# Rubric

## Part 1: Cancer

### Task 1 (6 points)

1. (+6.0) All of the below
2. (+2.0) Dataset loaded correctly (breast cancer dataset).
3. (+2.0) Train/test split done correctly into X_train, X_test, y_train, y_test.
4. (+2.0) Number of benign (1) vs malignant (0) in training data computed/printed.

---

### Task 2 (8 points)

1. (+8.0) All of the below
2. (+2.0) Random predictions generated with probabilities from training class proportions.
3. (+2.0) Confusion matrix shown with correct labels.
4. (+2.0) Accuracy, precision, recall correctly computed.
5. (+2.0) Short written interpretation of accuracy/precision/recall in context of benign/malignant.

---

### Task 3 (8 points)

1. (+8.0) All of the below
2. (+2.0) Features scaled with StandardScaler (fit on train only, applied to both train/test).
3. (+2.0) Logistic regression fit with penalty=None, random_state=2025, max_iter=10000.
4. (+2.0) Training and test accuracy both reported.
5. (+2.0) Results printed clearly with values.

---

### Task 4 (9 points)

1. (+9.0) All of the below
2. (+1.0) Run study across all six C values [0.001, 0.01, 0.1, 1.0, 10.0, 100.0].
3. (+1.0) 5-fold CV mean accuracy stored in val_scores.
4. (+1.0) Sparsity ratio calculated with threshold 0.0001.
5. (+2.0) Validation accuracy curve and sparsity curve clearly shown.

6. (+2.0) Table of results (C, accuracy, sparsity) printed.
7. (+2.0) Written explanation identifying 2–3 reasonable C choices.

---

### Task 5 (Bonus)

1. (+5.0) All of the below
2. (+3.0) Both sparse logistic regression (chosen C) and decision tree (max_depth=3) trained.
3. (+1.0) Test accuracy, precision, recall, confusion matrices for both reported.
4. (+1.0) Short written comparison included.

---

## Part 2: Logistic

### Task 1 (8 points)

1. (+8.0) All of the below
2. (+2.0) Import LogisticRegression from sklearn.linear_model
3. (+1.0) Import accuracy_score from sklearn.metrics.
4. (+1.0) Logistic regression model initialized with penalty=None parameter
5. (+1.0) Model fitted on X_train and y_train
6. (+1.0) Make correct predictions on training set (X_train)
7. (+1.0) Make correct calculation of training accuracy
8. (+1.0) Make correct calculation of test accuracy

---

### Task 2 (16 points)

1. (+16.0) All of the below
2. (+2.0) (predict_proba method) Compute linear predictions using X @ self.weights
3. (+2.0) (predict_proba method) Apply sigmoid function to linear predictions
4. (+2.0) (predict method) Use predict_proba method to get probabilities
5. (+1.0) (predict method) Apply 0.5 threshold to convert probabilities to class labels
6. (+2.0) (fit method) Weights initialized using np.random.normal with correct shape (n_features,)
7. (+1.0) (fit method) Compute linear predictions (X @ self.weights)
8. (+1.0) (fit method) Apply sigmoid to get predicted probabilities
9. (+1.0) (fit method) Compute prediction errors (predictions - y)

10. (+1.0) (fit method) Compute gradient using X.T @ errors / n_samples
11. (+2.0) (fit method) Update weights using gradient descent rule: weights - lr * gradient

---

### Task 3 (10 points)

1. (+10.0) All of the below
2. (+1.0) Create instance of BinaryLogisticRegression class
3. (+1.0) Call fit method on training data (X_train, y_train)
4. (+2.0) Make predictions on training set using DIY model
5. (+2.0) Make predictions on test set using DIY model
6. (+2.0) Calculate accuracy for test predictions
7. (+2.0) Calculate accuracy for training predictions

---

### Task 4 (Bonus)

1. (+5.0) All of the below
2. (+1.0) Test all seven specified learning rates: [100, 10, 1, 0.1, 0.01, 0.001, 0.0001]
3. (+2.0) Evaluate accuracy on training set and number of iterations for each learning rate
4. (+2.0) Results clearly displayed in an organized format along with written explanation of learning rate importance considering both performance and computational complexity

---

## Part 3: Toxic

### Task 1 (8 points)

1. (+8.0) All of the below
2. (+2.0) Load data using pandas, csv module, or appropriate method
3. (+2.0) Create target variable y with correct 3-class encoding (0=normal, 1=toxic, 2=severely toxic)
4. (+1.0) Target variable logic correct (severely toxic prioritized over toxic, and correct class assignments)
5. (+1.0) Fit and transform text data into bag-of-words representation
6. (+1.0) Print or display the shape of matrix X
7. (+1.0) Final X matrix has the right shape or dimensions

---

### Task 2 (10 points)

1. (+10.0) All of the below
2. (+1.0) Use 30% test size or random_state=2025
3. (+1.0) Compute validation curve for the specified C values [0.001, 0.01, 0.1, 0.5, 1.0]
4. (+1.0) Use cv=3 for cross-validation in validation curve
5. (+1.0) Visualize or display validation curve results
6. (+1.0) Select optimal C value based on validation curve results
7. (+1.0) Create LogisticRegression model with selected C value
8. (+1.0) Fit model on training data (X_train, y_train)
9. (+1.0) Make predictions on test set
10. (+1.0) Compute confusion matrix for test set predictions
11. (+1.0) Visualize or display confusion matrix

---

## Task 3 (11 points)

1. (+11.0) All of the below
2. (+1.0) (Resampling Method) Implement upsampling or downsampling to create balanced training data
3. (+1.0) (Resampling Method) Does not apply resampling to test set (should only resample training data)
4. (+1.0) (Resampling Method) Create balanced training set with equal class representation
5. (+2.0) (Class Weighting Method) Use class_weight='balanced' parameter in LogisticRegression
6. (+1.0) Use the same C value selected in Task 2 for both methods
7. (+1.0) Discuss false positives vs false negatives in content moderation context
8. (+1.0) Clearly select one method as the final model
9. (+1.0) Confusion Matrix Visualization: display confusion matrices for both resampling and class weighting methods
10. (+2.0) Provide clear reasoning for method selection based on error types and content moderation needs

---

## Task 4 (6 points)

1. (+6.0) All of the below
2. (+1.0) Clearly state a position on whether ML models should be used for toxic speech detection/removal
3. (+2.0) Response is significantly not shorter than 2-3 paragraphs
4. (+3.0) Provide reasoning or evidence to support their position