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**	
1.	**Introduction to Data Science**			**3**		**3**			**1**
1.1.	Overview of data science	D, E	● Definition, History of Data Science	0.25					
1.2.	Jargons of Data Science	D,E	● Introduction to Jargon used in Data Science with Definitions; Artificial Intelligence, Machine Learning, Mathematics, Statistics, Data Mining, Feature Engineering	0.5					
1.3.	Modern Data Ecosystem	D, E, Dw	● Components of Modern Data Ecosystem along with Key Players	0.5					
1.4.	Data science lifecycle	D, E, Dw	● Definition of the data science lifecycle and it's phase	0.75					

1.5.	Trends, Markets and Applications of Data Science	D, E	● Market Demands Key Trends in data science landscape ● Global Market Growth of Data science ● Examples of Data science applications	0.5					
1.6.	Tools and technologies in data science	D, E	● List of tools and technologies used in the data science landscape	0.25					
1.7.	Data scientist and their roles	D, E	● Definition. Explain the roles of Data scientist	0.25					
2.	**Math for Data Science**			**10**	**5**	**9**			2,3,4
2.1.	Introduction to linear algebra for data science	D,E	● Definition of linear algebra and it's examples	0.5					
2.2.	Vectors, matrices, and matrix factorization	D,E	● Definition, Example, ● Operations in Linear algebra (addition, subtraction, dot product, matrix multiplication, determinants and Inverse) ● System of Linear equations and ways to solve them, ● Example of Matrix Factorization: Singular Value Decomposition, Non negative matrix factorization	0.5					
2.3.	Gradient descent for optimization	D,E,N	● Definition, Minimizing a simple quadratic function with gradient descent, ● Minimizing mean square error using gradient descent **Numerical Problem** ● simple quadratic function	1	1				
2.4.	Introduction to Probability and Random Variable	D,E,N	**Introduction to Probability** ● Basic probability concepts: Events, outcomes, and probability rules. ● Conditional probability and Bayes' Theorem.	1	1				

			Random Variables ● Discrete and continuous random variables. ● Expectation, variance, and standard deviation. **Numerical Problem** ● Conditional probability and Bayes' Theorem.						
2.5.	Probability Distributions: Normal, Bernoulli, Binomial, Poisson	D,E,N	**Common Distributions and its properties and applications in Data Science** ● Normal Distribution: ● Bernoulli Distribution: ● Binomial Distribution: ● Poisson Distribution: **Numerical Problem** ● Simple Numerical calculations excluding complex derivations	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	● Concept of point estimate, parameter and statistics ● Definition and example	1					
● 2.8.	● Normal approximation; Hypothesis Testing Procedures: Tests about the mean of a normal population	D,E,N	● Normal Approximation: Discrete to Continuous ● Introduction to Hypothesis Testing ● Hypothesis Testing Procedure ● Type I and Type II errors. ● One sample Z-test and t-test ● P-values and significance levels. ● Numerical Problem ● Two- tailed test	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	● Two sample z-test and t-test ● Confidence Intervals and its interpretation ● Numerical on Confidence Intervals	1	0.5				
2.10.	ANOVA	D,E,N	● One-way ANOVA: Comparing means across multiple groups. ● Assumptions and interpretation of ANOVA results. ● Definition of Two-way ANOVA ● **Numerical one-way ANOVA.**	2	1				
3.	**Data Understanding and Preprocessing**			**10**	**3**	9			5,6,7
3.1.	Types of data: structured, unstructured, semi-structured	D, I, Dm	● Structured, unstructured & semi-structured data ● Quantitative versus qualitative data ● The four levels of data **Flavors of data (Principles of Data Science, chapter 2 - Packt Publishing)**	1.25					
3.2.	Data Preprocessing Requirements	D, Dm	● Real world data, ● need of data preprocessing, ● general steps of data preprocessing ● tools in data preprocessing	0.75					
3.3.	Data sources and collection methods	D, I	● Need of data collection ● Primary and secondary sources and its characteristics ● Internal, external, and open-source data ● Collection methods and steps	1.5					
3.4.	Data Cleaning and Preparation	D, I	● Data cleaning, Handling missing values, Handling outliers (z-score, IQR), measure of data quality ● Benefits and Challenges in Data Cleaning and preparation.	1					

3.5.	Data Wrangling and Associated Tools	D, Dw	• Data Wrangling Process, • Open source and other tools for data wrangling process	1					
3.6.	Data Enrichment, Validation, and Publishing	D, I, Dm	• Data Enrichment and different methods associated with it, Usages in different sectors (financial, social media data, customer, ecommerce, etc.) • Data Validation and different types of validation • Data Publishing and different methods associated with it	1.5					
3.7.	Data Transformation and Normalization	D, I, N	• Data Transformation, Categorical Encoding : One hot, ordinal and label • Normalization and its importance • Type of normalization: min-max, z-score • Numerical: Normalization	1	0.5				
3.8.	Dimensionality Reduction Linear Factor Model, PCA	D, I, Dw, N	• Data Dimensionality, Curse of Dimensionality, • Dimension reduction technique • Linear Factor Model and its relation to PCA • Examples and Steps in finding PCA, • Application and Limitation of PCA **Numerical** • **PCA, Eigenvalues and Eigenvectors.**	2	2.5				
4.	**Data Analysis**			8	1	9			8,9,10
4.1.	Data Analytics : Descriptive, Diagnostic, Predictive and Prescriptive Analytics	D,E	• Types of analytics: Descriptive, Diagnostic, Predictive, Prescriptive. • Examples and applications	1.5					

4.2.	Exploratory Data Analysis using Descriptive Statistics	D,E,N	● Measures of central tendency ● Measures of dispersion ● Data distribution and histograms. ● Identifying outliers and missing values. ● Basic data summaries and visualizations. ● Correlation analysis and data relationships.	2					
4.3.	Data Visualization	D,E,Dm	● Importance and objectives of visualization. ● Types of Data Visualization (Explanatory and Exploratory) ● Infographics and Visualization	1					
4.4.	Data visualization techniques	Dw,E,I	● Data Visualization Techniques: Charts, plots, and dashboards. ● Creating bar plots, histograms, line graphs, scatter plots. ● Heatmaps, box plots, Time Series plot and pie charts. ● Multi-dimensional visualizations (e.g., pair plots, 3D plots). ● Common visualization tools (Matplotlib, Seaborn, Tableau, Power BI, etc.). **Numerical:** ● **Box plot**	1	1				
4.5.	Principles of effective data visualization	D,E,Dm	● Clarity, accuracy, consistency and simplicity in visual design. ● Choosing the right chart for the data. ● (Relationship, Comparison, Distribution and Composition) ● Avoiding misleading visualizations.	1					

			• Following the Principal of Data Visualization Design (Data Science for Dummies, chapter 9 - Wiley)						
4.6.	Feature Engineering and other aspects of Data manipulation	D,E,D m	• Importance of Feature Engineering for Predictive Modeling and Analytics, • Feature Engineering: Extraction, Selection and Reduction, • Feature Selection Methods, • Correlation and Causation • Feature Representation in texts and images	1.5					
5.	**Regression and Predictive Modeling**			**5**	**3**	**6**			11, 12
5.1.	Empirical Models, Simple Linear Regression, MLE and Least Square Estimator	D,E,D m,Dr, N, Dw	• Concept and use of empirical models. • Regression and predictive analytics. • Differences between regression and classification tasks. • Simple Linear Regression, Derivation, Assumptions, Application, Interpretation, **Numerical** • Concept of Maximum Likelihood Estimation (MLE) for model fitting. • Concept of Least Squares Estimator. • Comparing MLE and Least Squares in regression.	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	• Understanding multiple predictors. • Matrix notation for MLR. • Assumptions and diagnostics in MLR. • Derivation of regression coefficients using matrix algebra.	2.25	1.5				

	variables, Selection of variables, and Model Building		• Computational advantages of the matrix approach. • Extending linear regression to non-linear relationships. • Fitting polynomial models to data. • Avoiding overfitting in polynomial regression. • Categorical Regressors and Handling categorical variables in regression models. • Use of indicator (dummy) variables. • Feature selection techniques Eg: Forward selection, backward elimination, and stepwise selection. • Model Performance: residuals analysis, MSE, RMSE, MAE, R^2, Adjusted R^2, AIC, BIC • **Numerical upto 3 variables**						
5.3.	Logistic regression	D,E,Dm,N	• Differences between linear and logistic regression. • Logistic regression as a classification model. • Sigmoid function and probability interpretation. • Relationship between log-odds and coefficients. • Case studies using regression models in domains such as healthcare, finance, and marketing.	1.25					
6.	**Modeling and validation processes**			**6**	**3**	**9**			13, 14
6.1.	Introduction to Machine Learning	D,E	• Definition and goals of machine learning. • Rule-based programming vs ML	0.5					

			● Real-world applications of machine learning.						
6.2.	Introduction to Supervised, Unsupervised & Reinforcement Learning.	D,E,D w	● Definition, Applications and Algorithms for Supervised Learning ● Details of ANN, Naive Bayes ● Definition, Applications and Algorithms for Unsupervised Learning ● Details of K means clustering and density-based clustering ● Definition and Application of Reinforcement Learning ● Basic concept: agent, reward, environment and policy	3					
6.3.	Modeling Process, Training /Validating model, Cross Validation methods, Predicting new observations Interpretation	D,Dr, E,N	● Steps in the Modeling Process **Training and Validating Models** ● Concepts of training, validation, and hyper parameter tuning . ● Training and validation loss ● overfitting and underfitting. ● Bias-variance tradeoff. **Cross-Validation Methods** ● Types of cross-validation: K-fold, leave-one-out, stratified. ● Advantages and limitations of cross-validation. **Predicting New Observations** ● Model deployment and inference in unseen data.	1.5					
6.4.	Measures for Model Performance and Evaluation: Classification accuracy, Confusion matrix, Sensitivity, Specificity, Precision, Recall, F-score,	D,E,N	**Classification Performance Measures** ● Classification accuracy. ● Confusion matrix and derived metrics: o Sensitivity (Recall), Specificity.	1	3				

	ROC curve, Clustering performance measures, other measures		o Precision, F-score. o ROC curve and AUC **Clustering Performance Measures** ● Internal and External measures: (Silhouette, Rand index) ● Other Measures for regression and classification ● **Numerical**						
7.	**Ethics and Recent Trends**			3					15
7.1.	Ethical considerations in data science	D,E	● Introduction to Ethics in Data Science ● Key Ethical Principles ● Transparency, accountability, and fairness. ● Avoiding bias in data and algorithms. ● Ethical implications of automation and AI. ● Responsible AI Principles	0.5					
7.2.	Data privacy regulations	E, I	**Overview of Data Privacy** ● Importance of protecting user data. ● Concepts of data ownership and consent. **Global Data Privacy Regulations** ● General Data Protection Regulation (GDPR) ● k-Anonymity, Consent **Implications for Data Science** ● Compliance in data collection and processing. ● Strategies for anonymizing and securing data.	0.5					
7.3.	Responsible data usage	E	**Principles of Responsible Data Usage**	1					

			● Ensuring data accuracy and integrity. ● Preventing misuse of data and algorithms. ● Social and cultural implications of data-driven decisions. ● Diversity in Data, Explainability **Ethical AI and Machine Learning** ● Building explainable and interpretable models. ● Avoiding harmful or discriminatory outcomes. ● Tools for fairness evaluation (e.g., IBM AI Fairness 360).						
7.4.	The Five Cs	D, Dw, E	**The Five Cs Framework**	0.5					
7.5.	Future Trends	D, E, Dw	**Emerging Technologies** ● Advances in AI and machine learning (e.g., generative AI, federated learning). ● Growth of edge computing and IoT data analytics. **Sustainability in Data Science** ● Reducing the carbon footprint of AI/ML models. ● Green data centers and energy-efficient computing. **Social Impact of Data Science** ● AI for social good: Applications in healthcare, climate, and education. ● Challenges in addressing algorithmic bias and inequality. ● Skills for Future Data Scientists	0.5					

			● Evolving roles and the importance of continuous learning. ● Interdisciplinary approaches integrating ethics, law, and technology. ● Case studies on ethical dilemmas in data science (e.g., biased algorithms, privacy breaches).						

SAMPLE QUESTION

SAMPLE FINAL EXAM QUESTION

- **Attempt ALL questions.**

<table>
<tr><th>Q. N</th><th>Question</th><th>Marks</th></tr>
<tr><td>1.</td><td>What is Data Science? Elaborate about the data science lifecycle.</td><td>1+4 =5</td></tr>
<tr><td>2.</td><td>Why is math important in Data Science? List the cases where we use linear algebra, statistics and calculus.</td><td>2+4=6</td></tr>
<tr><td>3.</td><td>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.</td><td>2+4=6</td></tr>
<tr><td>4.</td><td>Here is the customer complaint record of a service company for twelve consecutive days, and answer the following questions using this data.
<table>
<tr><td>Day</td><td>1</td><td>2</td><td>3</td><td>4</td><td>s</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>
i. Draw a box plot with proper labeling of all calculated values
ii. Normalize Day#9 complaints using both min-max and z-score normalization methods.</td><td>2+4=6</td></tr>
<tr><td>5.</td><td>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.</td><td>2+5=7</td></tr>
</table>

<table>
<tr><td></td><td></td><td>
<table>
<tr><td colspan="2">Customer ID</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td></tr>
<tr><td rowspan="3">Customer Preferences</td><td>Product A</td><td>4</td><td>5</td><td>6</td><td>7</td><td>8</td></tr>
<tr><td>Product B</td><td>5</td><td>4</td><td>3</td><td>6</td><td>7</td></tr>
<tr><td>Product C</td><td>2</td><td>3</td><td>2</td><td>4</td><td>5</td></tr>
</table>
</td><td></td></tr>
<tr><td>6.</td><td>What is Data Visualisation? Explain the principles of effective visualization. How is the right visualization chart chosen ?</td><td>2+2+2=6</td></tr>
<tr><td>7.</td><td>Compare Linear Regression with the Logistics Regression. Explain how the Logistic regression acts as a classifier with examples.</td><td>3+3=6</td></tr>
<tr><td>8.</td><td>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:
<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>
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.</td><td>2+4=6</td></tr>
<tr><td>9.</td><td>Write short notes on the following: -
i. Data Wrangling
ii. Exploratory Data Analysis
iii. Responsible Data Usage</td><td>(3*3)
=12</td></tr>
</table>