A PRACTICAL PROBLEM

Instruction:

a) *Consider the training and test datasets below for a loan risk problem, and answer the following questions by showing your written detailed work step-by-step on paper (then scan to a PDF) or compose a PDF file (push it to the GitHub repo below due to your progress).*

b) *Compose a cheat sheet containing all necessary terminologies, methods, models, and formulas for solving these questions (push it to the GitHub repo below due to your progress). [This is very important for your final examination.]*

c) *Write a notebook to solve these questions using Python on a GitHub repo, commit messages using an issue for each question, and submit the link (this week, see the deadline) so that I can check your progress and results.*

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. 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. 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. 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. Implement batch gradient descent to find the optimal weights for predicting CreditScore using Age as input. Starting with initial parameters $\theta_0=500$, $\theta_1=5$, compute the cost function and one iteration of gradient descent updates using learning rate $\alpha=0.01$. Interpret the direction of the parameter updates.

Question 5. Perform multiple linear regression to predict CreditScore using Age and Education. Using the training data, calculate the normal equation solution $(X^TX)^{-1}X^Ty$, and interpret the resulting coefficients' meanings.

Question 6. Calculate the mean squared error and R^2 value for your linear regression model from the previous question when predicting the CreditScore values in the training set. Based on these metrics, assess whether linear regression is appropriate for this relationship.

Question 7. Formulate the logistic regression model for predicting RiskLevel using Age and CreditScore. Calculate the initial prediction for T1 using weights $w_0=0.5$, $w_1=-0.02$, $w_2=0.01$, then compute the cost function value.

Question 8. Using the prediction from the previous question, calculate the gradient vector for one step of gradient descent optimization. Explain how regularization would modify these gradients and why it might be necessary for this dataset.

Question 9. Design a perceptron to classify T1 by showing the input normalization and prediction calculation using weights [0.3, 0.4] and bias 0.1, then explain why normalization is necessary for neural networks.

Question 10. For a single hidden layer neural network classifying T1, demonstrate one complete forward pass calculation and explain how the error would propagate backward if the prediction was incorrect. Consider a neural network with two input neurons (normalized Age and CreditScore), one hidden layer with two neurons (using sigmoid activation), and one output neuron (using sigmoid activation) classifying loan risk. Given T1 (Age=37, CreditScore=705) with normalized values [0.375, 0.583], perform one forward pass with weights W1=[[0.3, 0.5], [0.4, -0.2]], biases b1=[0.1, -0.1], output weights W2=[0.6, -0.4], and output bias b2=0.2. Then calculate the gradients for all weights using backpropagation, assuming the target output is 1 (Low Risk) and using a learning rate of 0.1.

Question 11. 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 12. 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 13. Using predictions from your perceptron (Question 9) and Naive Bayes (Question 1111) models, calculate precision and recall metrics, then recommend which metric is more important for loan risk assessment.

Question 14. 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.

Question 15. Calculate the optimal margin hyperplane for separating the Low and High risk classes using only Age and CreditScore features. First, identify at least two support vectors from the training data, then explain how the margin would be affected if you were to use a soft margin SVM instead of a hard margin approach.

Question 16. Using the loan risk dataset, determine the nonlinear decision boundary by applying a kernel SVM with a radial basis function (RBF). Calculate the kernel matrix for the first three training samples using $\gamma=0.1$, then explain how this transformation helps classify samples that aren't linearly separable in the original feature space.

Question 17. Calculate the Euclidean distances between test case T1 and all training samples using normalized Age and CreditScore features. For k=3, determine T1's classification and explain how the choice of k affects the decision boundary.

Question 18. Compare distance-weighted k-NN with standard k-NN by calculating T1's risk classification probabilities using both methods with k=3. Explain which approach would be more robust for this dataset and why.

Question 19. Construct a simple Hidden Markov Model for modeling credit risk progression over time with states {Low, Medium, High}. Using the Viterbi algorithm, calculate the most likely sequence of hidden states given observations [710, 650, 680] and transition probabilities P(Low→Low)=0.7, P(Low→Medium)=0.3, P(Medium→Medium)=0.6, P(Medium→High)=0.4, P(High→High)=0.8, P(High→Medium)=0.2.

Question 20. Calculate the probability of observing the sequence [705, 645] using the forward algorithm with the Hidden Markov Model Hidden Markov Model defined in the previous question. Explain how this model could help predict future credit behavior.
