Strathmore
UNIVERSITY

STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCES
MASTER OF SCIENCE IN STATISTICAL SCIENCES
END OF SEMESTER EXAMINATION
STA 8203: PREDICTIVE MODELING AND STATISTICAL LEARNING

DATE: 22nd April, 2022 Time: 2 Hours

Instructions

1. This examination consists of **FIVE** questions and an appendix to one of the questions.
2. Answer **Question ONE (COMPULSORY)** and any other **TWO** questions.

Question 1 (20 Marks)

a) Exploratory Data Analysis (EDA) is an approach/philosophy employed in data analysis. Explain how this approach is employed, how it differs from classical and Bayesian analysis and hence enumerate the EDA assumptions.

(**7** Marks)

b) Data mining tasks can be categorized into 5 main kinds of tasks. Describe each of these task, explaining how each is performed

(**9** Marks)

c) Distinguish between supervised and unsupervised learning

(**4** Marks)

Question 2 (20 Marks)

a) Cross-validation is an important tool in predictive modeling. Describe how the following cross-validation techniques work: Leave-One-Out Cross-Validation; and k-Fold Cross-Validation. (**9** Marks)

b) What are the advantages and disadvantages of k-fold cross-validation relative to:

i) The validation set approach. (**3** Marks)

ii) LOOCV? (**3** Marks)

c) Suppose that we use some data mining method to make a prediction for the response **Y** for a particular value of the predictor **X**. Carefully describe how we might estimate the standard deviation of our prediction. (**5** Marks)

Question 3 (20 Marks)

a) Consider the linear regression model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim MVN(\mathbf{0}, \sigma^2 \boldsymbol{I})$.

i) Explain what the hat-matrix is.

ii) Explain how standardized residuals are used in regression diagnostics and use a mathematical approach to show that

$$Z = \frac{e_i}{s.e.(e_i)} = \frac{e_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}} \sim N(0,1)$$

iii) Explain how outliers, high-leverage values, and influential observations are identified on the basis of the hat-matrix.

(**12** Marks)

b) The Dixon and the generalized (extreme Studentized deviate) ESD (Rosner) tests are approaches used in exploratory data analysis. Distinguish between them and explain in (mathematical) detail how each approach works.

(**8** Marks)

Question 4 (20 Marks)

a) There are many possible classification techniques, or classifiers, that one classifier might use to predict a qualitative response. Two most widely-used classifiers are logistic regression, linear discriminant analysis. Distinguish between these two methods.

(**7** Marks)

b) A team of researcher at **CDC-Kenya** would like to develop *a predictive model* for HIV serostatus in Kenya using two explanatory variables: age at first sex; and sex for gifts. The data used in the study was the most recent Kenya Aids Indicator Survey. A description of variables considered is presented in the Table 1.

Table 1 Variables considered in the study

Variable Name	Levels	Meaning
Hivstatus	0=Negative; 1=Positive	HIV serostatus
age1stsex	1=Before 18 years; 2=18+ years; 3=Don't know; 4=Refused	Age at first sex
sex4gifts_ever	1=No; 2=Yes	Ever received sex for gifts in the past year

Appendix 1 presents a summary of logistic regression models fitted to these data.

- **Fit 1**: a logistic regression model with age at first sex as the only predictor

- **Fit 2**: a logistic regression model with sex for gifts as the only predictor
- **Fit 3**: a logistic regression model with age at first sex and sex for gifts as the predictors

Table 2 Confusion matrices for the three models considered

Fit 1: age1stsex as the only predictor

	Predicted status	
Actual status	Negative	Positive
Negative	5877	5101
Positive	398	390

Fit 2: sex4gifts_ever as the only predictor

	Predicted status	
Actual status	Negative	Positive
Negative	376	10602
Positive	37	611

Fit 3: age1stsex and sex4gifts_ever predictors

	Predicted status	
Actual status	Negative	Positive
Negative	6005	497
Positive	408	240

i) From the results present in Fit 1: a logistic regression model with age at first sex as the only predictor

- **Fit 2**: a logistic regression model with sex for gifts as the only predictor
- **Fit 3**: a logistic regression model with age at first sex and sex for gifts as the predictors

ii) Table 2, compare the 3 models. Compare your results.

(**4** Marks)

iii) Based on the results presented in part (i) above and the information provided in the Appendix, develop a table of effects with relevant unadjusted and adjusted odds ratios. Comment on your results.

(**4** Marks)

iv) For the best fitting model, compute the following measures: sensitivity, specificity and the false positive rate.

(**4** Marks)

Question 5 (20 Marks)

a) Distinguish between partitioning and hierarchical clustering approaches.

(6 Marks)

b) The K-means and PAM are two common partitioning cluster analysis approaches in the literature. Describe each algorithm and provide any advantages or disadvantages of one over the other.

(8 Marks)

c) Distinguish between Agglomerative clustering and Divisive clustering algorithms.

(6 Marks)

Appendix

```
> fit1=glm(hivstatus~age1stsex,family=binomial,data=kais)
> fit2=glm(hivstatus~sex4gifts_ever,family=binomial,data=kais)
> fit3=glm(hivstatus~age1stsex+sex4gifts_ever,family=binomial,data=kais)
> summary(fit1)

Call:
glm(formula = hivstatus ~ age1stsex, family = binomial, data = kais)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4512  -0.3590  -0.3590  -0.3093   2.4753

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           -2.70965    0.05651 -47.951  < 2e-16 ***
age1stsex2 18+ years  -0.30609    0.08596  -3.561  0.00037 ***
age1stsex3 Don't know  0.09789    0.14419   0.679  0.49720
age1stsexRefused       0.47605    0.61012   0.780  0.43523
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5000.9  on 11625  degrees of freedom
Residual deviance: 4984.4  on 11622  degrees of freedom
  (2094 observations deleted due to missingness)
AIC: 4992.4

Number of Fisher Scoring iterations: 5

> summary(fit2)

Call:
glm(formula = hivstatus ~ sex4gifts_ever, family = binomial,
    data = kais)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4333  -0.3348  -0.3348  -0.3348   2.4124

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -2.8537     0.0416 -68.592  < 2e-16 ***
sex4gifts_everYes   0.5350     0.1772   3.019  0.00254 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5000.9  on 11625  degrees of freedom
Residual deviance: 4992.9  on 11624  degrees of freedom
  (2094 observations deleted due to missingness)
AIC: 4996.9

Number of Fisher Scoring iterations: 5

> summary(fit3)

Call:
glm(formula = hivstatus ~ age1stsex + sex4gifts_ever, family = binomial,
    data = kais)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5649  -0.3542  -0.3542  -0.3069   2.4815

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           -2.73731    0.05778 -47.377  < 2e-16 ***
age1stsex2 18+ years  -0.29449    0.08612  -3.419 0.000628 ***
age1stsex3 Don't know  0.10286    0.14426   0.713 0.475849
age1stsexRefused       0.48423    0.61043   0.793 0.427621
sex4gifts_everYes      0.49857    0.17769   2.806 0.005018 **
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5000.9  on 11625  degrees of freedom
Residual deviance: 4977.5  on 11621  degrees of freedom
  (2094 observations deleted due to missingness)
AIC: 4987.5

Number of Fisher Scoring iterations: 5
```