NEPAL COLLEGE OF INFORMATION TECHNOLOGY

SPRING SEMESTR 2025

Level: Bachelor
Program: BEIT
Course: **Data Science and Analytics**

Year: 2025
Full Marks: 100
Pass Marks: 45
Time: 3 hrs.

Candidates are required to answer in their own words as far as practicable.
The figures in the margin indicate full marks.
Attempt all the questions.

1. a) Explain the concept of Knowledge Discovery from Data (KDD) and how it relates to the Data Science Pipeline. 7

   b) Differentiate between structured and unstructured data and Distinguish between numeric and categorical variables. 4+4

2. a) Explain Pearson correlation and how it measures relationship between two numeric variables show it by considering the example. 7

   b) Illustrate the empirical distribution of numeric and categorical data using the following: 8
   1. For numeric: Draw and interpret a histogram. Explain the significance of normal and power-law distributions.
   2. For categorical: Use a bar plot and explain Zipf's law with an example.

3. a) A real estate analyst collected data from 3 houses in a neighborhood to predict house prices based on two features: square footage (in 100s of square feet) and number of bedrooms. 7

Square footage(X1)	Number of bed rooms(X2)	House price (in $1000s) (Y)
15	3	240
18	4	290
20	3	310

Find the regression equation: $Y = \beta 0 + \beta 1X1 + \beta 2X2$

b) A data analyst builds a multivariate linear regression model to predict house prices based on features such as square footage, number of bedrooms, and distance to the city center. 3+5

1. Explain the significance of R2 and Adjusted R2 in evaluating the performance of the model. Why might Adjusted R2 be a better metric when dealing with multiple predictors?
2. While analyzing the model, the analyst suspects that some predictors (e.g., square footage and number of rooms) are highly correlated.(Explain in term of multicollinearity and how can it affect the model)
3. Describe how Variance Inflation Factor (VIF) is used to detect multicollinearity and interpret what high VIF values imply.

4. a) Derive the Nadaraya-Watson kernel regression estimator for the conditional mean E[Y□X=x]. Clearly show how the formula is obtained using kernel weighting .The performance of the Nadaraya-Watson estimator heavily depends on the choice of the bandwidth parameter. Explain the role of the bandwidth in kernel regression and discuss strategies for selecting an optimal bandwidth (e.g., cross-validation, rule-of-thumb methods). 7 8

   b) You are given the following two-dimensional data set consisting of 6 observations:

   (2,1),(3,5),(4,3),(5,6),(6,7),(7,8)
   Using the Principal Component Analysis (PCA) algorithm, compute the first principal component of the dataset

5. a) You are given the following dataset of customers, where the target variable is whether the customer purchased a product (`Yes` or `No`): 7

Age	Income	Purchased
young	high	no
young	medium	yes
middle	high	yes
senior	Low	no
senior	medium	yes
middle	Low	yes

Using the CART (Classification and Regression Tree) algorithm, construct a binary classification tree (up to two levels deep) for predicting whether a customer will purchase the product.

b) Apply the K-Means algorithm with K=2 on the following dataset:(160,55),(175,65),(172,60),(180,70),(178,68),(165,58) Use the first two data points as the initial centroids: and(160,55)and(175,65) Perform the clustering for two iterations, and at each step:

- Assign points to the nearest centroid
- Recompute the centroids
- Show the cluster membership after each iteration

8

6. a) What is time series analysis? Explain four components of time series: trend, seasonal variation, cyclical variation and irregular variation. Describe autocorrelation and stationarity with respect to time series analysis. 3+4

b) Explain the differences between the ARMA(Auto-Regressive Moving Average) and ARIMA(Auto-Regressive Integrated Moving Average) models. Define the components of each model:Auto-Regressive (AR) Moving Average(MA) Integration (I). 8

7. Write short notes on: **(Any two)** 2×5

a) permutation tests, partial and partial dependence plots
b) Causation vs. correlation
c) Statistical Significance: p-value
d) logistic regression

Pokhara University

Everest Engineering College

Final Internal Assessment
Spring- 2025

Level: Bachelor — **F.M. 100**
Program: BE IT(6th Semester) — **P.M. 45**
Faculty: Science & Technology — **Time: 3hrs**
Section: A/B
Subject: Data Science and Analytics

Attempt all the questions.

1 a) Why do we require data transformation? Explain by giving realistic examples. 7

b) You are given thirty numbers: 3.22, 5.38, 4.95, 4.87, 6.67, 4.46, 4.55, 5.42, 5.85, 4.81, 6.5, 5.66, 4.32, 5.21, 6.07, 3.97, 5.3, 5.37, 5.17, 5.35, 4.47, 5.98, 5.93, 5.5, 6.73, 4.33, 6.8, 4.76, 5.37, 6.15. Do you think the distribution is uniform, exponential, or normal? Justify using histogram plot. 8

2 a) Define a power law distribution and give real-life examples for it. Why power laws are called heavy-tailed distribution? Explain by comparing power laws to normal and exponential distribution? 8

b) Suppose you surveyed 5 male students and found their average daily spending on refreshments is Rs. 730, with a sample standard deviation of Rs. 200. A previous study shows that female students spend Rs. 810 per day on average. Using a t-test, determine if male students' daily spending is significantly different from that of female students. (Use $|t| > 1.8$ as the threshold for significance.) 7

3 a) How is the multivariate linear regression problem formulated in matrix notation? How is the solution obtained? Describe. 8

b) You are given the following data.

x	0.5	1.0	1.5	2.0
y	-2.3	0.5	9.3	20.8

Estimate the value of y at $x = 1.3$ using the Nadaraya-Watson regression estimator. Take bandwidth $h = 0.2$ 7

4 a) Explain Principal Component Analysis as a tool of identifying the latent variables in the data. 7

b) Suppose that while building a decision tree classifier, the current node corresponds to the following data. 8

A	B	Class label
T	F	Black
T	T	Black
T	T	Black
T	F	White
T	T	Black
F	F	White
F	F	White
F	F	White
T	T	White
T	F	White

Should the decision tree algorithm use the attribute A or the attribute B for splitting? Justify. You can use any purity index, Gini or entropy, for the problem.

5 a) Perform one iteration of k-means clustering given the following data. 7

Items	x	y
A	5	3
B	-1	1
C	1	-2
D	-3	-2

Take the number of centroids k=2. Assume that initially A and B are allocated to cluster 1 and C and D are allocated to cluster 2.

b) Why is logistic regression required? Describe how logistic regression models the probability $P(y = 1|\, x)$ where y is a binary variable. 8

6 a) What is a stationary time series? Describe the ARMA model. 7

b) How can we be sure that an event A causes an event B? Differentiate correlation and causation with an appropriate example. 8

7 Write short notes on: (Any two) 2*5=10

a) Granger Causality

b) Statistical significance

c) The KDD process

Best Wishes

College of Engineering and Management

Nepalgunj – 10,Banke

Level : Bachelor
Programmer : BE IT
Course : Data Science & Analytics

Semester :- Spring (New)
Final Internal Assessment

Year : 2025
Full Marks : 100
Pass Marks : 45
Time : 3 hrs.

Candidates are required to give their answer in their own words as far as practicable.
The figure in the margin indicate full marks.
Attempt all the questions

1. a. Define Data science. Explain knowledge discovery form Database process in brief. (8)
   b. Difference between structured & unstructured data. Explain the steps in data processing. (7)
2. a. Define OLS estimation. Suppose we have following data on hours studied (x) & exam scores (Y). (8)

Hour studied (X)	Exam score (Y)
2	60
4	70
6	80
8	90

   Calculate regression equation (y).
3. a. Define Multicollinearity & variance inflation factor. Compute the principal components for the following 2D data. (7)
   $x=(x_1, x_2) = (2,4),(3,4),(4,8),(5,10)$
   b. Define Decision Tree. Explain about CART algorithm in detail. (7)
4. a. Suppose you have data on whether a student purchase a product the student's pocket money per month (in units of Rs 1000) and the product rading (between 1 to 5). Using statistical software, you obtained the following logistic regression, model: og $\left(p/1-p\right)$ = 5+1.2 pocket money+0.8 rating. Where 'p' is the probability that the student purchase the product. (8)

i. Logistic Regression model as a classis, suppose a product rating has rating 1 & a student purchase the product. What if pocket money is 2 & rating is 3?

ii. Suppose a product has rating 2. What is the probability that a student with pocket money 2 will purchase the product.

b. A mall wants to segment its customers based an annual income & spending score (1-100) to design targeted marketing campaigns. (7)

Customer	Annual Income & 1000	Spending score (1-100)
1	15	39
2	15	81
3	55	65
4	65	25
5	85	75

Using K-means clustering. Find the final chutes.

5. a. Below are the values of two covariates x_1 & x_2. (8)

X_1	X_2
0.50	0.25
-0.14	0.43
0.65	0.94
1.52	1.64
-0.23	-0.37
-0.23	-0.18
1.58	1.36
0.77	0.87
-0.47	-0.60
-0.47	-0.41

Calculate the Kart pearson correlation coefficient between x_1 to x_2.

b. The following table presents a hypothetical cross-tabulation between food preference & gender. (7)

	Female	Male	Other
Likes buff momo	8	75	1
Likes Chicken momo	63	3	0
Likes Veg. momo	23	19	3

Do you think food preference are correlated with gender? Justify using chi-squared test with significant level 0.05. The critical value for the chi-squared static with 4-degree of freedom at a significant level of 0.052 is 9.4877.

6. a. You are given for series A&B, the following values for time t= 1,2…6, . (8)

Time t	1	2	3	4	5	6
A	5	7	9	11	13	15
B	3	7	5	8	6	7

Which of these two series is more likely to be stationary? Justify your answer

b. Explain time series analysis of two models ARMA & ARIMA. (7)

7. Write short notes on: (Any two) (2×5)
   a. Causation vs Correlation
   b. DAG
   c. Data integration

COSMOS COLLEGE
OF MANAGEMENT AND TECHNOLOGY

Term Test II

Date: 2082/04/05			
Level	BE	Full Marks	70
Programme	BEIT	Time	
Semester	VI	2 hrs	

Subject: - Data Science & Analytics

- ✓ *Candidates are required to give their answers in their own words as far as practicable.*
- ✓ ***Attempt any 5 questions of 7 marks and 5 of 8 marks and 2 short notes.***
- ✓ *The figures in the margin indicate **Full Marks**.*
- ✓ *Assume suitable data if necessary.*

1. a) Briefly explain the Knowledge Discovery from Database (KDD) process with its major phases. A numeric variable has the following values: 11.49, 9.59, 12.94, 16.57, 8.30, 8.30, 16.74, 13.30, 6.31, 10.83, 10.47, 10.64, 13.24, 7.09, 7.28, 13.44, 9.99, 8.09, 9.07, 6.30. Discretize the values in five levels based on the frequencies. 4+3 =7

b) How does data integration address the challenge of combining data from multiple sources, multiple formats and forms? A numeric variable has the following values.Standardize the values.

Values: 46.93, 138.07, 64.77, 51.46, 15.38, 15.37, 11.32, 223.21, 120.23, 19.85, 19.80, 12.39, 55.79, 86.31, 3.99, 101.51, 30.71, 94.28, 102.39, 18.04 4+4 =8

2. a) Differentiate discrete and continuous numeric variables with examples. What are nominal and ordinal categorical variables? Explain with examples. 7

b) You are given the following dataset of 5 observations with two numeric variables Age and Income:

CustomerID	Age	Income
001	25	50000
002	35	62000
003	30	58000
004	22	48000
005	28	54000

i) Calculate the Pearson correlation coefficient between Age and Income. Show all necessary intermediate steps.

ii) ii) Interpret the strength and direction of the correlation. Support your interpretation with an appropriate plot and briefly explain how the visualization confirms the relationship.

iii) Pearson correlation may not capture non-linear relationships between variables. Why? Explain with an appropriate example. 8

3. a) The following table presents a hypothetical cross-tabulation between 'food preferences' and 'gender'.

	Female	Male	Other
Likes buff momo	8	75	1
Likes chicken momo	63	3	0
Likes veg momo	23	19	3

Do you think food preferences are correlated with gender? Justify using chi-squared test with significance level 0.05. The critical value for the Chi-Squared statistic with 4 degrees of freedom at a significance level of 0.05 is 9.4877. 7

b) Explain non-parametric regression and how it differs from linear regression. What is the derivation approach for Nadaraya-Watson kernel regression estimator? 8

7·460192X10^10

A401

0·98772 #

4. a) Describe using matrix formulation how a multivariate linear regression is estimated. 7

b) The following table shows the values of variables x1, x2, and y. The fourth column shows the estimated value of y obtained by regressing y on x1 and x2.

x1	x2	y	Fitted y
1.0	2.0	8.1	8.0
2.0	1.5	8.4	8.5
1.5	2.5	11.3	11.0
2.5	2.0	11.7	11.5
3.0	1.0	9.2	9.0
1.2	1.8	8.3	8.4
2.2	2.2	12.2	12.0
1.8	2.0	10.0	10.2
2.8	1.5	10.8	10.6
2.0	2.5	12.9	12.5

Calculate the values of R-squared and Adjusted R-squared for the regression model. 8

5. a) Describe a stationary time series. Why is the stationarity assumption required in time series modeling? Explain. 7

b) Given the dataset with six observations:
(185, 72), (170, 56), (168, 60), (179, 68), (182, 72), (188, 77),

i) Apply the K-Means clustering algorithm with K=2 clusters for two iterations, using the first two points (185,72) and (170,56) as the initial centroids. Show all calculations, including cluster assignments and updated centroids after each iteration.

ii) Interpret the clustering results and discuss whether the algorithm has converged after two iterations. Explain the concept of convergence in K-Means and its significance 8

6. a) Define autocorrelation and stationarity in time series data. Why are these properties important when fitting ARMA and ARIMA models? 7

b) How can Principal Component Analysis be considered as a method of identifying latent variables? Explain providing a real-word example. 8

7. Write short notes on: (Any two) 5*2 = 10
   a) Empirical Distribution
   b) AIC and BIC in model selection
   c) Permutation Tests for Variable Importance