BIRLA INSTITUTE OF TECHNOLOGY, MESRA, RANCHI
(END SEMESTER EXAMINATION)

CLASS: IMSC **SEMESTER: VIII**
BRANCH: CQEDS **SESSION: SP/2025**

SUBJECT: ED415 CATEGORICAL DATA ANALYSIS AND STATISTICS IN BAYESIAN PARADIGM

TIME: 3 Hours **FULL MARKS: 50**

INSTRUCTIONS:
1. The question paper contains 5 questions each of 10 marks and total 50 marks.
2. Attempt all questions.
3. The missing data, if any, may be assumed suitably.
***4*. Before attempting the question paper, be sure that you have got the correct question paper.**
5. Tables/Data hand book/Graph paper etc. to be supplied to the candidates in the examination hall.

			CO	BL
Q.1(a)	When the 2010 General Social Survey asked subjects in the US whether they would be willing to accept cuts in their standard of living to protect the environment, 486 of 1374 subjects said yes. a) Estimate the population proportion of those who would say yes. Construct and interpret a 99% and 95% confidence interval for this proportion. b) Comment on the confidence intervals' behaviour when the proportion estimate is near 0 or 1. Is it possible to construct the improved confidence interval?	[6]	1	4
Q.1(b)	Identify the natural response variable and the explanatory variables with proper justification, and identify the most appropriate scale for both the response variable and the explanatory variables, nominal or ordinal. a) Attitude toward gun control (favour, oppose), gender (female, male), mother's education (high school, college). b) Heart disease (yes, no), blood pressure, and cholesterol level. c) Race (white, nonwhite), religion (Catholic, Jewish, Muslim, Protestant, none), vote for president (Democrat, Republican, Green), and annual income. d) The number of runs scored by a batsman in an IPL match, the following variables were recorded: Runs scored in the match, Batting position (e.g., opener, middle order, lower order), Opposition team, Match location (home, away, neutral), Number of balls faced, Bowler type most faced (fast, spin), Toss result (won, lost).	[4]	1	3
Q.2	A sports analyst is studying the effectiveness of a new technique to predict whether an IPL batsman will score more than 50 runs in a match. The predictions made by the model and the actual outcomes are shown in the contingency table below:			

	Predicted: >50 runs	Predicted: ≤50 runs	Total
Actually >50 runs	60	20	80
Actually ≤50 runs	30	90	120
Total	90	110	200

Answer the following:

	a) What is the probability that the model correctly predicts a batsman scoring more than 50 runs? *(1 mark)* b) Compute the sensitivity and specificity of the model. *(2 marks)* c) Find the positive predictive value (PPV) and negative predictive value (NPV). *(2 marks)* d) What are the odds that a batsman scores more than 50 runs given that the model predicts it? *(1 mark)* e) Calculate the odds ratio for the prediction and actual performance association. *(2 marks)* f) If a batsman is predicted to score ≤50 runs, what is the conditional probability that he scores ≤50 runs? *(1 mark)* g) Interpret the odds ratio obtained in part (e) in the context of prediction accuracy. *(1 mark)*	[10]	2	4

PTO

Q.3	a) Explain why we need Generalised Linear Models instead of traditional linear regression. Describe at least two limitations of linear regression that GLMs address. Give one real-world example where a GLM is more appropriate than linear regression, clearly describing the response variable and how the GLM handles it. (4 marks) b) State the key components of a Generalised Linear Model. (2 marks) c) When dealing with categorical response variables, explain: What type of GLM suits nominal and ordinal response data? (4 marks)	[10]	3	4
Q.4	A researcher is analysing the lifetimes (in hours) of a certain type of LED light bulb. It is assumed that the lifetime XX of such a bulb follows an Exponential distribution with rate parameter $\lambda > 0$ $$f(x \mid \lambda) = \lambda e^{-\lambda}, \quad x > 0$$ Suppose the lifetimes of n=5 randomly chosen LED bulbs were observed as: x = 1050, 970, 1120, 880, 1010 a) Assume the prior for λ is a Gamma distribution, that is $\lambda \sim$ Gamma(α, β),with density $\pi(\lambda)$ Take $\alpha = 2,\ \beta = 1000,$ find the posterior distribution and obtain the Bayes estimator of λ under squared error loss.(4 marks) b) Derive the Jeffreys prior for λ, find the corresponding posterior, and obtain the Bayes estimator under squared error loss. (2 marks) c) Using the given data x, calculate the numerical values of: (4 marks) i. The MLE of λ. ii. The Bayes estimator of λ under the conjugate prior. iii. The Bayes estimator of λ using the Jeffreys prior.	[10]	4	4
Q.5(a)	A manufacturing company claims that its new machine produces defect-free items with a probability of at least 0.95. To assess this claim, a quality control team inspects a random sample of 20 items produced by the machine and finds two defective items. (a) Define the null and alternative hypotheses regarding the defect-free probability p.(1 mark) (b) Assuming the number of defect-free items follows a Binomial distribution and using a Beta (1,1) prior for p under the alternative hypothesis, compute the Bayes Factor comparing the defined null and alternative hypotheses. (5 marks) (c) Interpret the Bayes factor and decide on the manufacturer's claim. (2 marks)	[8]	5	4
Q.5(b)	How the Bayesian approach is different from the frequentist approach.	[2]	5	3

::::::26/04/2025::::::E