - \text{RandomAccuracy}}{1 - \text{RandomAccuracy}}$$

where $\text{RandomAccuracy} = \frac{(\text{TP+FN})\times(\text{TP+FP})+(\text{TN+FP})\times(\text{TN+FN})}{(\text{Total Number of Data})^2}$. (5 marks)

Ans. $\text{Accuracy} = \frac{579+143}{579+127+71+143} = \frac{722}{920} = 0.7848$ ..... [2 marks]

$$\text{RandomAccuracy} = \frac{(579+127)(579+71)+(143+71)(143+127)}{(579+127+71+143)^2} = 0.6104 \quad \text{[1.5 marks]}$$

$$\text{Kappa statistic} = \frac{0.7848-0.6104}{1-0.6104} = 0.4476 \quad \text{.....[1.5 marks]}$$

(iii) Use proper examples to discuss whether the accuracy is a good performance metric for an imbalanced data. (3 marks)

Ans. The accuracy is a **not** a performance metric for imbalanced data because it is not able to it is not able to identify the bad predictive model which identifies the majority correctly which the minority very poorly as illustrated below.

..... [1 mark]

Model A	Actual	
Prediction	0	1
0	900	100
1	0	0

Model B	Actual	
Prediction	0	1
0	800	0
1	100	100

Model A and Model B both give an accuracy of 0.9. However, Model A cannot predict 1 at all while Model B can predict 1 very well. ..... [2 marks]

(c) When the linear regression model is applied to study the relation between average hourly earnings (y_i) against the years of education (`educ`), the predictions $\widehat{y}_i$ in Table 1.2 are obtained.

Table 1.2: Comparison between results from a regression model to actual data

educ	y_i	$\widehat{y}_i$
14	8.8	7.2
10	5.1	4.7
16	8.3	8.4
18	10.0	9.6
6	2.9	2.2
12	2.9	5.9

(i) Calculate the sum of squared errors (SSE). (3 marks)

Ans. $SSE = (8.8-7.2)^2+(5.1-4.7)^2+\ldots+(2.9-5.9)^2 = 12.38$. [3 marks]

(ii) the coefficient of determination, R^2. (3 marks)

Ans. $SST = (8.8-6.3333)^2+(5.1-6.3333)^2+\ldots+(2.9-6.3333)^2$

$= 48.49333$ [1.5 mark]

$$R^2 = 1-\frac{SSE}{SST} = 1-\frac{12.38}{48.49333} = 0.7447$$ [1.5 mark]

(d) Write down **four** major categories of **unsupervised learning methods** and provide a concrete application for each category. (6 marks)

Ans. The four major categories of unsupervised learning methods are

(i) Dimensionality reduction . . . . . [0.5 mark]

(ii) Cluster analysis . . . . . [0.5 mark]

(iii) Finding association rules . . . . . [0.5 mark]

(iv) Anomaly detection / Visualisation / feature extraction, etc. . . . . [0.5 mark]

Biologists employ dimensionality reduction to determine the relations between various living species based on the (complete or fragment of) genetic information. . . . [1 mark]

Clustering is the process of grouping the given data into different clusters or groups. E-commerce websites like Amazon use clustering algorithms to implement the user-specific recommendation system. . . . . [1 mark]

(i) Market segmentation divides the consumers of the market into some groups. In a group, consumers will be similar to each other based on some predefined set of characteristics. If two customers are not similar based on these characteristics, they are in different groups. Companies use this clustered data and the features of the customers to decide their market strategies, like which group of customer they should target or which group of customers needs more advertising etc. etc.

(ii) There are millions of people in social networking websites and analysing their behaviours sounds really fun. Clustering plays the role here. This idea of social networks analysis can be extended to real life social scenarios.

(iii) Search engines and many other websites use clustering to group similar web pages, videos, songs etc. and improve results for their users.

Finding Association Rules is the process of finding associations between different parameters in the available data. It discovers the probability of the co-occurrence of items in a collection, such as people that buy X also tend to buy Y. It is used in supermarket item placement (association rules) and logistics. . . . . [1 mark]

Anomaly detection: The identification of rare items, events or observations which brings suspicions by differing significantly from the normal data. . . . . [1 mark]

[Total : 25 marks]

Q2. When a bank receives a loan application, the bank has to make a decision whether to go ahead with the loan approval or not based on the applicant's profile. Two types of risks are associated with the bank's decision:

- If the applicant is a good credit risk, i.e. is likely to repay the loan, then not approving the loan to the person results in a loss of business to the bank;
- If the applicant is a bad credit risk, i.e. is not likely to repay the loan, then approving the loan to the person results in a financial loss to the bank.

To minimise loss from the bank's perspective, the bank needs a predictive model regarding who to give approval of the loan and who not to based on an applicant's demographic and socio-economic profiles.

Suppose the response variable `Y` is 0 when the loan is approved and is 1 when the loan is not approved. Suppose the features of the data are listed below:

- X_1 (categorical): Status of existing checking account (A11, A12, A13, A14);
- X_2 (integer): Duration in months
- X_3 (integer): Credit amount
- X_4 (integer): Instalment rate in percentage of disposable income
- X_5 (binary): foreign worker (yes, no)

UECM3993 PREDICTIVE MODELLING May 2024 Marking Guide

(a) When the data is trained with a logistic regression model, the statistical estimates below are obtained:

```
Call:
glm(formula = Y ~ ., family = binomial, data = d.f.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8613  -0.7239  -0.5115   0.9647   2.0657

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.561e-01  9.771e-01   0.364 0.715529
X1A12       -6.719e-01  6.889e-01  -0.975 0.329365
X1A13       -1.792e+01  2.795e+03  -0.006 0.994884
X1A14       -2.275e+00  6.754e-01  -3.369 0.000755 ***
X2           3.834e-02  3.039e-02   1.262 0.207052
X3          -1.965e-05  1.406e-04  -0.140 0.888871
X4          -1.336e-01  2.634e-01  -0.507 0.612044
X5no        -1.733e+01  1.935e+03  -0.009 0.992854
---
Signif.:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(i) Write down the mathematical expression of the logistic regression model in the conditional probability form. (4 marks)

Ans. The mathematical expression of the logistic regression model is

$$P(Y=1|\mathbf{x}) = \frac{1}{1+\exp(-\beta\cdot\mathbf{x})} \qquad \text{[2 marks]}$$

where

$$\beta\cdot\mathbf{x} = 0.3561 - 0.6719x_1^{A12} - 17.92x_1^{A13} - 2.275x_1^{A14} + 0.03834x_2 - 1.965\times10^{-5}x_3 - 0.1336x_4 - 17.33x_5 \qquad \text{[2 marks]}$$

(ii) Calculate the conditional probability of $Y=1$ and the conditional probability of $Y=0$ for a foreign worker when the status of existing checking account of the customer is A11, the duration is 6 months, the credit amount is 1169 and the instalment rate of disposable income is 4%. (6 marks)

Ans. We tabulate the information for calculation:

	X_1	X_2	X_3	X_4	X_5	$\beta\cdot\mathbf{x}$
	A11	6	1169	4	yes	
	0	3.834×10^{-2}	-1.965×10^{-5}	-0.1336	0	
0.3561	0	0.23004	-0.02297085	-0.5344	0	0.02876915

. . . . . [5 marks]

Therefore,

$$P(Y=1|X) = \frac{1}{1+\exp(-(0.02876915))} = 0.5071918 \qquad \text{[0.5 mark]}$$

$$P(Y=0|X) = 1 - 0.5071918 = 0.4928082 \qquad \text{[0.5 mark]}$$

(b) When the data is trained with a naive Bayes model with Laplace smoothing, the statistical estimates below are obtained:

```
 A priori probabilities:

    0     1
0.625 0.375

 Tables:

X1              0           1
  A11 0.18518519 0.41176471
  A12 0.18518519 0.35294118
  A13 0.05555556 0.02941176
  A14 0.57407407 0.20588235

X2            0        1
  mean 18.86000 25.30000
  sd   11.29206 15.33117

X3            0        1
  mean 2940.040 3490.167
  sd   2254.614 3213.598

X4            0        1
  mean 3.060000 3.033333
  sd   1.095631 1.098065

X5             0           1
  yes 0.92307692 0.96875000
  no  0.07692308 0.03125000
```

State the naive Bayes model for this problem using conditional probabilities and estimate the posterior probabilities for $Y = 0$ and $Y = 1$ for a foreign worker when the status of existing checking account of the customer is A11, the duration is 6 months, the credit amount is 1169 and the instalment rate of disposable income is 4%. (8 marks)

Ans. The naive Bayes model for the problem with $Y = j$, where $j = 0, 1$ is ... [1 mark]

$$P(Y = j|X_1, X_2, X_3, X_4, X_5) \propto P(Y = j)P(X_1|Y = j)P(X_2|Y = j)P(X_3|Y = j)\times P(X_4|Y = j)P(X_5|Y = j).$$

From this model, we can build a table for the computation:

j	$P(Y = j)$	$X_1 = A11\|Y = j$	$X_2 = 6\|Y = j$	$X_3 = 1169\|Y = j$	$X_4 = 4\|Y = j$	$X_5 = yes\|Y = j$
0	0.625	0.18518519	0.0184714	12.9972×10^{-5}	0.2520039	0.92307692
1	0.375	0.41176471	0.0117817	9.5638×10^{-5}	0.2466008	0.96875000

..... [5 marks]

$$P(X_2 = 6|Y = 0) = \frac{1}{\sqrt{2\pi}(11.29206)} \exp(-\frac{1}{2}(\frac{6 - 18.86}{11.29206})^2) = 0.0184714, \quad \ldots$$

The products are

$$P(Y = 0|X) \propto 6.463698 \times 10^{-8}, \quad P(Y = 1|X) \propto 4.156474 \times 10^{-8}. \quad \text{[1 mark]}$$

and the posterior probabilities are

$$P(Y = 0|X) = 0.6086246, \quad P(Y = 1|X) = 0.3913754 \quad \text{[1 mark]}$$

(c) When the data is trained with a CART model the text representation of the CART is obtained:

```
node), split, n, deviance, yval, (yprob)
      * denotes terminal node

 1) root 80 105.900 0 ( 0.6250 0.3750 )
   2) X1: A13,A14 38  33.150 0 ( 0.8421 0.1579 )
     4) X4 < 2.5 12   0.000 0 ( 1.0000 0.0000 ) *
     5) X4 > 2.5 26  28.090 0 ( 0.7692 0.2308 )
      10) X2 < 30 20  16.910 0 ( 0.8500 0.1500 )
        20) X3 < 1550.5 10  12.220 0 ( 0.7000 0.3000 ) *
        21) X3 > 1550.5 10   0.000 0 ( 1.0000 0.0000 ) *
      11) X2 > 30 6   8.318 0 ( 0.5000 0.5000 ) *
   3) X1: A11,A12 42  57.360 1 ( 0.4286 0.5714 )
     6) X3 < 3266.5 29  40.170 0 ( 0.5172 0.4828 )
      12) X3 < 1499 16  19.870 1 ( 0.3125 0.6875 )
        24) X4 < 2.5 5   0.000 1 ( 0.0000 1.0000 ) *
        25) X4 > 2.5 11  15.160 1 ( 0.4545 0.5455 ) *
      13) X3 > 1499 13  14.050 0 ( 0.7692 0.2308 )
        26) X3 < 2243.5 7   0.000 0 ( 1.0000 0.0000 ) *
        27) X3 > 2243.5 6   8.318 0 ( 0.5000 0.5000 ) *
     7) X3 > 3266.5 13  14.050 1 ( 0.2308 0.7692 )
      14) X3 < 6595.5 8   0.000 1 ( 0.0000 1.0000 ) *
      15) X3 > 6595.5 5   6.730 0 ( 0.6000 0.4000 ) *
```

Apply the CART model to predict Y for a foreign worker when the status of existing checking account of the customer is A11, the duration is 6 months, the credit amount is 1169 and the instalment rate of disposable income is 4%. You need to write down your steps. (3 marks)

Ans.
- $X_1 = A11$, go to item 3) ..... [0.5 mark]
- $X_3 = 1169 < 3266.5$, go to 6) ..... [0.5 mark]
- $X_3 = 1169 < 1499$, go to 12) ..... [0.5 mark]
- $X_4 = 4 > 2.5$, $\boxed{Y = 1}$. ..... [1.5 marks]

(d) Suppose the confusion matrix for logistic regression is given in Table 2.1, the confusion matrix for naive Bayes model is given in Table 2.2, the confusion matrix for CART model is given in Table 2.3, if your objective is to identify the applicant with good credit risk and reject applicants with bad credit risk, state the performance metrics that meets your requirement and evaluate if the models are acceptable based on appropriate performance metrics calculations.

Table 2.1: Confusion matrix for Logistic Regression (0 is positive)

	Actual	
Prediction	0	1
0	466	98
1	184	172

Table 2.2: Confusion matrix for naive Bayes model (0 is positive)

	Actual	
Prediction	0	1
0	556	174
1	94	96

Table 2.3: Confusion matrix for CART model (0 is positive)

	Actual	
Prediction	0	1
0	446	142
1	204	128

(4 marks)

Ans. Since the data are **imbalanced** (650 zeros vs 270 ones), accuracy is not a good performance metric:

- Accuracy of logistic regression = 0.6934783
- Accuracy of naive Bayes model = 0.7086957
- Accuracy of CART model = 0.623913

None of the three models are acceptable because if we predict all to be zeros, we get an accuracy of $650/(650+270) = 0.7065217$. ..... [3 marks]

A better performance metric is the Kappa statistic which captures the recalls and the precision. ..... [1 mark]

[Total : 25 marks]

Q3. Given the three-dimensional data in Table 3.1.

Table 3.1: Three-dimensional data

Obs.	x_1	x_2	x_3
A	4	5	3
B	7	4	3
C	4	10	2
D	0	2	6
E	4	7	1
F	1	3	4

(a) Write down the mathematical formula of the Minkowski distance of order $r(\geq 1)$ for two vectors (x_1, x_2, x_3) and (y_1, y_2, y_3). (2 marks)

Ans. $\left(|x_1 - y_1|^r + |x_2 - y_2|^r + |x_3 - y_3|^r\right)^{1/r}$, $r \geq 1$. ..... [2 marks]

(b) Suppose a partial (Euclidean) distance matrix of Table 3.1 is given below:

```
          A        B        C        D
A   0
B   3.1623
C   5.0990   6.7823
D   5.8310   7.8740   9.7980
E   2.8284   4.6904   3.1623   8.1240
F   3.7417   6.1644   7.8740   2.4495
```

(i) Calculate the Euclidean distance of the point E to the point F and write down the complete distance matrix for the three-dimensional data in Table 3.1. (3 marks)

Ans. $\left(|4-1|^2 + |7-3|^2 + |1-4|^2\right)^{1/2} = \sqrt{9+16+9} = 5.8310$ ... [2 marks]

The complete distance matrix for the three-dimensional data in Table 3.1 is

```
          A        B        C        D        E        F
A         0
B   3.1623         0
C   5.0990   6.7823         0
D   5.8310   7.8740   9.7980         0
E   2.8284   4.6904   3.1623   8.1240         0
F   3.7417   6.1644   7.8740   2.4495   5.8310         0
```

..... [1 mark]

(ii) Use the complete distance matrix in part (i) to perform hierarchical clustering analysis with **complete linkage** and then draw the **dendrogram** of the hierarchical clustering. (10 marks)

Ans. The first height is 2.4495, D and F are merged as follows: .....[1 mark]

	A	B	C	D,F	E
A	0				
B	3.1623	0			
C	5.0990	6.7823	0		
D,F	5.8310	7.8740	9.7980	0	
E	2.8284	4.6904	3.1623	8.1240	0

..... [2 marks]

The second height is 2.8284, A and E are merged as follows: ..... [1 mark]

	A,E	B	C	D,F
A,E	0			
B	4.6904	0		
C	5.0990	6.7823	0	
D,F	8.1240	7.8740	9.7980	0

..... [1 mark]

The third height is 4.6904, AE and B are merged as follows: .....[1 mark]

	AE,B	C	D,F
AE,B	0		
C	6.7823		
D,F	8.1240	9.7980	0

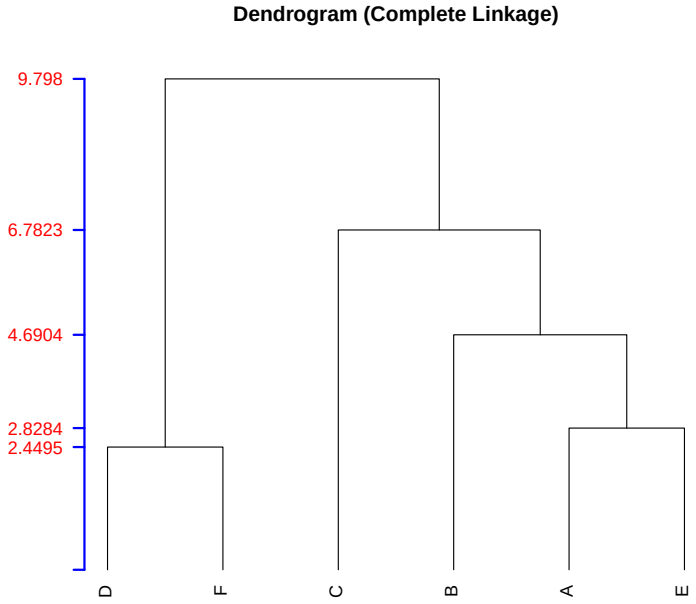
..... [1 mark]

The fourth height is 6.7823, AEB and C are merged as follows:

	AEB,C	D,F
AEB,C	0	
D,F	9.7980	0

..... [1 mark]

The dendrogram is drawn below:



Marks are deducted for poor labelling or terribly drawn lines (i.e. there should not have terrible crossings) ..... [2 marks]

(c) Perform k-means clustering algorithm using the Euclidean distance on the data from Table 3.1 with B and C as the initial centres until **two clusters** are found. Write down the stable cluster centres. You may round the numbers in your calculations to 4 decimal places. (6 marks)

Ans. Given the initial centres B(7, 4, 3), C(4, 10, 2) which correspond to cluster 1 and cluster 2.

Step 1 : Update table based on distance to cluster centres (the distance can be obtained from part (b)(i))

x_1	x_2	x_3	dist.1	dist.2	clust.centre
4	5	3	3.1623	5.0990	1
7	4	3	0	6.7823	1
4	10	2	6.7823	0	2
0	2	6	7.8740	9.7980	1
4	7	1	4.6904	3.1623	2
1	3	4	6.1644	7.8740	1

. . . . . [2 marks]

The new cluster centres are

$$Centre1 = (3, 3.5, 4), \quad Centre2 = (4, 8.5, 1.5). \qquad \text{[1 mark]}$$

Step 2 : Update table based on distance to cluster centres

x_1	x_2	x_3	dist.1	dist.2	clust.centre
4	5	3	2.0616	3.8079	1
7	4	3	4.1533	5.6125	1
4	10	2	6.8739	1.5811	2
0	2	6	3.9051	8.8600	1
4	7	1	4.7170	1.5811	2
1	3	4	2.0616	6.7454	1

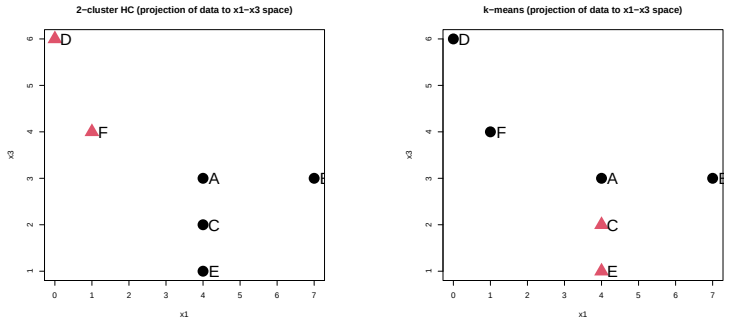
. . . . . [2 marks]

The stable cluster centres are

$$Centre1 = (3, 3.5, 4), \quad Centre2 = (4, 8.5, 1.5). \qquad \text{[1 mark]}$$

(d) With **appropriate symbols/labels**, sketch the projections of the data in Table 3.1 to the subspace spanned by the variables x_1 and x_3 with the cluster labels from the hierarchical clustering analysis from part (b) and the cluster labels from the k-means clustering from part (c). (4 marks)

Ans. The hierarchical clustering is shown on the left and the k-means is shown on the right. . . . . $[2 \times 2 = 4$ marks]



[Total : 25 marks]

PART B: Answer **ONE** question.

Q4. (a) Given the training data with three numeric features "bill length" (unit: mm), "bill depth" (unit: mm), "flipper length" (unit: mm) and the label "species" in Table 4.1.

Table 4.1: Training data of the penguin data with three different labels of penguins — Adelie, Chinstrap and Gentoo.

Obs.	bill length	bill depth	flipper length	species
A	41.1	19.1	188	Adelie
B	35.9	19.2	189	Adelie
C	36.0	17.9	190	Adelie
D	43.4	14.4	218	Gentoo
E	50.0	15.2	218	Gentoo
F	44.5	14.7	214	Gentoo
G	50.6	19.4	193	Chinstrap
H	45.7	17.0	195	Chinstrap

(i) Write down the min-max scaling for all the features in Table 4.1 which transform Table 4.1 to Table 4.2.

Table 4.2: Scaled training data from Table 4.1

Obs.	bill length	bill depth	flipper length	species
A	0.3537	0.94	0.0000	Adelie
B	0.0000	0.96	0.0333	Adelie
C	0.0068	0.70	0.0667	Adelie
D	0.5102	0.00	1.0000	Gentoo
E	0.9592	0.16	1.0000	Gentoo
F	0.5850	0.06	0.8667	Gentoo
G	1.0000	1.00	0.1667	Chinstrap
H	0.6667	0.52	0.2333	Chinstrap

(3 marks)

Ans. $S_1(x) = \dfrac{x-35.9}{14.7}$ ..... [1 mark]

$S_2(x) = \dfrac{x-14.4}{5}$ ..... [1 mark]

$S_3(x) = \dfrac{x-188}{30}$ ..... [1 mark]

(ii) Given the results listing of LDA for Table 4.2:

```
lda(species ~ ., data = D.train.s)

Prior probabilities of groups:
   Adelie Chinstrap    Gentoo
    0.375     0.250     0.375

Group means:
          bill_length bill_depth flipper_length
Adelie         0.1202     0.8667          0.0333
Chinstrap      0.8333     0.7600          0.2000
Gentoo         0.6848     0.0733          0.9556
```

By using Table 4.2, suppose the unscaled group covariance matrix for the

species Adelie is

$$\begin{bmatrix} 0.0818 & 0.0248 & -0.0116 \\ 0.0248 & 0.0419 & -0.0080 \\ -0.0116 & -0.0080 & 0.0022 \end{bmatrix},$$

the unscaled group covariance matrix for the species Gentoo is

$$\begin{bmatrix} 0.1157 & 0.0379 & 0.0133 \\ 0.0379 & 0.0131 & 0.0018 \\ 0.0133 & 0.0018 & 0.0119 \end{bmatrix},$$

find the unscaled group covariance matrix for the species Chinstrap and the estimated common covariance matrix $\widehat{C}$. (5 marks)

Ans. The unscaled group covariance for the species Chinstrap is

$$\begin{bmatrix} 1.0000-0.83335 & 0.6667-0.83335 \\ 1.00-0.76 & 0.52-0.76 \\ 0.1667-0.2 & 0.2333-0.2 \end{bmatrix} \begin{bmatrix} 1.0000-0.83335 & 1.00-0.76 & 0.1667-0.2 \\ 0.6667-0.83335 & 0.52-0.76 & 0.2333-0.2 \end{bmatrix}$$
$$= \begin{bmatrix} 0.16665 & -0.16665 \\ 0.24000 & -0.24000 \\ -0.03330 & 0.03330 \end{bmatrix} \begin{bmatrix} 0.16665 & 0.24 & -0.0333 \\ -0.16665 & -0.24 & 0.0333 \end{bmatrix}$$
$$= \begin{bmatrix} 0.0555 & 0.0800 & -0.0111 \\ 0.0800 & 0.1152 & -0.0160 \\ -0.0111 & -0.0160 & 0.0022 \end{bmatrix}$$

. . . . . [3 marks]

The estimated common covariance matrix

$$\widehat{C} = \frac{1}{8-3} \begin{bmatrix} 0.0818+0.1157+0.0555 & 0.0248+0.0379+0.0800 & -0.0116+0.0133-0.0111 \\ 0.0248+0.0379+0.0800 & 0.0419+0.0131+0.1152 & -0.0080+0.0018-0.0160 \\ -0.0116+0.0133-0.0111 & -0.0080+0.0018-0.0160 & 0.0022+0.0119+0.0022 \end{bmatrix}$$
$$= \begin{bmatrix} 0.05050 & 0.02854 & -0.00188 \\ 0.02854 & 0.03404 & -0.00444 \\ -0.00188 & -0.00444 & 0.00326 \end{bmatrix}$$

. . . . . [2 marks]

(iii) Suppose the inverse matrix of the estimated common covariance matrix $\widehat{C}$ is

$$\begin{bmatrix} 39.33 & -36.50 & -27.03 \\ -36.50 & 69.60 & 73.74 \\ -27.03 & 73.74 & 391.59 \end{bmatrix}.$$

By finding the posterior probabilities or otherwise, predict the species of penguin with a bill length in 51.3 mm, a bill depth in 19.2 mm and a flipper length in 193 mm. (7 marks)

Ans. To perform prediction, one needs to scale the features using functions from part (i):

$$\mathbf{x}^* = \frac{51.3-35.9}{14.7} = 1.0476, \quad \frac{19.2-14.4}{5} = 0.96, \quad \frac{193-188}{30} = 0.1667$$

[1 mark]

Suppose we are estimating the posterior probabilities for j (being one of the penguin species), we will be using the formula

$$P(Y=j|\mathbf{x}) \propto P(Y=j)\exp(-\frac{1}{2}(\mathbf{x}-\mu_j)\widehat{C}^{-1}(\mathbf{x}-\mu_j)^T). \quad \text{[1 mark]}$$

Let $A = Adelie$, $C = Chinstrap$ and $G = Gentoo$.

$$P(Y = A|\mathbf{x}^*) \propto 0.375 \exp(-\frac{1}{2} \begin{bmatrix} 1.0476 - 0.1202 \\ 0.96 - 0.8667 \\ 0.1667 - 0.0333 \end{bmatrix}^T \widehat{C}^{-1} \begin{bmatrix} 0.9274 \\ 0.0933 \\ 0.1334 \end{bmatrix}^T)$$

$$= 0.375 \exp(-\frac{30.2321}{2})$$

[2 marks]

$$P(Y = C|\mathbf{x}^*) \propto 0.250 \exp(\begin{bmatrix} 1.0476 - 0.8333 \\ 0.96 - 0.7600 \\ 0.1667 - 0.2000 \end{bmatrix}^T \widehat{C}^{-1} \begin{bmatrix} 0.2143 \\ 0.2000 \\ -0.0333 \end{bmatrix}^T) \quad \text{[1 mark]}$$

$$= 0.250 \exp(-\frac{1.299226}{2})$$

$$P(Y = G|\mathbf{x}^*) \propto 0.375 \exp(\begin{bmatrix} 1.0476 - 0.6848 \\ 0.96 - 0.0733 \\ 0.1667 - 0.9556 \end{bmatrix}^T \widehat{C}^{-1} \begin{bmatrix} 0.3628 \\ 0.8867 \\ -0.7889 \end{bmatrix}^T) \quad \text{[1 mark]}$$

$$= 0.375 \exp(-\frac{192.4342}{2})$$

Since $P(Y = G|\mathbf{x}^*)$ and $P(Y = A|\mathbf{x}^*)$ are very small, the penguin species is predicted to be Chinstrap. ..... [1 mark]

(b) Given the two-dimensional data in Table 4.3.

Table 4.3: Two-dimensional data

x_1	x_2
4.4	4.8
1.2	11.1
5.0	9.7
7.8	9.6
5.9	6.9
5.8	9.2
3.6	7.6

Suppose the covariance matrix of the data in Table 4.2 is

$$\begin{bmatrix} 4.3014 & -0.7186 \\ -0.7186 & 4.4848 \end{bmatrix},$$

write down the principal components of a data $\mathbf{x}$ related to the data in Table 4.3 by first finding all normalised eigenvectors of the PCA. (10 marks)

Ans. From the quadratic equation

$$\begin{vmatrix} 4.3014-\lambda & -0.7186 \\ -0.7186 & 4.4848-\lambda \end{vmatrix} = \lambda^2 - 8.7862\lambda + 18.7745 = 0 \qquad \text{[3 marks]}$$

we obtain the eigenvalues of the covariance matrix:

$$\lambda = 5.1175,\ 3.6687 \qquad \text{[1 mark]}$$

The normal eigenvector corresponding to $\lambda = 5.1175$ is obtained from

$$\begin{bmatrix} 4.3014-5.1175 & -0.7186 \\ -0.7186 & 4.4848-5.1175 \end{bmatrix} \mathbf{x}_1 = \mathbf{0}$$
$$\Rightarrow \mathbf{x}_1 = \frac{1}{\sqrt{((-0.7186)^2+(-0.8161)^2}} \begin{bmatrix} -0.7186 \\ -(-0.8161) \end{bmatrix} = \begin{bmatrix} -0.6608518 \\ 0.7505165 \end{bmatrix} \qquad \text{[2 marks]}$$

By orthogonality, the normal eigenvector corresponding to $\lambda = 3.6687$ is

$$\mathbf{x}_2 = \begin{bmatrix} 0.7505165 \\ 0.6608518 \end{bmatrix} \qquad \text{[1 mark]}$$

To write down the principal components, we first find the column average

$$(4.814286,\ 8.414286) \qquad \text{[1 mark]}$$

The first principal component is

$$PC_1(\mathbf{x}) = -0.6608518(x_1 - 4.814286) + 0.7505165(x_2 - 8.414286) \qquad \text{[1 mark]}$$

The second principal component is

$$PC_2(\mathbf{x}) = 0.7505165(x_1 - 4.814286) + +0.6608518(x_2 - 8.414286) \qquad \text{[1 mark]}$$

[Total : 25 marks]

Q5. (a) Given the training data with four numeric features "bill length" (unit: mm), "bill depth" (unit: mm), "flipper length" (unit: mm), "body mass" (unit: g) and the label "species" in Table 5.1.

Table 5.1: Training data of the penguin data with three types of penguins — Adelie, Chinstrap and Gentoo.

Obs.	bill length	bill depth	flipper length	body mass	species
A	41.1	19.1	188	4100	Adelie
B	50.6	19.4	193	3800	Chinstrap
C	45.7	17.0	195	3650	Chinstrap
D	43.4	14.4	218	4600	Gentoo
E	44.5	14.7	214	4850	Gentoo
F	35.9	19.2	189	3800	Adelie
G	36.0	17.9	190	3450	Adelie
H	50.0	15.2	218	5700	Gentoo

(i) Use the supervised learning model kNN (k=3) with the Euclidean distance to predict the species of a penguin with a bill length of 38.9 mm, a bill depth of 17.8 mm, a flipper length of 181 mm and a body mass of 3625 g. You may round the distance to 2 decimal places. (9 marks)

Ans. By calculating the Euclidean distance from the point (46.5, 14.8, 217, 5200) to the points in the training data, we can obtain the following table:

Obs.	bill length	bill depth	flipper length	body mass	species	Dist
A	41.1	19.1	188	4100	Adelie	475.0584
B	50.6	19.4	193	3800	Chinstrap	175.8080
C	45.7	17.0	195	3650	Chinstrap	29.4598
D	43.4	14.4	218	4600	Gentoo	975.7181
E	44.5	14.7	214	4850	Gentoo	1225.4611
F	35.9	19.2	189	3800	Adelie	175.2140
G	36.0	17.9	190	3450	Adelie	175.2553
H	50.0	15.2	218	5700	Gentoo	2075.3612

. . . . . [8 marks]

The 3 nearest neighbours are observations C, F and G, which correspond to species Chinstrap, Adelie and Adelie. Therefore, the prediction of the species of the penguin is Adelie. . . . . [1 mark]

(ii) Based on your calculation in part (i), explain the problem with predictive modelling and what data preparation step is required to resolve the problem. (2 marks)

Ans. Problem: the data are not scaled — "body mass" has a large variation compare to other features . . . . . [1 mark]

Solution: introduce min-max scaling or standardisation to bring all the features to similar variations. . . . . [1 mark]

(b) Given the unlabelled iris flower data in Table 5.2.

Table 5.2: Unlabelled iris flower data.

Obs.	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
A	4.4	3.0	1.3	0.2
B	5.5	2.6	4.4	1.2
C	7.7	2.8	6.7	2.0
D	4.6	3.1	1.5	0.2
E	5.6	2.5	3.9	1.1
F	6.2	3.4	5.4	2.3

Suppose the results of the principal component analysis from R are listed below:

```
Standard deviations (1, .., p=4):
[1] 2.56885138 0.42639622 0.32438097 0.07723882

Rotation (n x k) = (4 x 4):
                    PC1          PC2         PC3         PC4
Sepal.Length  0.45219359  0.42363786  0.7307630  0.2864216
Sepal.Width  -0.01145146 -0.64043843  0.5769823 -0.5067533
Petal.Length  0.83095985  0.01028704 -0.3510437 -0.4314721
Petal.Width   0.32387581 -0.64051835 -0.0992227  0.6891992
```

(i) State the proportions of variance explained (PVE) and state the number of principal components to be considered if a targeted cumulative PVE (CPVE) of 90% is set. You may round all your calculations to 2 decimal places.

(5 marks)

Ans. The PVE (rounded to 2 decimal places) are

$$\frac{(2.57^2, 2.57^2, 0.32^2, 0.08^2)}{2.57^2+0.43^2+0.32^2+0.08^2} = \frac{(6.6, 0.18, 0.1, 0.01)}{6.89} = (0.96,\ 0.03,\ 0.01,\ 0.00)$$

[4 marks]

Since the first principal component explains 96% of the variation of the data, only 1 principal component should be considered. . . . . [1 mark]

(ii) Write down the first two principal components of the data in Table 5.2 and then perform appropriate calculations to sketch the biplot of the data in Table 5.2 with proper labels. You may round all the numbers to 2 decimal places.

(9 marks)

Ans. To find the principal components, we first calculate the column means of the data in Table 5.2:

$$(5.67,\ 2.9,\ 3.87,\ 1.17).$$ [2 marks]

Next we write down the first two principal components:

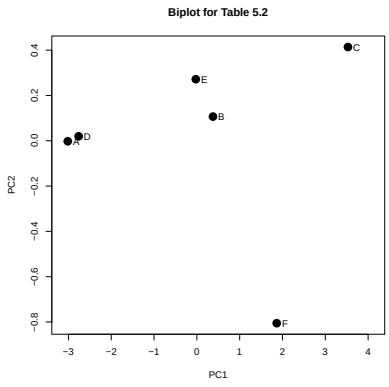
$$PC_1 = 0.45(x_1 - 5.67) - 0.01(x_2 - 2.9) + 0.83(x_3 - 3.87) + 0.32(x_4 - 1.17)$$
$$PC_2 = 0.42(x_1 - 5.67) - 0.64(x_2 - 2.9) + 0.01(x_3 - 3.87) - 0.64(x_4 - 1.17)$$

[2 marks]

Using the above formulas, we can obtain the first two principal components and the biplot:

Obs.	PC1	PC2
A	-3.016	-0.0023
B	0.376	0.1067
C	3.529	0.4137
D	-2.761	0.0197
E	-0.025	0.2717
F	1.865	-0.8053

[3 marks]



[2 marks]

[Total : 25 marks]