U.S.N.										

B.M.S. College of Engineering, Bengaluru-560019

Autonomous Institute Affiliated to VTU

February 2025 Semester End Main Examinations

Programme: B.E. **Semester: IV**

Branch: Artificial Intelligence and Machine Learning **Duration: 3 hrs.**

Course Code: 22AM4PCPSM **Max Marks: 100**

Course: Probability and Statistics for Machine Learning

Instructions: 1. Answer any FIVE full questions, choosing one full question from each unit.
2. Missing data, if any, may be suitably assumed.

Important Note: Completing your answers, compulsorily draw diagonal cross lines on the remaining blank pages. Revealing of identification, appeal to evaluator will be treated as malpractice.

		UNIT - I	*CO*	*PO*	Marks
1	a)	In a community, 70% of the people use social media platform X, 50% use social media platform Y, and 30% use both platforms. i) What percentage of people use either platform X or platform Y or both? ii) What percentage of people use neither platform X nor platform Y?	*CO1*	*PO1*	**06**
	b)	At a college, the probability that a student enrolls in a science class is 0.4, and the probability that a student enrolls in a history class is 0.6. The probability that a student enrolls in a science class given that they are enrolled in a history class is 0.3. i) What is the probability that a student enrolls in both science and history classes? ii) What is the probability that a student enrolls in either science or history classes? iii) Is enrollment in science and history classes independent? iv) Is enrollment in science and history classes mutually exclusive?	*CO1*	*PO1*	**08**
	c)	Define sample space and event. Prove that sample space of N possible outcomes yields 2^N possible events with an example.	*CO1*	*PO2*	**06**
		OR			
2	a)	A computer program is tested by 3 independent tests. When there is an error, these tests will discover it with probabilities 0.2, 0.3, and 0.5, respectively. Suppose that the program contains an error. What is the probability that it will be found by at least one test?	*CO1*	*PO2*	**06**
	b)	Discuss the various definitions of probability, including the Classical, Frequentist and Axiomatic approaches. Highlight the merits and demerits of each definition with suitable examples.	CO2	*PO1*	**08**
	c)	In random arrangement of the letters of the word 'ENGINEERING', what is the probability that vowels always occur together.	CO2	*PO2*	**06**

			UNIT - II			
	3	a)	In a factory, the probability that a machine produces a defective item is 0.1. If 10 items are produced by the machine, what is the probability that: i) Exactly 2 items are defective? ii) At most 2 items are defective? iii) More than 2 items are defective?	CO*1*	*PO2*	**06**
		b)	Time between arrival of customers at a coffee shop is to be modelled, and the average time between arrivals of customers is 8 minutes ($\lambda = 1/8$). i) Calculate the probability that the next customer will arrive within 4 minutes of the previous one? ii) Compute the probability that the next customers will arrive in the interval of 4 to 10 minutes?	CO*1*	*PO2*	**08**
		c)	For the Bernoulli distribution derive mean, variance and establish relationship between them.	CO*1*	*PO2*	**06**
			OR			
	4	a)	A program consists of two modules. The number of errors, X1, in the 1st module and the number of errors, X2, in the 2nd module have the joint distributions P (0,0) =P (0,1) =P (1,0) = 0.2, P (1,1) =P (1,2) =P (1,3) = 0.1, P (0,2) =P (0,3) = 0.05. The maximum number errors in X1 are 1 and X2 is 3 Find: i) Marginal Distributions of total number of errors. ii) The Probability of number of errors in the first module. iii) The distribution of the total number of errors in the program.	CO2	*PO4*	**06**
		b)	Obtain the mean and variance of Exponential distribution.	*CO1*	*PO2*	**08**
		c)	The score on a quiz is uniformly distributed between 40 and 80.A student is randomly selected and is known to have scored above 60. What is the probability that the student scored between 65 and 75?	CO1	*PO2*	**06**
			UNIT - III			
	5	a)	Prove unbiasedness, consistency and asymptotic normality of sample mean.	CO*2*	*PO3*	**08**
		b)	Given Data 8,12,15,18,19,22,25,30,31,150 The Mean and Standard Deviation of the dataset are 33 and 41.76, respectively. i) Identify the quartiles. ii) Identify outliers using the 1.5 * IQR rule. iii) Delete the detected outliers and identify the mean, quartiles, and standard deviation again. iv) Make a conclusion about the effect of outliers on basic descriptive statistics.	CO*2*	*PO3*	**08**
		c)	Illustrate the relationship between Mean, Median and Mode.	CO*2*	*PO2*	**04**
			OR			
	6	a)	A researcher analyses test scores of 200 students. She calculates the mean, median, and standard deviation and creates histograms.	CO3	*PO3*	**06**

		She then uses the data to predict overall student performance and tests score differences between two departments. i. Identify the descriptive statistics used. ii. Identify the inferential statistics used. iii. Explain the difference between descriptive and inferential statistics using this example.			
	b)	An incomplete frequency distribution is given as follows: <table><tr><td>x</td><td>10-20</td><td>20-30</td><td>30-40</td><td>40-50</td><td>50-60</td><td>60-70</td><td>70-80</td><td>Total</td></tr><tr><td>f</td><td>12</td><td>30</td><td>?</td><td>65</td><td>?</td><td>25</td><td>19</td><td>230</td></tr></table> Calculate the missing values by considering mean to be 70.	CO3	*PO2*	**10**
	c)	Explain why standard deviation is used to measure the dispersion of data, even though variance is also available as a measure.	CO1	*PO1*	**04**
		UNIT - IV			
7	a)	The lifetime of a certain type of electronic component follows a Gamma distribution with unknown shape parameter α and scale parameter β. Assume that five components are tested independently, and their lifetimes are recorded as follows: 75, 85, 90, 80, and 95 hours. i) Calculate the moment estimators for α and β based on the given data. ii) Solve α and β using the measured lifetimes.	CO2	*PO3*	**10**
	b)	The weights of a particular type of package are normally distributed with a known population standard deviation of 4.2 pounds. A random sample of 50 packages has an average weight of 25.3 pounds. Estimate the population mean with: i) 85% confidence. ii) 95% confidence. iii) 99% confidence. Also conclude what effect an increase in confidence level has on the critical value, margin of error, and width of the confidence interval?	CO2	*PO3*	**10**
		OR			
8	a)	A study was conducted to assess the preference for two different online learning platforms X and Y. 300 random users were surveyed about their preference for Platform X, and 165 users reported a preference for Platform X. Another 300 random users were surveyed about their preference for Platform Y, and 180 users reported a preference for Platform Y. At 1% level of significance, determine if there is a difference in the proportion of users who prefer the two platforms.	CO2	*PO3*	**10**
	b)	Derive the maximum likelihood estimator of a geometric distribution function $f(x,p) = p(1-p)^{x-1}$.	CO2	*PO3*	**10**

			UNIT - V			
	9	a)	Given a logistic regression model logit(p) = $b_{0+}b_1$ * study hours. The threshold for classification is 0.5. Based on the following study hours, classify each student as "Pass" (1) or "Fail" (0): Student name: A, B, C, D, E Study Hours: 10, 15, 20, 25, 30	CO3	PO3	**10**
		b)	Given the following data: Observation 1: Attribute 1 = 4, Attribute 2 = 5 Observation 2: Attribute 1 = 3, Attribute 2 = 6 Observation 3: Attribute 1 = 2, Attribute 2 = 7 Observation 4: Attribute 1 = 1, Attribute 2 = 8 With the help of PCA, reduce its dimensionality.	CO3	PO3	**10**
			OR			
	10	a)	At a tire company, 200 customers were asked to record the tire pressure of their vehicles and the number of years they have used the tires. The results are summarized in the table: Tire Pressure (psi): Sample Mean = 35, Standard Deviation = 5 Years Used: Sample Mean = 6, Standard Deviation = 2 The sample correlation coefficient between tire pressure and years of use is $r = -0.25$. i) Compute the least squares regression line which describes how the number of miles per gallon depends on the mileage. What does the obtained slope and intercept mean in this situation? ii) Use R^2 to evaluate its goodness of fit. Is this a good model?	CO3	PO3	**10**
		b)	The following statistics were obtained from a sample of size n=60 • The predictor variable X has a mean of 40 and a variance of 10; • The response variable Y has a mean of 15 and a variance of 5; • The sample covariance between X and Y is 7. i) Estimate the linear regression equation for predicting Y based on X. ii) Compute the ANOVA table.	CO3	PO3	**10**

****