E-commerce Customer Satisfaction Prediction

Consider the following dataset for an e-commerce company that wants to predict customer satisfaction based on product reviews. Answer 10 questions (each question is worth 10 marks) by showing your detailed work step-by-step.

Training Dataset (8 entries):

ID	Text Sentiment	Product Visibility	Image Quality	Satisfaction
1	Positive	High	Good	Satisfied
2	Negative	High	Fair	Unsatisfied
3	Neutral	Low	Poor	Unsatisfied
4	Positive	Medium	Fair	Satisfied
5	Negative	Low	Good	Unsatisfied
6	Neutral	High	Good	Satisfied
7	Positive	High	Poor	Satisfied
8	Negative	Medium	Poor	Unsatisfied

Test Dataset (3 entries):

ID	Text Sentiment	Product Visibility	Image Quality	Satisfaction
9	Positive	Low	Good	?
10	Neutral	Medium	Fair	?
11	Negative	Low	Poor	?

Question 1:

A machine learning model made the following predictions on the training dataset:

ID	Actual Satisfaction	Predicted Satisfaction
1	Satisfied	Satisfied
2	Unsatisfied	Unsatisfied
3	Unsatisfied	Unsatisfied
4	Satisfied	Satisfied
5	Unsatisfied	Satisfied
6	Satisfied	Satisfied
7	Satisfied	Unsatisfied
8	Unsatisfied	Unsatisfied

a) Construct the confusion matrix for this model.

b) Calculate the accuracy, precision, recall, and F1-score for the "Satisfied" class. Which metric would be most appropriate if we want to ensure customers with satisfaction issues are properly identified? Explain your reasoning.

Question 2:

a) Calculate the initial entropy of the training dataset with respect to Customer Satisfaction.

b) Calculate the information gain for each of the three features (Text Sentiment, Product Visibility, and Image Quality). Which feature should be selected as the first split in a decision tree? Show all your calculations.

Question 3:

Suppose in the training dataset, entry #6 is missing its "Product Visibility" value.

a) Describe three different strategies for handling this missing value, and explain the potential impact of each strategy on the model's performance.

b) If you choose to predict the missing value using a decision tree based on the other two features, construct a small decision tree to predict "Product Visibility" and determine what value would be assigned to entry #6. Show your work.

Question 4:

a) Design a single perceptron to classify the data into "Satisfied" or "Unsatisfied" based on the three features. Assign appropriate numerical values to the categorical features and explain your encoding scheme. Draw the perceptron with initial small random values.

b) Explain why a single perceptron might not be able to properly classify this dataset. Design a minimal neural network (with input layer, one hidden layer, and output layer) that could better handle this classification task. Specify the activation functions you would use and explain your choices.

Question 5:

Consider a simple neural network with:

- 3 input nodes (one for each feature, appropriately encoded)
- 2 hidden nodes with sigmoid activation
- 1 output node with sigmoid activation for binary classification

a) Starting with the weights given below, perform one step of forward propagation for training example #1:

- w_input_to_hidden = $[[0.1, 0.2], [0.3, -0.1], [0.2, 0.1]]$
- w_hidden_to_output = $[[0.4, -0.3]]$
- Encode Positive=1, Neutral=0, Negative=-1 for Text Sentiment
- Encode High=1, Medium=0.5, Low=0 for Product Visibility
- Encode Good=1, Fair=0.5, Poor=0 for Image Quality

Show all calculations and the final output.

b) If the actual target for example #1 is 1 (Satisfied), calculate the loss using binary cross-entropy and perform one step of backpropagation to update the weight from the first hidden node to the output node. Use a learning rate of 0.1. Show all your calculations.

Question 6:

a) Using the training dataset, calculate the prior probabilities P(Satisfied) and P(Unsatisfied).

b) Using Naive Bayes, calculate the probability that test example #9 is "Satisfied" vs. "Unsatisfied." Apply Laplace smoothing with α=1 to handle zero probabilities. Show all your calculations and determine the final prediction.

Question 7:

a) Explain what the Bayes Optimal Classifier is and how it relates to the Naive Bayes classifier. What assumptions does Naive Bayes make that the Bayes Optimal Classifier doesn't?

b) For our dataset, identify a specific case where the independence assumption of Naive Bayes might be violated. Calculate the joint probability P(Text Sentiment, Product Visibility |

Satisfaction) for this case both with and without the independence assumption to demonstrate the potential difference.

Question 8:

a) Design a genetic algorithm for feature selection in our dataset:

- Define a suitable chromosome representation
- Specify fitness function
- Describe selection, crossover, and mutation operators
- Define termination criteria

Explain how your design choices are appropriate for this specific dataset.

b) Trace through two generations of your genetic algorithm starting with an initial population of four randomly generated chromosomes. Show how selection, crossover, and mutation would work, and how the best solution evolves.

Question 9:

a) The training dataset has equal numbers of "Satisfied" and "Unsatisfied" examples, but in reality, 80% of all customers are satisfied. Discuss the potential consequences of this sampling bias and propose two methods to address this issue.

b) A decision tree model perfectly classifies all training examples but performs poorly on new data. Explain how decision tree pruning could help with this overfitting problem. Based on the training dataset, suggest specific pruning that could be applied and explain its expected impact on model performance.

Question 10:

a) Explain the concept of k-fold cross-validation to evaluate model performance more reliably. Suggest a specific cross-validation strategy for this dataset and explain why it's appropriate.

b) If a model achieves 75% accuracy on the training data and 67% accuracy on a separate validation set, interpret these results. Is the model likely underfitting, overfitting, or appropriately fit? Explain why and suggest ways to improve the model.