U.S.N.										

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

Autonomous Institute Affiliated to VTU

June 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)		Explain the concepts of an event and sample space in probability, and illustrate with an example.	*CO2*	*PO1*	**4**
	b)		If two dice are thrown, what is the probability that the sum is i. Greater than 8 ii. Neither 7 nor 11.	*CO2*	*PO1*	**6**
	c)		A bag contains 10 gold and 8 silver coins. Two successive drawings of 4 coins from the bag, are made such that: i. coins are replaced before the second trial ii. coins are not replaced before the second trial Find the probability that the first drawing will give 4 gold and the second drawing will give 4 silver coins.	*CO2*	*PO1*	**10**
			OR			
2	a)		Justify the following with suitable examples: i. Can events be mutually exclusive and exhaustive. ii. Can events be mutually exclusive and independent.	*CO2*	*PO3*	**06**
	b)		How do you handle uncertainty in machine learning models using probability?	*CO1*	*PO1*	**06**
	c)		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*	**08**
			UNIT - II			
3	a)		A random variable X has the following probability distribution function. **x** | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 **p(x)** | 0 | k | $2k$ | $2k$ | $3k$ | k^2 | $2k^2$ | $7k^2+k$ i. Find k. ii. Evaluate P ($X \geq 6$), P ($X < 6$) and P ($0 < X < 5$) iii. Compute E ($3X + 2$)	*CO1*	*PO1*	**10**
	b)		Calculate the Mean and Variance of Exponential distribution with parameter θ and probability density function (pdf) given as	*CO1*	*PO1*	**10**

			$f(x) = \begin{cases} \theta e^{-\theta x} & , x \geq 0 \\ 0, & otherwise \end{cases}$.			
			OR			
	4	a)	State the Central Limit Theorem and elaborate on its applications for the large sample case.	*CO2*	*PO3*	**04**
		b)	What are the characteristics of a normal distribution, and how can it be applied to real-world scenarios? Given that the mean height of soldiers follows a normal distribution with a mean of 68.22 inches and a variance of 10.8 in², how many soldiers in a regiment of 1,000 would you expect to be taller than 6 feet?	*CO3*	*PO2*	**10**
		c)	The joint probability distribution of X and Y is given by Y/X | -1 | 0 -1 | 1/6 | 1/3 0 | 0 | 1/3 1 | 1/6 | 0 Compute the marginal distribution of X and also obtain its variance.	*CO1*	*PO2*	**06**
			UNIT - III			
	5	a)	Calculate Q_1, Q_3, D_6, P_{85} for the following data: x: 10, 11, 12, 13, 14, 15, 16, 17, 18 f: 3, 4, 5, 12, 10, 7, 5, 2, 1	*CO2*	*PO1*	**08**
		b)	Calculate the Median, Variance and Standard Deviation of the patients with the blood sugar levels given below: 50, 51, 62, 62, 70, 72, 72, 54, 62	*CO2*	*PO1*	**05**
		c)	In a study a random sample of 500 pineapples was taken from a large consignment, The average weight of the sampled pineapples was found to be 1.2 kg with a standard deviation of 0.3 kg. i. Identify population and the sample in the context of this study. Explain what each term represents in this scenario. ii. Calculate the standard error (S.E.) of the mean weight of pineapples in the sample.	*CO2*	*PO1*	**07**
			OR			
	6	a)	Explain the procedure to identify outliers using the Interquartile Range (IQR) method.	*CO2*	*PO1*	**05**
		b)	Define a parameter and a statistic in descriptive statistics. Explain their differences. Based on the scenario below, identify the population, sample, parameter, and statistic: Imagine you are a data analyst studying the average income of households in Bengaluru. You randomly select 200 households to collect income data.	*CO2*	*PO1*	**07**
		c)	What is the limitation of sample mean? With an example and necessary diagram, analyze whether it is reliable or not. What is the alternate measure that can be used for estimation?	*CO2*	*PO4*	**08**
			UNIT - IV			
	7	a)	Find the maximum likelihood estimate (MLE) for the parameter λ of Poisson distribution based on a sample of size n.	*CO1*	*PO1*	**06**
		b)	A geneticist is interested in the average Hemoglobin level of males with a certain blood disorder. In a random sample of 100 males, the mean	*CO1*	*PO1*	**05**

			Hemoglobin level is found to be 12.5 g/dL with a standard deviation of 1.8 g/dL. Compute the 95% confidence interval for the Mean Hemoglobin level of males with this blood disorder.			
		c)	Time magazine conducted a telephone poll of 800 adult Americans, asking 'Do you believe stricter regulations on data privacy are necessary to ensure ethical AI development?" The results of the survey are as given below <table><tr><th></th><th>AI engineers</th><th>Others</th></tr><tr><td>Sample</td><td>605</td><td>195</td></tr><tr><td>Responded yes</td><td>351</td><td>41</td></tr></table> Test whether there is significant difference between the opinions of AI engineers and others?	*CO1*	*PO1*	**09**
			OR			
	8	a)	Elaborate on the general steps involved in using the method of moments to obtain parameter estimates.	*CO2*	*PO1*	**06**
		b)	Differentiate the following: i. Acceptance and Rejection region. ii. Probabilities of Type I and Type II error.	*CO2*	*PO1*	**06**
		c)	Two groups A and B consists of 100 people each who have a disease. A serum is given to group A but not to group B. It is found that in groups A and B, 75 and 65 people, respectively, recover from the disease. Test the hypothesis that the serum helps to cure the disease.	*CO2*	*PO1*	**08**
			UNIT - V			
	9	a)	Given Rocket Propellant Data, it is suspected that shear strength is related to the age (in weeks) of the batch of sustainer propellant. Six observations on shear strength and the age of the corresponding batch of propellant have been collected and are as given below in the form of table: <table><tr><th>Observation</th><th>Shear Strength, y_i, (psi)</th><th>Age of Propellant, x_i (weeks)</th></tr><tr><td>1</td><td>2158.70</td><td>15.50</td></tr><tr><td>2</td><td>1678.15</td><td>23.75</td></tr><tr><td>3</td><td>2316.00</td><td>8.00</td></tr><tr><td>4</td><td>2061.30</td><td>17.00</td></tr><tr><td>5</td><td>2207.50</td><td>5.50</td></tr><tr><td>6</td><td>1708.30</td><td>19.00</td></tr></table> Estimate the least square estimates $\hat{\beta}_0$(intercept parameter) and $\hat{\beta}_1$(slope parameter) and fit simple linear regression line $y = \hat{\beta}_0 + \hat{\beta}_1 x$.	*CO3*	*PO2*	**10**
		b)	Discuss the role of dimension reduction in data analysis. Identify and briefly describe three common techniques used for this purpose, along with their typical applications.	*CO3*	*PO1*	**05**
		c)	Following is the data given about students' marks. Classify whether student will be admitted to a university or not based on their exam scores. Consider the threshold as 0.5 to classify the students as admitted (1) or not admitted (0) by considering the predicted probabilities. Assume the estimated coefficients are $b_0 = -6.0$ and $b_1 = 0.15$. <table><tr><th>Student name</th><th>Marks scored</th></tr><tr><td>A</td><td>34</td></tr><tr><td>B</td><td>89</td></tr><tr><td>C</td><td>67</td></tr></table>	*CO3*	*PO2*	**05**

			OR			
	10	a)	For the dataset with house price being a function of house area and plot size. **House Area(ft^2)** | **Plot Size (ft^2)** | **House Price(US$)** 1474 | 10000 | 150000 1560 | 9500 | 145000 1584 | 12000 | 160000 1645 | 13000 | 170000 1712 | 13500 | 190000 i. Employ the following linear regression model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ Where, x_1 and x_2 denote the variables of house area and plot size, respectively. ii. Predict the value of house with area of 1,560 ft^2 and plot size of 13,750 ft^2.	*CO3*	*PO2*	**10**
		b)	Describe the R^2 value in regression analysis, including its meaning, interpretation, and role in evaluating model fit. Also, discuss any limitations associated with R^2.	*CO3*	*PO2*	**05**
		c)	The purity of oxygen produced by a fractional distillation process is thought to be related to the percentage of hydrocarbons in the main condenser of the processing unit. Five samples are shown below. (Consider $\beta_0 = 34.33$ and $\beta_1 = 10.98$). Sample | Percentage of Hydrocarbons (x) | Purity of Oxygen (y) 1 | 2.0 | 90.5 2 | 3.0 | 88.2 3 | 4.0 | 85.0 4 | 5.0 | 82.1 5 | 6.0 | 79.8 Calculate the 95% confidence interval for the slope of the regression line that predicts the purity of oxygen based on the percentage of hydrocarbons.	*CO3*	*PO2*	**05**

****