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: 2024/2025	**SESSION: 1**
DURATION: 2 Hours	**MARKS: 60**

SUPPLEMENTARY/SECOND OPPORTUNITY EXAMINATION QUESTION PAPER	
EXAMINER(S):	**MR. SEBASTIAN MUKUMBIRA**
MODERATOR (S):	**MS. EMILIA SHIKEENGA**

THIS EXAMINATION PAPER CONSISTS OF 5 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: Exploratory Data Analysis (EDA) [15 marks]

A dataset contains information on the sales of a retail store for a year, including columns for Date, Product, Quantity Sold, and Revenue. The first few rows of the dataset are as follows:

Date	Product	Quantity Sold	Revenue
2023-01-01	A	30	600
2023-01-01	B	20	400
2023-01-02	A	25	500
2023-01-02	B	15	300
2023-01-03	A	35	700

(a) Calculate the total revenue generated from each product for the first three days of sales. **(3 marks)**

(b) Determine the average quantity sold per day for each product. **(4 marks)**

(c) Identify the day with the highest total revenue and specify the revenue amount. **(4 marks)**

(d) Discuss two potential insights you can draw from the EDA of this dataset. **(2 marks)**

(e) Propose one data cleaning step that could improve the quality of this dataset. **(2 marks)**

Question 2: Hypothesis Testing [15 marks]

A retail company wants to know if there is a significant association between the type of advertisement used and the purchase decision made by customers. The data collected is summarised in the following contingency table:

Advertisement Type	Purchase (Yes)	Purchase (No)
Online	40	10
TV	30	20
Print	20	30

(a) State the null and alternative hypotheses for the above mentioned market research. **(2 marks)**

(b) Calculate the expected frequencies for each cell in the contingency table. Show your calculations. **(5 marks)**

(c) Perform the Chi-square test. Calculate the Chi-square statistic and determine the p-value. What can you conclude? **(5 marks)**

Use the following formula for chi-square statistic calculation:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Where O = observed frequencies, E = expected frequencies.

(d) Discuss one assumption of the Chi-square test that should be considered. **(2 marks)**

(e) Suggest a follow-up analysis if a significant association is found. **(1 mark)**

Question 3: Clustering Analysis [15 marks]

Consider the following dataset representing customers based on their annual income and spending score:

Customer ID	Annual Income	Spending Score
1	15,000	39
2	20,000	81
3	25,000	6
4	30,000	77
5	35,000	40

(a) Calculate the Euclidean distance between Customer 1 and Customer 2 based on the provided features. **(3 marks)**

(b) Explain the concept of K-means clustering and its application in customer segmentation. **(4 marks)**

(c) If you were to group the customers into clusters, how would you determine the optimal number of clusters? Discuss. **(4 marks)**

(d) Demonstrate one iteration of the K-means clustering algorithm for this dataset, assuming three clusters. Your answer should include:

- Selection of initial centroids
- Assignment of data points to clusters
- Recalculation of cluster centroids

Clearly explain your process and show all necessary calculations. **(3 marks)**

(e) Discuss one potential drawback of using K-means clustering. **(1 mark)**

Question 4: Decision Trees [15 marks]

Consider a dataset that includes features like Age, Income, and Loan Status (Approved or Rejected) for loan applications:

Age	Income	Loan Status
25	30000	Approved
30	40000	Approved
35	50000	Rejected
40	60000	Approved
45	70000	Rejected

(a) Explain the concept of a decision tree and how it is used for classification tasks. **(3 marks)**

(b) Calculate the Gini impurity for the Loan Status variable in the dataset. Show your calculations. **(5 marks)**

Use the following formula for Gini impurity calculation:

$$G = 1 - \sum_{i=1}^{n} p_i^2$$

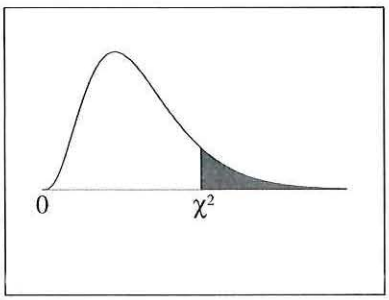
where p_i is the probability of class i.

(c) How would you handle missing values in a dataset before training a decision tree model? Discuss one method. **(3 marks)**

(d) Describe how overfitting can occur in decision trees and propose one technique to mitigate it. **(3 marks)**

(e) Provide an example of a real-world application where decision trees could be effectively used. **(1 mark)**

Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169