

STATISTICAL MACHINE LEARNING FOR DATA SCIENCE LAB																									
Course Code		22CDL72				CIE Marks		50																	
L: T:P:S		0:0:1:0				SEE Marks		50																	
Hrs. / Week		2				Total Marks		100																	
Credits		03				Exam Hours		03																	
Course outcomes: At the end of the course, the student will be able to:																									
22CDL72.1	Analyze datasets using statistical measures and visualizations such as percentiles, IQR, bar plots, and correlation matrices to uncover trends, variability, and outliers.																								
22CDL72.2	Apply different types of features encoding and feature engineering on real datasets.																								
22CDL72.3	Interpret regression models - linear, spline, and Poisson regression to analyze relationships and make predictions.																								
22CDL72.4	Implement classification and dimensionality reduction techniques for imbalanced datasets using logistic regression, Fisher's LDA, and model evaluation metrics.																								
Mapping of Course Outcomes to Program Outcomes and Program Specific Outcomes:																									
	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2											
22CDL72.1	2	2	-	3	2	-	-	-	-	-	-	2	3	3											
22CDL72.2	3	-	2	2	-	-	-	-	-	-	-	2	3	3											
22CDL72.3	2	3	-	2	2	-	-	-	-	-	-	2	3	3											
22CDL72.4	2	2	2	3	2	-	-	-	-	-	-	2	3	3											
Pgm. No.	List of Programs									Hours	Cos														
Prerequisite Programs																									
	<ul style="list-style-type: none"> Basic Python Programming (Lists, Functions, File Handling, Libraries like NumPy and Pandas). Fundamentals of Statistics (Mean, Median, Standard Deviation, Percentiles). Basic Data Visualization using Python (Matplotlib, Seaborn). CSV File Handling and Data Preprocessing Techniques. 										2	NA													
PART-A																									
	Consider the monthly electricity bills (in ₹) of 25 households in a residential neighbourhood: Electricity Bills (₹) = [850, 900, 950, 980, 1000, 1020, 1050, 1075, 1100, 1125, 1150, 1175, 1200, 1250, 1300, 1350, 1400, 1450, 1500, 1600, 1700, 1800, 1900, 2100, 2300]																								
1	Tasks: <ol style="list-style-type: none"> Compute the 25th percentile (Q1) and 75th percentile (Q3) of the electricity bill data. Calculate the Interquartile Range (IQR) using the formula: $IQR = Q3 - Q1$. Discuss how the IQR helps in understanding spending variability among households and how it can indicate outliers such as unusually high consumption. 										2	22CDL72.1													
2	Analyze a dataset containing housing price and attributes like area, number of bedrooms, furnishing status etc. Generate appropriate visualizations to explore the association between various attributes. What can be inferred? Dataset link: https://www.kaggle.com/datasets/yasserh/housing-prices-dataset										2	22CDL72.1													

3	<p>A dataset contains information about smartphones, including their battery capacity (mAh), screen size (inches), and price (in ₹). Use a pair plot or correlation matrix to explore the relationships between these variables.</p> <ul style="list-style-type: none"> • Which variables show the strongest correlations? • What practical insights can you draw from these relationships about smartphone features and pricing? <p>Dataset Link: https://www.kaggle.com/datasets/nishantdeswal1810/smartphones</p>	2	22CDL72.1
4	<p>Estimate the average daily step count of users of a fitness tracking app. The underlying daily step count data is known to be right-skewed (i.e., most users have moderate activity, but a few have very high step counts). Consider 10 different random samples, each consisting of 50 users, and calculate the sample mean step count for each sample.</p> <ul style="list-style-type: none"> • Plot the distribution of these sample mean 	2	22CDL72.1
5	<p>Perform Weight of Evidence (WOE) encoding on various features of titanic dataset. https://www.kaggle.com/c/titanic</p>	2	22CDL72.2
6	<p>Take 2 documents each from sports and politics category and perform TF-IDF and Bag of Words encoding.</p>	2	22CDL72.2
PART-B			
7	<p>Fit a polynomial of degree 1, 2, 3, 4 and 5 for synthetically generated data of 2nd degree polynomial. Illustrate model capacity, overfitting, and underfitting using the results. Identify the model without overfitting or underfitting.</p>	2	22CDL72.2
8	<p>Use SVM classifier with linear and RBF kernel for binary classification using titanic dataset. Compare the results.</p>	2	22CDL72.3
9	<p>Perform regression experiments on housing price dataset. Notice that the relationship between square footage and price is nonlinear. Fit a spline regression model to allow the relationship between square footage and price to change at 2,000 square feet. Explain how spline regression can capture different behaviours of the relationship before and after 2,000 square feet</p>	2	22CDL72.3
10	<p>Perform LASSO and Ridge regression on diabetes dataset, which is built-in in scikit-learn, to illustrate the effect of regularization.</p>	2	22CDL72.3
11	<p>Generate 2-dimensional synthetic data samples (use <code>make_blob</code> function in <code>sklearn.datasets</code>) for a Gaussian Mixture Model with 3 clusters. Run EM algorithm on the dataset to estimate the parameters of Gaussian. Compare the results with actual parameters used to generate the data.</p>	2	22CDL72.4
12	<p>A dataset contains two numerical features and a binary target variable with imbalanced classes (e.g., 90% of Class 0 and 10% of Class 1).</p> <ol style="list-style-type: none"> Compute the covariance matrix of the input features to understand their linear relationship. Apply Fisher's Linear Discriminant Analysis to reduce the dimensionality of the data and observe how well the classes are separated in the projected space. Fit a logistic regression model (a Generalized Linear Model) to classify the target variable using the input features, incorporating a strategy to handle class imbalance. Interpret the model coefficients and compute the corresponding odds ratios for each feature. Evaluate the model using appropriate metrics (precision, recall, F1-score) and 	2	22CDL72.4

	discuss its performance on the minority class.																																		
PART-C Beyond Syllabus Virtual Lab Content (To be done during Lab but not to be included for CIE or SEE)																																			
1. https://www.calculatorsoup.com/calculators/statistics/quartile-calculator.php 2. https://onlinestatbook.com/rvls.html 3. https://statpages.info/																																			
CIE Assessment Pattern (50 Marks - Lab)																																			
<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th colspan="2" style="text-align: center;">RBT Levels</th> <th style="text-align: center;">Test (s)</th> <th style="text-align: center;">Weekly Assessment</th> </tr> <tr> <td></td> <td></td> <td style="text-align: center;">20</td> <td style="text-align: center;">30</td> </tr> </thead> <tbody> <tr> <td style="text-align: center;">L1</td> <td style="text-align: center;">Remember</td> <td style="text-align: center;">-</td> <td style="text-align: center;">-</td> </tr> <tr> <td style="text-align: center;">L2</td> <td style="text-align: center;">Understand</td> <td style="text-align: center;">-</td> <td style="text-align: center;">5</td> </tr> <tr> <td style="text-align: center;">L3</td> <td style="text-align: center;">Apply</td> <td style="text-align: center;">5</td> <td style="text-align: center;">10</td> </tr> <tr> <td style="text-align: center;">L4</td> <td style="text-align: center;">Analyze</td> <td style="text-align: center;">10</td> <td style="text-align: center;">10</td> </tr> <tr> <td style="text-align: center;">L5</td> <td style="text-align: center;">Evaluate</td> <td style="text-align: center;">5</td> <td style="text-align: center;">5</td> </tr> <tr> <td style="text-align: center;">L6</td> <td style="text-align: center;">Create</td> <td></td> <td></td> </tr> </tbody> </table>				RBT Levels		Test (s)	Weekly Assessment			20	30	L1	Remember	-	-	L2	Understand	-	5	L3	Apply	5	10	L4	Analyze	10	10	L5	Evaluate	5	5	L6	Create		
RBT Levels		Test (s)	Weekly Assessment																																
		20	30																																
L1	Remember	-	-																																
L2	Understand	-	5																																
L3	Apply	5	10																																
L4	Analyze	10	10																																
L5	Evaluate	5	5																																
L6	Create																																		
SEE Assessment Pattern (50 Marks - Lab)																																			
<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th colspan="2" style="text-align: center;">RBT Levels</th> <th style="text-align: center;">Exam Marks Distribution (50)</th> </tr> </thead> <tbody> <tr> <td style="text-align: center;">L1</td> <td style="text-align: center;">Remember</td> <td style="text-align: center;">-</td> </tr> <tr> <td style="text-align: center;">L2</td> <td style="text-align: center;">Understand</td> <td style="text-align: center;">10</td> </tr> <tr> <td style="text-align: center;">L3</td> <td style="text-align: center;">Apply</td> <td style="text-align: center;">10</td> </tr> <tr> <td style="text-align: center;">L4</td> <td style="text-align: center;">Analyze</td> <td style="text-align: center;">20</td> </tr> <tr> <td style="text-align: center;">L5</td> <td style="text-align: center;">Evaluate</td> <td style="text-align: center;">10</td> </tr> <tr> <td style="text-align: center;">L6</td> <td style="text-align: center;">Create</td> <td></td> </tr> </tbody> </table>				RBT Levels		Exam Marks Distribution (50)	L1	Remember	-	L2	Understand	10	L3	Apply	10	L4	Analyze	20	L5	Evaluate	10	L6	Create												
RBT Levels		Exam Marks Distribution (50)																																	
L1	Remember	-																																	
L2	Understand	10																																	
L3	Apply	10																																	
L4	Analyze	20																																	
L5	Evaluate	10																																	
L6	Create																																		
* SEE EXAM: Students will be assigned one program from Part A and one program from Part B.																																			
Suggested Learning Resources: <p>Text Books:</p> <ol style="list-style-type: none"> Peter Bruce, Andrew Bruce and Peter Gadeck, "Practical Statistics for Data Scientists", 2nd edition, O'Reilly Publications, 2020 Debasis Samanta: Classic Data Structures, 2nd Edition, PHI, 2009, ISBN-13: 978-8120337312. Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, An Introduction to Statistical Learning: with Applications in R, 2nd Edition, Springer, 2021, ISBN-13: 978-1071614174. Kevin P. Murphy, Machine Learning: A Probabilistic Perspective, MIT Press, 2012, ISBN-13: 978-0262018029 <p>Reference Books:</p> <ol style="list-style-type: none"> Christopher M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006, ISBN-13: 978-0387310732 Hastie, Tibshirani, and Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition, Springer, 2009, ISBN-13: 978-0387848570. Larry Wasserman, All of Statistics: A Concise Course in Statistical Inference, Springer, 2004, ISBN-13: 978-0387402727 <p>Web links and Video Lectures (e-Resources):</p> <ol style="list-style-type: none"> Statistical learning for Reliability Analysis: https://nptel.ac.in/courses/106105239 Engineering Statistics: https://nptel.ac.in/courses/127101233 																																			