

## Micro Syllabus

### Subject: Foundation of Data Science II/I (3-1-3)

Marks Allotment	Theory	Practical	Total
Sessional	40		40
Final	60		60
Total	100	50	150
Time Allotment	Lecture	Tutorial	Practical
Weekly Hours			
Depth Codes	C – Circuit	D – Definition	Dm – Demonstration
Dr – Derivation	Dw – Drawing	E – Explanation	I – Illustration
N – Numerical	P – Proof	Pr – Programming	S – State

Unit	Topic/ Sub topic	Depth Code	Description of Depth	Actual Plan			Plan for this Semester		Week
				L	T	P	L	T	
<b>1.</b>	<b>Introduction to Data Science</b>			<b>3</b>		<b>3</b>			<b>1</b>
1.1.	Overview of data science	D, E	<ul style="list-style-type: none"> <li>Definition, History of Data Science</li> </ul>	0.25					
1.2.	Jargons of Data Science	D,E	<ul style="list-style-type: none"> <li>Introduction to Jargon used in Data Science with Definitions; Artificial Intelligence, Machine Learning, Mathematics, Statistics, Data Mining, Feature Engineering</li> </ul>	0.5					
1.3.	Modern Data Ecosystem	D, E, Dw	<ul style="list-style-type: none"> <li>Components of Modern Data Ecosystem along with Key Players</li> </ul>	0.5					
1.4.	Data science lifecycle	D, E, Dw	<ul style="list-style-type: none"> <li>Definition of the data science lifecycle and it's phase</li> </ul>	0.75					

1.5.	Trends, Markets and Applications of Data Science	D, E	<ul style="list-style-type: none"> <li>Market Demands Key Trends in data science landscape</li> <li>Global Market Growth of Data science</li> <li>Examples of Data science applications</li> </ul>	0.5					
1.6.	Tools and technologies in data science	D, E	<ul style="list-style-type: none"> <li>List of tools and technologies used in the data science landscape</li> </ul>	0.25					
1.7.	Data scientist and their roles	D, E	<ul style="list-style-type: none"> <li>Definition. Explain the roles of Data scientist</li> </ul>	0.25					
<b>2.</b>	<b>Math for Data Science</b>			<b>10</b>	<b>5</b>	<b>9</b>			2,3,4
2.1.	Introduction to linear algebra for data science	D,E	<ul style="list-style-type: none"> <li>Definition of linear algebra and it's examples</li> </ul>	0.5					
2.2.	Vectors, matrices, and matrix factorization	D,E	<ul style="list-style-type: none"> <li>Definition, Example,</li> <li>Operations in Linear algebra (addition, subtraction, dot product, matrix multiplication, determinants and Inverse)</li> <li>System of Linear equations and ways to solve them,</li> <li>Example of Matrix Factorization: Singular Value Decomposition, Non negative matrix factorization</li> </ul>	0.5					
2.3.	Gradient descent for optimization	D,E,N	<ul style="list-style-type: none"> <li>Definition, Minimizing a simple quadratic function with gradient descent,</li> <li>Minimizing mean square error using gradient descent</li> </ul> <p><b>Numerical Problem</b></p> <ul style="list-style-type: none"> <li>simple quadratic function</li> </ul>	1	1				
2.4.	Introduction to Probability and Random Variable	D,E,N	<p><b>Introduction to Probability</b></p> <ul style="list-style-type: none"> <li>Basic probability concepts: Events, outcomes, and probability rules.</li> <li>Conditional probability and Bayes' Theorem.</li> </ul>	1	1				

			<b>Random Variables</b> <ul style="list-style-type: none"> <li>Discrete and continuous random variables.</li> <li>Expectation, variance, and standard deviation.</li> </ul> <b>Numerical Problem</b> <ul style="list-style-type: none"> <li>Conditional probability and Bayes' Theorem.</li> </ul>						
2.5.	Probability Distributions: Normal, Bernoulli, Binomial, Poisson	D,E,N	<b>Common Distributions and its properties and applications in Data Science</b> <ul style="list-style-type: none"> <li>Normal Distribution:</li> <li>Bernoulli Distribution:</li> <li>Binomial Distribution:</li> <li>Poisson Distribution:</li> </ul> <b>Numerical Problem</b> <ul style="list-style-type: none"> <li>Simple Numerical calculations excluding complex derivations</li> </ul>	1	1				
2.6.	Descriptive and inferential statistics	D,E	Definition and difference	1					
2.7.	Central limit theorem and Sample distribution concepts	D, Dr, I, E	<ul style="list-style-type: none"> <li>Concept of point estimate, parameter and statistics</li> <li>Definition and example</li> </ul>	1					
2.8.	Normal approximation; Hypothesis Testing Procedures: Tests about the mean of a normal population	D,E,N	<ul style="list-style-type: none"> <li>Normal Approximation: Discrete to Continuous</li> <li>Introduction to Hypothesis Testing</li> <li>Hypothesis Testing Procedure</li> <li>Type I and Type II errors.</li> <li>One sample Z-test and t-test</li> <li>P-values and significance levels.</li> <li><b>Numerical Problem</b> <ul style="list-style-type: none"> <li><b>Two- tailed test</b></li> </ul> </li> </ul>	1	0.5				

2.9.	The t-test, Z-tests for differences between two populations means, The two-sample t-test, Confidence interval for mean of normal population	D,E,N	<ul style="list-style-type: none"> <li>Two sample z-test and t-test</li> <li>Confidence Intervals and its interpretation</li> <li>Numerical on Confidence Intervals</li> </ul>	1	0.5				
2.10.	ANOVA	D,E,N	<ul style="list-style-type: none"> <li>One-way ANOVA: Comparing means across multiple groups.</li> <li>Assumptions and interpretation of ANOVA results.</li> <li>Definition of Two-way ANOVA</li> <li><b>Numerical one-way ANOVA.</b></li> </ul>	2	1				
<b>3.</b>	<b>Data Understanding and Preprocessing</b>			<b>10</b>	<b>3</b>	9			5,6,7
3.1.	Types of data: structured, unstructured, semi-structured	D, I, Dm	<ul style="list-style-type: none"> <li>Structured, unstructured &amp; semi-structured data</li> <li>Quantitative versus qualitative data</li> <li>The four levels of data</li> </ul> <b>Flavors of data (Principles of Data Science, chapter 2 - Packt Publishing)</b>	1.25					
3.2.	Data Preprocessing Requirements	D, Dm	<ul style="list-style-type: none"> <li>Real world data,</li> <li>need of data preprocessing,</li> <li>general steps of data preprocessing</li> <li>tools in data preprocessing</li> </ul>	0.75					
3.3.	Data sources and collection methods	D, I	<ul style="list-style-type: none"> <li>Need of data collection</li> <li>Primary and secondary sources and its characteristics</li> <li>Internal, external, and open-source data</li> <li>Collection methods and steps</li> </ul>	1.5					
3.4.	Data Cleaning and Preparation	D, I	<ul style="list-style-type: none"> <li>Data cleaning, Handling missing values, Handling outliers (z-score, IQR), measure of data quality</li> <li>Benefits and Challenges in Data Cleaning and preparation.</li> </ul>	1					

3.5.	Data Wrangling and Associated Tools	D, Dw	<ul style="list-style-type: none"> <li>• Data Wrangling Process,</li> <li>• Open source and other tools for data wrangling process</li> </ul>	1					
3.6.	Data Enrichment, Validation, and Publishing	D, I, Dm	<ul style="list-style-type: none"> <li>• Data Enrichment and different methods associated with it, Usages in different sectors (financial, social media data, customer, ecommerce, etc.)</li> <li>• Data Validation and different types of validation</li> <li>• Data Publishing and different methods associated with it</li> </ul>	1.5					
3.7.	Data Transformation and Normalization	D, I, N	<ul style="list-style-type: none"> <li>• Data Transformation, Categorical Encoding : One hot, ordinal and label</li> <li>• Normalization and its importance</li> <li>• Type of normalization: min-max, z-score</li> <li>• Numerical: Normalization</li> </ul>	1	0.5				
3.8.	Dimensionality Reduction Linear Factor Model, PCA	D, I, Dw, N	<ul style="list-style-type: none"> <li>• Data Dimensionality, Curse of Dimensionality,</li> <li>• Dimension reduction technique</li> <li>• Linear Factor Model and its relation to PCA</li> <li>• Examples and Steps in finding PCA,</li> <li>• Application and Limitation of PCA</li> </ul> <p><b>Numerical</b></p> <ul style="list-style-type: none"> <li>• <b>PCA, Eigenvalues and Eigenvectors.</b></li> </ul>	2	2.5				
<b>4.</b>	<b>Data Analysis</b>			8	1	9			8,9,10
4.1.	Data Analytics : Descriptive, Diagnostic, Predictive and Prescriptive Analytics	D,E	<ul style="list-style-type: none"> <li>• Types of analytics: Descriptive, Diagnostic, Predictive, Prescriptive.</li> <li>• Examples and applications</li> </ul>	1.5					

4.2.	Exploratory Data Analysis using Descriptive Statistics	D,E,N	<ul style="list-style-type: none"> <li>Measures of central tendency</li> <li>Measures of dispersion</li> <li>Data distribution and histograms.</li> <li>Identifying outliers and missing values.</li> <li>Basic data summaries and visualizations.</li> <li>Correlation analysis and data relationships.</li> </ul>	2					
4.3.	Data Visualization	D,E,D <sub>m</sub>	<ul style="list-style-type: none"> <li>Importance and objectives of visualization.</li> <li>Types of Data Visualization (Explanatory and Exploratory)</li> <li>Infographics and Visualization</li> </ul>	1					
4.4.	Data visualization techniques	Dw,E,I	<ul style="list-style-type: none"> <li>Data Visualization Techniques: Charts, plots, and dashboards.</li> <li>Creating bar plots, histograms, line graphs, scatter plots.</li> <li>Heatmaps, box plots, Time Series plot and pie charts.</li> <li>Multi-dimensional visualizations (e.g., pair plots, 3D plots).</li> <li>Common visualization tools (Matplotlib, Seaborn, Tableau, Power BI, etc.).</li> </ul> <p><b>Numerical:</b></p> <ul style="list-style-type: none"> <li><b>Box plot</b></li> </ul>	1	1				
4.5.	Principles of effective data visualization	D,E,D <sub>m</sub>	<ul style="list-style-type: none"> <li>Clarity, accuracy, consistency and simplicity in visual design.</li> <li>Choosing the right chart for the data. (Relationship, Comparison, Distribution and Composition)</li> <li>Avoiding misleading visualizations.</li> </ul>	1					

			<ul style="list-style-type: none"> <li>Following the Principal of Data Visualization Design (Data Science for Dummies, chapter 9 - Wiley)</li> </ul>						
4.6.	Feature Engineering and other aspects of Data manipulation	D,E,D m	<ul style="list-style-type: none"> <li>Importance of Feature Engineering for Predictive Modeling and Analytics,</li> <li>Feature Engineering: Extraction, Selection and Reduction,</li> <li>Feature Selection Methods,</li> <li>Correlation and Causation</li> <li>Feature Representation in texts and images</li> </ul>	1.5					
<b>5.</b>	<b>Regression and Predictive Modeling</b>			<b>5</b>	<b>3</b>	<b>6</b>			11, 12
5.1.	Empirical Models, Simple Linear Regression, MLE and Least Square Estimator	D,E,D m,Dr, N, Dw	<ul style="list-style-type: none"> <li>Concept and use of empirical models.</li> <li>Regression and predictive analytics.</li> <li>Differences between regression and classification tasks.</li> <li>Simple Linear Regression, Derivation, Assumptions, Application, Interpretation, <b>Numerical</b></li> <li>Concept of Maximum Likelihood Estimation (MLE) for model fitting.</li> <li>Concept of Least Squares Estimator.</li> <li>Comparing MLE and Least Squares in regression.</li> </ul>	1.5	1.5				
5.2.	Multiple Linear Regression, Matrix approach to Multiple Linear Regression, Polynomial Regression Models, Categorical Regressors, Indicator	D,E,D m,Dr, N	<ul style="list-style-type: none"> <li>Understanding multiple predictors.</li> <li>Matrix notation for MLR.</li> <li>Assumptions and diagnostics in MLR.</li> <li>Derivation of regression coefficients using matrix algebra.</li> </ul>	2.25	1.5				

	variables, Selection of variables, and Model Building		<ul style="list-style-type: none"> <li>• Computational advantages of the matrix approach.</li> <li>• Extending linear regression to non-linear relationships.</li> <li>• Fitting polynomial models to data.</li> <li>• Avoiding overfitting in polynomial regression.</li> <li>• Categorical Regressors and Handling categorical variables in regression models.</li> <li>• Use of indicator (dummy) variables.</li> <li>• Feature selection techniques Eg: Forward selection, backward elimination, and stepwise selection.</li> <li>• Model Performance: residuals analysis, MSE, RMSE, MAE, <math>R^2</math>, Adjusted <math>R^2</math>, AIC, BIC</li> <li>• <b>Numerical upto 3 variables</b></li> </ul>						
5.3.	Logistic regression	D,E,D m,N	<ul style="list-style-type: none"> <li>• Differences between linear and logistic regression.</li> <li>• Logistic regression as a classification model.</li> <li>• Sigmoid function and probability interpretation.</li> <li>• Relationship between log-odds and coefficients.</li> <li>• Case studies using regression models in domains such as healthcare, finance, and marketing.</li> </ul>	1.25					
<b>6.</b>	<b>Modeling and validation processes</b>			<b>6</b>	<b>3</b>	<b>9</b>			13, 14
6.1.	Introduction to Machine Learning	D,E	<ul style="list-style-type: none"> <li>• Definition and goals of machine learning.</li> <li>• Rule-based programming vs ML</li> </ul>	0.5					

			<ul style="list-style-type: none"> <li>Real-world applications of machine learning.</li> </ul>							
6.2.	Introduction to Supervised, Unsupervised & Reinforcement Learning.	D,E,D w	<ul style="list-style-type: none"> <li>Definition, Applications and Algorithms for Supervised Learning</li> <li>Details of ANN, Naive Bayes</li> <li>Definition, Applications and Algorithms for Unsupervised Learning</li> <li>Details of K means clustering and density-based clustering</li> <li>Definition and Application of Reinforcement Learning</li> <li>Basic concept: agent, reward, environment and policy</li> </ul>	3						
6.3.	Modeling Process, Training /Validating model, Cross Validation methods, Predicting new observations Interpretation	D,Dr, E,N	<ul style="list-style-type: none"> <li>Steps in the Modeling Process</li> </ul> <b>Training and Validating Models</b> <ul style="list-style-type: none"> <li>Concepts of training, validation, and hyper parameter tuning.</li> <li>Training and validation loss</li> <li>overfitting and underfitting.</li> <li>Bias-variance tradeoff.</li> </ul> <b>Cross-Validation Methods</b> <ul style="list-style-type: none"> <li>Types of cross-validation: K-fold, leave-one-out, stratified.</li> <li>Advantages and limitations of cross-validation.</li> </ul> <b>Predicting New Observations</b> <ul style="list-style-type: none"> <li>Model deployment and inference in unseen data.</li> </ul>	1.5						
6.4.	Measures for Model Performance and Evaluation: Classification accuracy, Confusion matrix, Sensitivity, Specificity, Precision, Recall, F-score,	D,E,N	<b>Classification Performance Measures</b> <ul style="list-style-type: none"> <li>Classification accuracy.</li> <li>Confusion matrix and derived metrics: <ul style="list-style-type: none"> <li>Sensitivity (Recall), Specificity.</li> </ul> </li> </ul>	1	3					

	ROC curve, Clustering performance measures, other measures		<ul style="list-style-type: none"> <li>o Precision, F-score.</li> <li>o ROC curve and AUC</li> </ul> <b>Clustering Performance Measures</b> <ul style="list-style-type: none"> <li>• Internal and External measures: (Silhouette, Rand index)</li> <li>• Other Measures for regression and classification</li> <li>• <b>Numerical</b></li> </ul>						
7.	<b>Ethics and Recent Trends</b>			3					15
7.1.	Ethical considerations in data science	D,E	<ul style="list-style-type: none"> <li>• Introduction to Ethics in Data Science</li> <li>• Key Ethical Principles</li> <li>• Transparency, accountability, and fairness.</li> <li>• Avoiding bias in data and algorithms.</li> <li>• Ethical implications of automation and AI.</li> <li>• Responsible AI Principles</li> </ul>	0.5					
7.2.	Data privacy regulations	E, I	<b>Overview of Data Privacy</b> <ul style="list-style-type: none"> <li>• Importance of protecting user data.</li> <li>• Concepts of data ownership and consent.</li> </ul> <b>Global Data Privacy Regulations</b> <ul style="list-style-type: none"> <li>• General Data Protection Regulation (GDPR)</li> <li>• k-Anonymity, Consent</li> </ul> <b>Implications for Data Science</b> <ul style="list-style-type: none"> <li>• Compliance in data collection and processing.</li> <li>• Strategies for anonymizing and securing data.</li> </ul>	0.5					
7.3.	Responsible data usage	E	<b>Principles of Responsible Data Usage</b>	1					

			<ul style="list-style-type: none"> <li>● Ensuring data accuracy and integrity.</li> <li>● Preventing misuse of data and algorithms.</li> <li>● Social and cultural implications of data-driven decisions.</li> <li>● Diversity in Data, Explainability</li> </ul> <p><b>Ethical AI and Machine Learning</b></p> <ul style="list-style-type: none"> <li>● Building explainable and interpretable models.</li> <li>● Avoiding harmful or discriminatory outcomes.</li> <li>● Tools for fairness evaluation (e.g., IBM AI Fairness 360).</li> </ul>						
7.4.	The Five Cs	D, Dw, E	<b>The Five Cs Framework</b>	0.5					
7.5.	Future Trends	D, E, Dw	<p><b>Emerging Technologies</b></p> <ul style="list-style-type: none"> <li>● Advances in AI and machine learning (e.g., generative AI, federated learning).</li> <li>● Growth of edge computing and IoT data analytics.</li> </ul> <p><b>Sustainability in Data Science</b></p> <ul style="list-style-type: none"> <li>● Reducing the carbon footprint of AI/ML models.</li> <li>● Green data centers and energy-efficient computing.</li> </ul> <p><b>Social Impact of Data Science</b></p> <ul style="list-style-type: none"> <li>● AI for social good: Applications in healthcare, climate, and education.</li> <li>● Challenges in addressing algorithmic bias and inequality.</li> <li>● Skills for Future Data Scientists</li> </ul>	0.5					

			<ul style="list-style-type: none"> <li>● Evolving roles and the importance of continuous learning.</li> <li>● Interdisciplinary approaches integrating ethics, law, and technology.</li> <li>● Case studies on ethical dilemmas in data science (e.g., biased algorithms, privacy breaches).</li> </ul>							
--	--	--	---	--	--	--	--	--	--	--

## Tutorial

The following tutorial activities ( including but not limited to those listed) should be conducted to cover all the required contents of this course.

- Solving data problems using Linear Algebra, vectors and matrices.
- Solving the problems related Probability and Statistics to Understand application in data science.
- Identifying the Data types and performing data cleaning, transformation, wrangling, and dimensionality reduction Including EDA and feature engineering.
- Solving the problem related to Linear and logistic regression.
- Understanding machine learning basics by model training, cross-validation, and performance evaluation.

## Practical

The following lab exercises (including but not limited to those listed) will give students hands-on experience in analyzing and visualizing data, as well as applying data science and machine learning tools to solve real-world problems:

- Get acquainted with data science tools and perform statistical analysis.
- Conduct hypothesis tests (e.g., t-tests, z-tests) on sample datasets to compare population means.
- Simulate and apply the Central Limit Theorem (CLT) to demonstrate how sample distributions converge to a normal distribution.
- Perform data wrangling and ETL processes on a dataset, followed by Exploratory Data Analysis (EDA).
- Utilize tools to create effective data visualizations (e.g., line charts, bar charts, heatmaps, box plots) to derive key insights from the dataset.
- Implement feature extraction and selection techniques, including experimenting with encoding methods like one-hot encoding and creating new features based on domain expertise.
- Develop a simple linear regression model, extend it to multiple linear regression with several variables, and visualize both the regression line and residual plots.
- Apply logistic regression and evaluate the model using metrics such as accuracy, precision, recall, and the ROC curve.
- Apply k-means clustering and assess cluster quality using evaluation metrics like the silhouette score.

By the end of the course, students will be required to submit a project where they develop a prototype to solve a real-world problem.

# **SAMPLE QUESTION**

## SAMPLE FINAL EXAM QUESTION

- Attempt ALL questions.

Q. N	Question	Marks																										
1.	What is Data Science? Elaborate about the data science lifecycle.	1+4 =5																										
2.	Why is math important in Data Science? List the cases where we use linear algebra, statistics and calculus.	2+4=6																										
3.	When do we use hypothesis testing? Thirty students were randomly selected to take the Data Science class. When investigating the average of their grades, the mean was 80, and the variance was 9. Find a 95% confidence interval for the average of their grades.	2+4=6																										
4.	<div>Here is the customer complaint record of a service company for twelve consecutive days, and answer the following questions using this data.</div> <table><tr><td>Day</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td><td>9</td><td>10</td><td>11</td><td>12</td></tr><tr><td>No. of Complaints</td><td>22</td><td>12</td><td>60</td><td>57</td><td>30</td><td>32</td><td>39</td><td>14</td><td>42</td><td>13</td><td>23</td><td>16</td></tr></table> <div><div>i. Draw a box plot with proper labeling of all calculated values</div><div>ii. Normalize Day#9 complaints using both min-max and z-score normalization methods.</div></div>	Day	1	2	3	4	5	6	7	8	9	10	11	12	No. of Complaints	22	12	60	57	30	32	39	14	42	13	23	16	2+4=6
Day	1	2	3	4	5	6	7	8	9	10	11	12																
No. of Complaints	22	12	60	57	30	32	39	14	42	13	23	16																
5.	Explain the use of eigenvalues and eigenvectors in Principal Component Analysis (PCA). For the purpose of justification and calculation, use the following data, where each row represents a customer and each column represents their preference score for different products and the goal is to reduce the dimensionality of the data (from 3 to 2 dimensions only) while retaining as much information as possible.	2+5=7																										

		<b>Customer ID</b>		1	2	3	4	5															
		Customer Preferences	Product A	4	5	6	7	8															
			Product B	5	4	3	6	7															
			Product C	2	3	2	4	5															
6.	What is Data Visualisation? Explain the principles of effective visualization. How is the right visualization chart chosen ?									2+2+2=6													
7.	Compare Linear Regression with the Logistics Regression. Explain how the Logistic regression acts as a classifier with examples.									3+3=6													
8.	<div>Discuss the different approaches for validating a classifier with calculating the accuracy of this Covid case test data. A confusion matrix for covid testing classifier is as follows:</div> <table><tr><td colspan="2" rowspan="2"></td><td colspan="2">Predicted Covid Cases</td></tr><tr><td>True</td><td>False</td></tr><tr><td rowspan="2">Actual Covid Cases</td><td>True</td><td>456</td><td>52</td></tr><tr><td>False</td><td>78</td><td>11569</td></tr></table> <div>Is accuracy sufficient to indicate the performance of this classifier? Justify with calculations and comparison of other parameters like precision, recall and F-1 scores.</div>											Predicted Covid Cases		True	False	Actual Covid Cases	True	456	52	False	78	11569	2+4=6
		Predicted Covid Cases																					
		True	False																				
Actual Covid Cases	True	456	52																				
	False	78	11569																				
9.	<div>Write short notes on the following: -</div> <div><div>i. Data Wrangling</div><div>iii. Responsible Data Usage</div><div>ii. Exploratory Data Analysis</div><div>iv. Coefficient of Determination</div></div>									<div>(4*3)</div> <div>=12</div>													

## SAMPLE FINAL EXAM QUESTION

- Attempt ALL questions.

Q.N	Questions	Mark												
1	Differentiate between Feature Engineering and Feature Selection in the context of Machine Learning. Provide examples to support your explanation.	6												
2	<div>The group of the students are tested for their reading speeds (words per minute).</div> <table><tr><td>Group 1</td><td>200</td><td>210</td><td>190</td></tr><tr><td>Group 2</td><td>200</td><td>230</td><td>225</td></tr><tr><td>Group 3</td><td>210</td><td>215</td><td>205</td></tr></table> <div>Determine if there is a significant difference in reading speeds among the groups using a one-way ANOVA test.</div>	Group 1	200	210	190	Group 2	200	230	225	Group 3	210	215	205	6
Group 1	200	210	190											
Group 2	200	230	225											
Group 3	210	215	205											
3	<div>Describe the differences between discrete and continuous random variables, giving examples of each.</div> <div>What are the key properties of the Poisson distribution and how is it used in data science applications?</div>	3+3												
4	Describe the sigmoid function used in the logistic regression and explain how it is related to the probability of a binary outcome.	3+3												

	Explain the relationship between log-odds and coefficients in a logistic regression model. How do the coefficients influence the predicted probabilities?																			
5	<p>Suppose a company's revenue (Y) depends on its advertising expenditure on the television (X1) and online platforms (X2). The following data is collected.</p> <table border="1"> <thead> <tr> <th>X1(TV in \$1000)</th><th>X2 (Online in \$1000)</th><th>Y (Revenue in \$1000)</th></tr> </thead> <tbody> <tr> <td>10</td><td>20</td><td>50</td></tr> <tr> <td>15</td><td>25</td><td>60</td></tr> <tr> <td>20</td><td>30</td><td>65</td></tr> <tr> <td>25</td><td>35</td><td>70</td></tr> <tr> <td>30</td><td>40</td><td>75</td></tr> </tbody> </table> <p>A. Fit a multiple regression model in the form of  <math>Y = A + BX1 + CX2</math>  B. Predict the revenue (Y) when X1=18 and X2 = 28.</p>	X1(TV in \$1000)	X2 (Online in \$1000)	Y (Revenue in \$1000)	10	20	50	15	25	60	20	30	65	25	35	70	30	40	75	3+3
X1(TV in \$1000)	X2 (Online in \$1000)	Y (Revenue in \$1000)																		
10	20	50																		
15	25	60																		
20	30	65																		
25	35	70																		
30	40	75																		
6	What is the difference between one-hot encoding, ordinal encoding and label encoding for categorical data? Provide examples where each method is most suitable.	3+3																		
7	<p>A binary classifier is used to predict whether a patient has a disease (positive) or not (negative). The confusion matrix for the classifier's prediction is as follows.</p> <table border="1"> <thead> <tr> <th></th><th>Predicted Positive</th><th>Predicted Negative</th></tr> </thead> <tbody> <tr> <td>Actual Positive</td><td>50</td><td>10</td></tr> <tr> <td>Actual Negative</td><td>2</td><td>35</td></tr> </tbody> </table> <p>A. Calculate the classification accuracy.  B. Compute sensitivity (recall) and specificity.  C. Determine precision and F1-score.</p>		Predicted Positive	Predicted Negative	Actual Positive	50	10	Actual Negative	2	35	2+2+2									
	Predicted Positive	Predicted Negative																		
Actual Positive	50	10																		
Actual Negative	2	35																		

8	<p>A dataset has been clustered into 3 clusters. For a particular data point <math>i</math>, the average intra-cluster distance (distance to another point in the same cluster) is <math>a=2.0</math> and average nearest-cluster distance (distance to the closest point in another cluster) is <math>b=4.0</math>.</p> <p>A. Compute the silhouette for this point.  B. Is clustering for this point well-defined or ambiguous?</p>	3+3
9	<p>Write short notes on the following:</p> <p>A. Ethical AI  B. Data Privacy  C. ROC Curve  D. Box plot</p>	3x4 = 12