COURSE MODULE OF THE SUBJECT TAUGHT FOR THE SESSION 2025-26 (ODD SEM)

Course Syllabi with CO's

Faculty Name:Dr Anitha D B			**Academic Year: 2025 – 2026**				
Department: CSE- Data Science							
Course Code	**Course Title**	**Core / Elective**	**Prerequisite**	**Contact Hours**			**Total Hrs/ Sessions**
				L	**T**	**P**	
BAD702	Statistical Machine Learning For Data Science	Core	Basics of Programming	3	-	2	40L+10P
Course Objectives	CLO1: Understand Exploratory Data Analysis CLO2: Explain Data and Sampling Distributions CLO3: To Analyse Statistical experiments and perform significance testing CLO4: To demonstrate how to perform regression analysis on the data CLO5: Explain Discriminant Analysis on the data.						

Topics Covered as per Syllabus

Module-1

Exploratory Data Analysis: estimates of locations and variability, exploring data distributions, exploring binary and categorical data, exploring two or more variables.

Textbook: Chapter 1

Module-2

Data and Sampling Distributions: Random sampling and bias, selection bias, sampling distribution of statistic, bootstrap, confidence intervals, data distributions: normal, long tailed, student's-t, binomial, Chi-square, F distribution, Poisson and related distributions.

Textbook: Chapter 2

Module-3

Statistical Experiments and Significance Testing: A/B testing, hypothesis testing, resampling, statistical significance & p-values, t-tests, multiple testing, degrees of freedom.

Textbook: Chapter 3

Module-4

Multi-arm bandit algorithm, power and sample size, factor variables in regression, interpreting the regression equation, Regression diagnostics, Polynomial and Spline Regression.

Textbook: Chapter 3 & 4

Module-5

Discriminant Analysis: Covariance Matrix, Fisher's Linear discriminant, Generalized Linear Models, Interpreting the coefficients and odd ratios, Strategies for Imbalanced Data.

Textbook: Chapter 5

Practical Component of IPCC

Sl.No.	**Experiments**
1	A dataset contains the prices of houses in a city. Find the 25th and 75th percentiles and calculate the interquartile range (IQR). How does the IQR help in understanding the price variability?

2	You are given a dataset with categorical variables about customer satisfaction levels (Low, Medium, High) and whether customers made repeat purchases (Yes/No). Create visualizations such as bar plots or stacked bar charts to explore the relationship between satisfaction level and repeat purchases. What can you infer from the data?
3	A dataset contains information about car models, including the engine size (in Liters), fuel efficiency (miles per gallon), and car price. Use a pair plot or correlation matrix to explore the relationships between these variables. Which variables seem to have the strongest relationships, and what might be the practical significance of these findings?
4	You want to estimate the mean salary of software engineers in a country. You take 10 different random samples, each containing 50 engineers, and calculate the sample mean for each. Plot the distribution of these sample means. How does the Central Limit Theorem explain the shape of this sampling distribution, even if the underlying salary distribution is skewed?
5	A researcher conducts an experiment with a sample of 20 participants to determine if a new drug affects heart rate. The sample has a mean heart rate increase of 8 beats per minute and a standard deviation of 2 beats per minute. Perform a hypothesis test using the t-distribution to determine if the mean heart rate increase is significantly different from zero at the 5% significance level.
6	A company is testing two versions of a webpage (A and B) to determine which version leads to more sales. Version A was shown to 1,000 users and resulted in 120 sales. Version B was shown to 1,200 users and resulted in 150 sales. Perform an A/B test to determine if there is a statistically significant difference in the conversion rates between the two versions. Use a 5% significance level.
7	You are comparing the average daily sales between two stores. Store A has a mean daily sales value of $1,000 with a standard deviation of $100 over 30 days, and Store B has a mean daily sales value of $950 with a standard deviation of $120 over 30 days. Conduct a two-sample t-test to determine if there is a significant difference between the average sales of the two stores at the 5% significance level.
8	A company collects data on employees' salaries and records their education level as a categorical variable with three levels: "High School", "Bachelor's", and "Master's". Fit a multiple linear regression model to predict salary using education level (as a factor variable) and years of experience. Interpret the coefficients for the education levels in the regression model.
9	You have data on housing prices and square footage and 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.
10	A hospital is using a Poisson regression model (a type of GLM) to predict the number of emergency room visits per week based on patient age and medical history. The model is given by: $\text{Log}(\lambda) = 2.5 - 0.03*\text{Age} + 0.5*\text{condition}$ where λ is the expected number of visits per week, Age is the patient's age, and condition is a binary variable (1 if the patient has a chronic condition, 0 otherwise). Interpret the coefficients of Age and condition. What is the expected number of visits per week for a 60-year-old patient with a chronic condition? How would the expected number of visits change if the patient did not have a chronic condition?
11	A bakery claims that its new cookie recipe is lower in calories compared to the old recipe, which had a mean calorie count of 200. You sample 40 new cookies and find a mean of 190 calories with a standard deviation of 15 calories. Perform a one-tailed t-test to determine if the new recipe has significantly fewer calories at a 5% significance level.

List of Textbooks

1. Peter Bruce, Andrew Bruce and Peter Gadeck, “Practical Statistics for Data Scientists”, 2nd edition, O’Reilly Publications, 2020.

Web links and Video Lectures (e-Resources)

- Statistical learning for Reliability Analysis: https://nptel.ac.in/courses/106105239
- Engineering Statistics: https://nptel.ac.in/courses/127101233

Activity Based Learning (Suggested Activities in Class)/ Practical Based Learning

Course (mini) project to demonstrate the concepts (10 marks)

Course Outcomes	**CO1:** Analyse data sets using techniques to estimate variability, exploring distributions, and investigating relationships between variables. **CO2:** Apply random sampling, confidence intervals, and recognize various data distributions on datasets. **CO3:** Perform significance testing and identify statistical significance. **CO4:** Apply regression analysis for prediction, interpret regression equations, and assess regression diagnostics. **CO5:** Perform discriminant analysis on the varieties of datasets

Internal Assessment Marks:

CIE marks for the **theory component** are 25 marks and that for the **practical component** is 25 marks.

- 25 marks for the theory component are split into 15 marks for Internal Assessment Tests and 10 marks for other assessment methods.
- 25 marks for the practical component are split into 15 marks for the conduction of the experiment and preparation of laboratory record, and 10 marks for the test to be conducted after the completion of all the laboratory sessions.
- The laboratory test (duration 02/03 hours) after completion of all the experiments shall be conducted for 50 marks and scaled down to 10 marks.

The Correlation of Course Outcomes (CO's) and Program Outcomes (PO's)

Subject Code	**BDS613B**	**TITLE:** Exploratory Data Analysis						**Faculty Name**	Dr Anitha D B			
List of Course Outcomes	**Program Outcomes**											**Total**
	PO 1	**PO 2**	**PO 3**	**PO 4**	**PO 5**	**PO 6**	**PO 7**	**PO 8**	**PO 9**	**PO 10**	**PO 11**	**PO 12**	
CO-1	2	2	-	-	2	-	-	-	2	2	-	-	10
CO-2	2	-	-	-	2	-	-	-	2	2	-	-	8
CO-3	2	2	2	-	2	-	-	-	2	2	-	-	12
CO-4	2	2	2	-	2	-	-	-	2	2	-	-	12
CO-5	2	2	-	-	2	-	-	-	2	2	-	-	10
Total	**10**	**08**	**04**	-	**10**	-	-	-	10	10	-	-	**52**

Note: 3 = Strong Contribution, 2 = Average Contribution, 1 = Weak Contribution, - = No Contribution

The Correlation of Course Outcomes (CO's) and Program Specific Outcomes (PSO's)

Subject Code	**BAD702**	**TITLE:** Statistical Machine Learning For Data Science	**Faculty Name**	Dr Anitha D B
List of Course Outcomes	**Program Specific Outcomes**			**Total**
	PSO-1	**PSO-2**	**PSO-3**	
CO-1	2	-	-	2
CO-2	2	2	-	4
CO-3	2	-	-	2
CO-4	2	2	-	4
CO-5	2	-	-	2
Total	**10**	**4**	-	**14**

www.kaggle