MIDTERM TEST

MACHINE LEARNING (CO3117)

Notes:

- At most 2 hand-written A4 cheat sheets is allowed.
- You may use calculators; Round numerical answers to 1 decimal place; Clearly state any assumptions you make
- Show all your calculations clearly.
- Can use pencils for drawing diagrams.

Instruction: Consider the training and test datasets below for a loan risk problem, answer 10 questions (each question is worth 1.0 marks) by showing you detailed work step-by-step.

Training Dataset (8 records)

ID	Age	CreditScore	Education	RiskLevel
1	35	720	16	Low
2	28	650	14	High
3	45	750	missing	Low
4	31	600	12	High
5	52	780	18	Low
6	29	630	14	High
7	42	710	16	Low
8	33	640	12	High

Test Dataset (2 records)

ID	Age	CreditScore	Education
T1	37	705	16
T2	30	645	missing

Question 1. *(L.O.1, L.O.2)* Calculate the information gain for splitting CreditScore at 650 in a decision tree classification task, then explain why you would or wouldn't choose this as the root node split.

Question 2. *(L.O.1, L.O.2)* For a regression decision tree predicting CreditScore, calculate the variance reduction when splitting on Age=35, and describe how this splitting criterion differs from information gain in classification trees.

Question 3. *(L.O.1, L.O.2)* Using both CreditScore and Age patterns in the training data, determine the probability of T2 being High Risk given its missing Education value, then propose a method to handle similar missing values in future cases.

Question 4: *(L.O.1, L.O.2)* Design a perceptron to classify T1 by showing the input normalization and prediction calculation using weights $[\mathbf{0.3}, \mathbf{0.4}]$ and bias 0.1, then explain why normalization is necessary for neural networks.

Question 5: *(L.O.1, L.O.2)* For a single hidden layer neural network classifying T1, demonstrate one complete forward pass calculation and explain how the error would propagate backwards if the prediction was incorrect.

Question 6. *(L.O.1, L.O.2)* Apply Naive Bayes to classify T1 by calculating all required probabilities using the training data, then compare this with a non-naive Bayesian approach by explaining their key differences.

Question 7. *(L.O.1, L.O.2)* For genetic algorithm-based feature selection, demonstrate a crossover operation between two example chromosomes you create, then explain how you would handle invalid offspring considering feature dependencies.

Question 8. *(L.O.3)* Identify potential sources of bias in the training dataset by analyzing the feature distributions, then propose two specific methods to reduce these biases with justification.

Question 9. *(L.O.3)* Using predictions from your perceptron (Question 4) and Naive Bayes (Question 6) models, calculate precision and recall metrics, then recommend which metric is more important for loan risk assessment.

Question 10. *(L.O.3)* Calculate the variance and entropy of the CreditScore feature for both risk classes, then use your results to explain how different ML models would handle this data distribution.

--- THE END ---