U.S.N.										

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

Autonomous Institute Affiliated to VTU

January / February 2025 Semester End Main Examinations

Programme: B.E. | **Semester: V**
Branch: Artificial Intelligence and Machine Learning | **Duration: 3 hrs.**
Course Code: 24AM5PCSML | **Max Marks: 100**
Course: STATISTICAL MODELING

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)	Define Simple Linear Regression and explain how it models relationships between variables, including a brief example.	CO1	PO1	**06**
	b)	Derive the least squares estimates for the parameters of a simple linear regression model.	CO1	PO2	**07**
	c)	Explain the steps involved in testing the hypothesis for the significance of slope parameter in a simple linear regression model.	CO2	PO2	**07**
		OR			
2	a)	What is the coefficient of determination (R^2)? How is it calculated, and what does it signify in a regression model?	CO1	PO2	**06**
	b)	The data regarding sales and advertisement expenditure of a firm is as follows:	CO1	PO2	**07**

Measure	**Sales** (in crores)	**Advertisement expenditure** (in crores)
Means	40	6
Standard deviations	10	1.5
Correlation coefficient	0.9	

If the firm targets sales of 60 crores, what should be the required advertisement expenditure? Use linear regression to calculate and explain.

	c)	Perform ANOVA for a simple linear regression model. Explain the decomposition of the total sum of squares and its relation to the regression and error sum of squares.	CO2	PO2	**07**

<table>
<tr><td></td><td></td><td></td><td>UNIT - II</td><td></td><td></td><td></td></tr>
<tr><td></td><td>3</td><td>a)</td><td>List and explain the assumptions of the Multiple Linear Regression model. Why are these assumptions important?</td><td>CO1</td><td>PO1</td><td>06</td></tr>
<tr><td></td><td></td><td>b)</td><td>Derive the least squares estimates for the parameters of a Multiple Linear Regression model and obtain the same for the following:
<table>
<tr><th>Execution Time in milliseconds (y)</th><th>Number of Elements (X1)</th><th>Input Complexity (X2)</th></tr>
<tr><td>78.5</td><td>7</td><td>26</td></tr>
<tr><td>74.3</td><td>1</td><td>29</td></tr>
<tr><td>104.3</td><td>11</td><td>56</td></tr>
<tr><td>87.6</td><td>11</td><td>31</td></tr>
<tr><td>95.9</td><td>7</td><td>52</td></tr>
<tr><td>109.2</td><td>11</td><td>55</td></tr>
</table></td><td>CO1</td><td>PO2</td><td>10</td></tr>
<tr><td></td><td></td><td>c)</td><td>State Gauss-Markov theorem.</td><td>CO1</td><td>PO1</td><td>04</td></tr>
<tr><td></td><td></td><td></td><td>OR</td><td></td><td></td><td></td></tr>
<tr><td></td><td>4</td><td>a)</td><td>What is multicollinearity in a regression model? Explain how the Variance Inflation Factor (VIF) is used to detect multicollinearity. Suggest remedies to address multicollinearity.</td><td>CO2</td><td>PO2</td><td>07</td></tr>
<tr><td></td><td></td><td>b)</td><td>Define heteroscedasticity in regression analysis. How can it be detected using residual plots? Suggest ways to address heteroscedasticity.</td><td>CO2</td><td>PO2</td><td>07</td></tr>
<tr><td></td><td></td><td>c)</td><td>Given the following dataset:
<table>
<tr><td>Observation</td><td>1</td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td></tr>
<tr><td>Rainfall (cms)</td><td>30</td><td>23</td><td>34</td><td>31</td><td>17</td><td>36</td></tr>
<tr><td>Yield (tons)</td><td>65</td><td>62</td><td>70</td><td>64</td><td>52</td><td>68</td></tr>
</table>
Calculate the Durbin-Watson d statistic to test positive autocorrelation and conclude. ($d_L = 0.61$ and $d_U = 1.40$)</td><td>CO2</td><td>PO2</td><td>06</td></tr>
<tr><td></td><td></td><td></td><td>UNIT - III</td><td></td><td></td><td></td></tr>
<tr><td></td><td>5</td><td>a)</td><td>Explain the importance of model diagnostics in regression analysis. How are added variable plots used to diagnose model issues?</td><td>CO2</td><td>PO1</td><td>06</td></tr>
<tr><td></td><td></td><td>b)</td><td>What are Hat Matrix Leverage values? How are they used to identify leverage data points in regression analysis?</td><td>CO2</td><td>PO2</td><td>07</td></tr>
<tr><td></td><td></td><td>c)</td><td>The DFBETAS matrix for a regression model with n=10 observations and 2 predictors are given below:</td><td>CO1</td><td>PO3</td><td>07</td></tr>
</table>

			$\begin{bmatrix} 0.02 & -0.15 \\ -0.08 & 0.10 \\ 0.05 & -0.02 \\ -0.30 & 0.25 \\ 0.12 & -0.05 \\ 0.07 & 0.04 \\ -0.50 & 0.60 \\ 0.20 & -0.10 \\ -0.05 & 0.08 \\ 0.03 & -0.04 \end{bmatrix}$ Compute the threshold for identifying influential observations and determine which data points are influential for each predictor.			
			OR			
	6	a)	Discuss the use of Studentized Deleted Residuals in identifying outliers. How do they differ from regular residuals?	*CO2*	*PO2*	**06**
		b)	Explain model validation criteria such as AIC, BIC, and Mallows' Cp. How are they used in model selection?	*CO3*	*PO1*	**06**
		c)	For the given dataset: (see table below) The fitted regression model and the error mean sum of square are given as $\hat{y} = 80.93 - 5.84X_1 + 11.32X_2$ and $MSE = 574.9$. i. Compute the fitted values and residuals ii. Identify the outliers in Y by computation of studentized residuals	*CO3*	*PO3*	**08**
			UNIT - IV			
	7	a)	Define a contingency table and explain how it is used in categorical data analysis.	*CO1*	*PO1*	**06**
		b)	Explain the concept of an odds ratio. How is it calculated and interpreted?	*CO1*	*PO2*	**07**
		c)	Describe the chi-squared test for independence. Provide the formula and explain its application.	*CO2*	*PO1*	**07**
			OR			
	8	a)	Explain the logistic regression model and its importance in analyzing categorical data.	*CO1*	*PO2*	**06**
		b)	Discuss how sensitivity, specificity, and ROC curves are used to summarize the predictive power of a logistic regression model.	*CO3*	*PO1*	**07**
		c)	How are parameters interpreted in logistic regression? Provide an example.	*CO3*	*PO1*	**07**

Dataset for 6 c):

i	**1**	**2**	**3**	**4**
y_i	301	327	246	187
x_{i1}	14	19	12	11
x_{i2}	25	32	22	15
h_{ii}	0.28	0.33	0.85	0.68

			UNIT - V			
	9	a)	Let $\{X_n, n \geq 0\}$ be a discrete time Markov Chain with state space {1,2,3,4} and the transition probability matrix A given as: $$A = \begin{pmatrix} 0.4 & 0.3 & 0.3 \\ 0.5 & 0.0 & 0.5 \\ 0.1 & 0.2 & 0.7 \end{pmatrix}$$ Given the initial probabilities $\pi = (0.25 \quad 0.15 \quad 0.6)$ Compute the following: i. $P(X_2 = 2 \| X_0 = 1)$ ii. $P(X_3 = 3, X_2 = 2)$ iii. $P(X_2 = 2, X_1 = 1 \| X_0 = 1)$	*CO3*	*PO2*	**06**
		b)	Describe the Forward-Backward algorithm and its applications in Hidden Markov Models.	*CO3*	*PO2*	**07**
		c)	Discuss the importance of Gaussian mixture models with Hidden Markov Models and their applications.	*CO3*	*PO2*	**07**
			OR			
	10	a)	Given a Hidden Markov Model (HMM) with the following parameters: Transition Probability Matrix (A), Emission Probability matrix (B) and initial probabilities (π) as: $\boldsymbol{A} = \begin{matrix} & p & q \\ p & 0.7 & 0.3 \\ q & 0.4 & 0.6 \end{matrix}$, $\boldsymbol{B} = \begin{matrix} & 0 & 1 \\ p & 0.5 & 0.5 \\ q & 0.1 & 0.9 \end{matrix}$ and $\boldsymbol{\pi} = [0.6 \quad 0.4]$ Compute the likelihood probability of the sequence {0, 1, 0} using backward probabilities obtained from the backward Algorithm.	*CO3*	*PO2*	**06**
		b)	Compare generative and discriminative classifiers. Provide examples where each is preferred.	*CO3*	*PO1*	**07**
		c)	Discuss the challenges in choosing the number of hidden states for an HMM and describe how smoothing and filtering techniques are applied.	*CO3*	*PO2*	**07**

******