Here’s the full content for the first question paper, including all questions and answers in one flow. I’ve also italicized the theory explanations for clarity.

Question Paper Date: 22/11/23

Question 1

Part of an annual transparency report published by a leading multinational technology company is as shown below:

Country	CR req	CR compl in %	UD req	UD compl in %	Hemi	HDI
Austria	21	100	134	32	Southern	High
Belgium	10	33	73	38	Northern	High
Brazil	224	67	703	82	Southern	Medium
Somalia	104	31	227	61	Southern	Poor
USA	92	63	5950	78	Northern	High

Variables Explanation:
- **CR req**: Content removal requests
- **CR compl**: Content removal compliance in %
- **UD req**: User data requests
- **UD compl**: User data compliance in %
- **Hemi**: Hemisphere
- **HDI**: Human Development Index

Question: *Identify each variable as Discrete Numerical, Continuous Numerical, Ordinal Categorical, or Regular Categorical with justification.*

Answer:
1. **Content Removal Requests (CR req)** - Discrete Numerical: Counts the number of requests, represented by whole numbers, making it discrete.
2. **Content Removal Compliance (CR compl)** - Continuous Numerical: This percentage shows compliance as a proportion, allowing any value within a range.
3. **User Data Requests (UD req)** - Discrete Numerical: Counts the number of data requests, so it is discrete.
4. **User Data Compliance (UD compl)** - Continuous Numerical: Given as a percentage, representing proportions, allowing a range of values.
5. **Hemisphere (Hemi)** - Regular Categorical: Represents geographic regions without any inherent order.
6. **Human Development Index (HDI)** - Ordinal Categorical: Although categorical, HDI levels (e.g., High, Medium, Poor) have an ordered ranking.

Question 2

A. *What are Type-I and Type-II errors in hypothesis testing? Which error should be minimized in court judgments? Justify.*

Theory:

- *Type-I Error (False Positive)*: Incorrectly rejecting a true null hypothesis.
- *Type-II Error (False Negative)*: Failing to reject a false null hypothesis.

Answer:
- In court, minimizing **Type-I Error** is critical because a Type-I Error would mean wrongly convicting an innocent person. It's more important to avoid punishing someone innocent than to potentially let a guilty person go free, making Type-I Errors more significant to minimize in legal contexts.

B. *A sample of 30 newspaper readers was asked about the total hours they spend per week reading newspapers. The sample had an average of 3.2 hours with a standard deviation of 1.74. Calculate the 95% confidence interval based on this data.*

Formula for Confidence Interval (CI):
$$
CI = \text{mean} \pm Z \times \left(\frac{s}{\sqrt{n}}\right)
$$

Given:
- Sample mean = 3.2 hours
- Standard deviation $s = 1.74$
- Sample size $n = 30$
- Z-score for 95% confidence level = 1.96

Solution:
$$
CI = 3.2 \pm 1.96 \times \left(\frac{1.74}{\sqrt{30}}\right) \approx 3.2 \pm 0.62
$$
Thus, the 95% confidence interval range is approximately **(2.58, 3.82) hours**.

Question 3

A. *Calculate the distance between points A (2,7,4) and D (3.2,4.8,5.8) using:*
  - *i. Manhattan Distance metric*
  - *ii. Euclidean Distance metric*

Theory:
- *Manhattan Distance*: The sum of the absolute differences of their Cartesian coordinates.
$$
d = |x_1 - x_2| + |y_1 - y_2| + |z_1 - z_2|
$$
- *Euclidean Distance*: The straight-line distance between two points in Euclidean space.
$$
d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}
$$

Solution:
1. **Manhattan Distance**:
   $$
   d = |2 - 3.2| + |7 - 4.8| + |4 - 5.8| = 1.2 + 2.2 + 1.8 = 5.2 \text{ units}
   $$
2. **Euclidean Distance**:
   $$
   d = \sqrt{(2 - 3.2)^2 + (7 - 4.8)^2 + (4 - 5.8)^2} = \sqrt{1.44 + 4.84 + 3.24} \approx 3.16 \text{ units}
   $$

\]

B. *For a given univariate function \(f(x) = 2x^4 - 8x^3 - 112x^2 + 1717 \), find the optimal local minimum and global minimum.*

Solution:
1. *Calculate \(f'(x) \) and \(f''(x) \)* to find critical points.
2. By testing values or solving, the optimal local minimum is identified around \(x \approx 4 \).

Question 4

A. *Following a linear regression process, the performance values are: SSR (Sum of Squares due to Regression) = 92.48; SSE (Sum of Squares due to Error) = 12.87. Calculate the value of \(R^2 \) and comment on the fit.*

Theory:
- *\(R^2 \) measures the proportion of variance in the dependent variable explained by the independent variable(s).*

Solution:
\[
SST = SSR + SSE = 92.48 + 12.87 = 105.35
\]
\[
R^2 = \frac{SSR}{SST} = \frac{92.48}{105.35} \approx 0.878
\]
Conclusion: An \(R^2 \) value of 0.878 indicates a strong fit.

B. *Given a logistic regression function with savings and loan status data, use the function:*
\[
\pi = \frac{1}{1 + e^{-(-4.07778 + 1.5046 \times \text{savings})}}
\]
- *Predict for a loan applicant with savings of Rs. 2.5 Lacs.*
- *Calculate the classification accuracy.*

Solution:
1. **Prediction for Rs. 2.5 Lacs**:
\[
\pi \approx \frac{1}{1 + e^{-(-4.07778 + 1.5046 \times 2.5)}} \approx 0.59
\]
- This suggests approval with 59% likelihood.

2. **Classification Accuracy**: Compare predicted vs. actual loan status and compute accuracy by matches/total cases.

Question 5

A. *The training data for a supervised classification is as follows: X1 (1.8, 1.6, 1), X2 (2.1, 1.8, 1), X3 (3.2, 2.4, 1), X4 (2.4, 2.6, 1), X5 (6.5, 4.2, 2), X6 (7.3, 4.3, 2), X7 (6.5, 4.2, 2), X8 (7.0, 4.8, 2). The test data point is at (4.6, 3.2). Use:*
1. *Nearest Neighbor to assign an appropriate class.*

2. *K-Nearest Neighbor with \(k = 3 \) to assign an appropriate class.*
3. *Weighted K-Nearest Neighbor with \(k = 3 \) to assign an appropriate class.*

B. *Consider 50 patterns split into 3 classes with 12, 28, and 10 samples. Calculate the Entropy impurity, Gini impurity, and Misclassification impurity at the node.*

Question 6

A. *Explain the 'Hold-out approach' used in planning the training and testing data.*

Theory:
- *Hold-out Approach:* This method involves splitting the data into separate training and testing sets. The model is trained on one portion (training set) and evaluated on the other (testing set) to assess performance.

B. *A confusion matrix for a classification model used to detect authentic and spam messages is as follows:*

| | Predicted Authentic |

 Predicted Spam |
|---|---|---|
| Actual Authentic | 1202 | 29 |
| Actual Spam | 5 | 54 |

Calculate Accuracy, Precision, Recall, and F-Score.

This includes all questions and answers for the first question paper. Let me know if there's anything else you'd like me to clarify!