NAMIBIA UNIVERSITY
OF SCIENCE AND TECHNOLOGY

FACULTY OF COMPUTING AND INFORMATICS

DEPARTMENT OF INFORMATICS

QUALIFICATION: Bachelor of Informatics Honours	
QUALIFICATION CODE: 08BIHW, 08BIFB	COURSE LEVEL: NQF LEVEL 8
COURSE: Data Science and Analytics	COURSE CODE: DSA821S
DATE: NOVEMBER 2024	SESSION: 1
DURATION: 2 Hours	MARKS: 60

FIRST OPPORTUNITY EXAMINATION QUESTION PAPER	
EXAMINER(S):	MR. SEBASTIAN MUKUMBIRA
MODERATOR (S):	MS. EMILIA SHIKEENGA

THIS EXAMINATION PAPER CONSISTS OF 3 PAGES
(INCLUDING THIS FRONT PAGE)

Instructions for the students

1. There are four Sections in this paper Section 1, Section 2, Section 3 and Section 4.
2. Answer ALL the questions in ALL Sections.
3. Write clearly and neatly.
4. Number the answers clearly.
5. Non-programmable calculators may be used.

Question 1: Regression Analysis [15 marks]

Answer the following questions:

(a) Explain the difference between linear regression and multiple regression. Provide an example for each. **(4 marks)**

(b) What is the purpose of the R-square (R^2) statistic in regression analysis, and how is it interpreted? **(3 marks)**

(c) What is overfitting in regression analysis, and how can it be prevented? **(3 marks)**

(d) Suppose you have the following regression equation to predict a student's exam score based on hours of study (Hours) and their attendance (Attendance, measured in days):

$$\text{ExamScore} = 50 + 3 \times \text{Hours} + 2 \times \text{Attendance}$$

(i) Interpret the coefficients for Hours and Attendance. **(3 marks)**

(ii) Predict the exam score for a student who studied for 10 hours and attended 20 days of classes. **(2 marks)**

Question 2: Association Analysis [15 marks]

An online retail store has collected data on user purchases across two product categories: Electronics and Clothing. The following incomplete table summarises the data:

	Electronics	~ Electronics	Clothing	~ Clothing	Total
Purchased	300	150		200	500
~ Purchased		100	400	50	800
Total	550		500	250	

(a) Complete the table. **(4 marks)**

(b) Calculate the support and confidence for the association rule 'Electronics Clothing'. Does the rule meet the thresholds of 10% minimum support and 50% minimum confidence? **(6 marks)**

(c) Calculate the lift for the association rule 'Electronics Clothing' and interpret the result. **(5 marks)**

Question 3: Machine Learning [15 marks]

Consider a dataset containing customer transaction history at an e-commerce company. You are tasked with using machine learning techniques to predict whether a customer will make a purchase in the next 30 days. The available features include customer demographics, browsing history, previous purchases, and time spent on the website.

(a) Explain the difference between supervised and unsupervised learning. Which type of learning would you use for this problem, and why? **(4 marks)**

(b) Define overfitting in the context of machine learning. What strategies can you use to prevent overfitting in this customer purchase prediction model? **(3 marks)**

(c) You decide to use logistic regression to solve this problem. What are the key assumptions made by logistic regression? Does this model have any limitations for predicting customer purchases? **(3 marks)**

(d) You have a dataset of 10,000 customers and decide to split the data into 80% training and 20% testing sets. Explain the purpose of this data split and how you would evaluate the model's performance. **(3 marks)**

(e) If you were to use a random forest model for this task, what would be the benefit of using this over logistic regression? Mention at least two advantages. **(2 marks)**

Question 4: Classification Analysis [15 marks]

A classification model was developed to predict whether students would pass or fail three subjects: Math, Science, and History. Based on the results of a test set of 25 students, the tables below show the confusion matrices for each subject:

Math	Predicted Pass	Predicted Fail	Total
Actual Pass	10	5	15
Actual Fail	2	8	10
Total	12	13	25

Science	Predicted Pass	Predicted Fail	Total
Actual Pass	8	6	14
Actual Fail	3	8	11
Total	11	14	25

History	Predicted Pass	Predicted Fail	Total
Actual Pass	12	3	15
Actual Fail	4	6	10
Total	16	9	25

(a) For Math predictions: **(3 marks)**

(i) Calculate the accuracy of the model. (1 mark)

(ii) Calculate the precision for predicting "Pass". (1 mark)

(iii) Calculate the recall for predicting "Pass". (1 mark)

(b) For Science predictions: **(3 marks)**

(i) Calculate the F1-score for predicting "Pass". (2 marks)

(ii) Explain the significance of the F1-score. (1 mark)

(c) For History predictions: **(3 marks)**

(i) Calculate the specificity of the model. (2 marks)

(ii) Interpret the result. (1 mark)

(d) Compare the performance of the classification model across all three subjects using accuracy. Which subject does the model perform best in, and why might this be the case? **(3 marks)**

(e) Analyse the recall for "Pass" predictions across all subjects. What does this metric tell us about the model's ability to correctly predict students who pass? **(3 marks)**