Question 1: Data Preprocessing

1(a): Identify and Resolve Data Quality Issues

Step 1: Understand the Dataset

- The dataset contains attributes like `NAME`, `AGE`, `HEIGHT`, `WEIGHT`, `BLOOD GROUP`, and `COVID-19 RESULT`.
- Before using the data for analysis or modeling, we need to clean it by addressing inconsistencies and errors.

Step 2: Identify Data Quality Issues

1. **Inconsistent Data Formats**:
   - Example: `WEIGHT` column has entries like `62kg`, `120lbs`.
   - **Problem**: These entries are in different units and formats, making them unusable for calculations.
   - **Solution**:
     - Convert all weights to a single unit (e.g., kilograms).
     - Formula: $\text{weight in kg} = \text{weight in lbs} \times 0.453592$.
     - Remove text like "kg" or "lbs" and keep only numerical values.
2. **Missing or Invalid Values**:
   - Example: `COVID-19 RESULT` column has invalid values like `1`, `0`.
   - **Problem**: These do not match valid results ("Positive"/"Negative").
   - **Solution**:
     - Replace invalid values with proper labels.
     - If unsure, mark them as missing (`NULL`) and handle them later.
3. **Outliers in Numerical Data**:
   - Example: `AGE = 350` is unrealistic.
   - **Problem**: This value does not align with real-world expectations.
   - **Solution**:
     - Replace such outliers with a realistic value (e.g., the mean or median age of the dataset).
4. **Inconsistent Text Formatting**:
   - Example: Names like `RAMA` vs. `Akbar`.
   - **Problem**: Inconsistent capitalization reduces clarity.
   - **Solution**:

- Standardize text capitalization (e.g., "Rama", "Akbar").

5. **Duplicate or Erroneous Records**:
   - Example: `NAME = SysUsr789` seems system-generated and invalid.
   - **Problem**: Such entries can skew results.
   - **Solution**:
     - Flag such records for review or remove them if confirmed invalid.

1(b): Generative vs. Discriminative Classifiers

What are Generative Classifiers?

- Generative classifiers model the **joint probability distribution** $P(x, y)$.
- They can:
  1. Predict classes.
  2. Detect outliers using density estimation.

What are Discriminative Classifiers?

- Discriminative classifiers focus on the **decision boundary** by modeling $P(y|x)$.
- They are efficient for classification but cannot perform density estimation.

Which to Use?

- **Recommendation**: Use **Generative Classifiers** (e.g., Naive Bayes).
- **Reason**: They allow both classification and outlier detection.

Question 2: Ridge Regression and Feature Scaling

2(a): High Bias vs. High Variance

Step 1: Understand the Problem

- Training error and validation error are both high.
- **Conclusion**: The model is underfitting (high bias).

Step 2: Solve High Bias

1. **Decrease Regularization** (λ):
   - Ridge regression penalizes large coefficients. A high λ overly restricts the model.
   - Lower λ to increase flexibility and reduce bias.
2. **Add Complexity**:
   - Include polynomial or interaction terms to better capture data patterns.
3. **Use Cross-Validation**:
   - Tune λ to find the optimal value.

2(b): Gradient Descent vs. Least Squares

When to Use Gradient Descent?

- Dataset: $n = 2,000,000$, $m = 300,000$.
- **Reason**:
  - Gradient descent is iterative, with a complexity of $O(n \cdot m)$ per iteration.
  - Least squares involves inverting a matrix ($O(m^3)$), which is infeasible for large m.

2(c): Importance of Feature Scaling

Why Scale Features?

1. **Faster Convergence**:
   - Without scaling, gradient updates become unbalanced, slowing optimization.
2. **Avoid Bias**:
   - Features with larger ranges dominate smaller ones.
3. **Improve Numerical Stability**:
   - Prevents overflow/underflow errors.

Techniques:

1. **Standardization**:

- Formula: $z = \frac{x-\mu}{\sigma}$.
- Ensures mean = 0, standard deviation = 1.

2. **Normalization**:
   - Formula: $x' = \frac{x-\min(x)}{\max(x)-\min(x)}$.
   - Scales values to [0, 1].

Question 3: Logistic Regression Analysis

Evaluate Statements:

1. **(a)**: "The model will work perfectly well on unseen data."
   - **Answer**: False.
   - **Reason**: Overfitting leads to poor generalization.
2. **(b)**: "If $J(\theta_0, \theta_1) = 0$, predictions match actual labels for training data."
   - **Answer**: True.
3. **(c)**: "For $J(\theta_0, \theta_1) = 0$, θ_0 and θ_1 must be 0."
   - **Answer**: False.
   - **Reason**: θ values depend on data.
4. **(d)**: "Cost function $J(\theta_0, \theta_1)$ cannot be 0."
   - **Answer**: False.
   - **Reason**: It can be 0 for perfectly separable data.

Question 4: Fraud Detection Performance Metrics

Contingency Table:

True Class	Predicted Fraud	Predicted Not Fraud
Fraud	60 (TP)	0 (FN)
Not Fraud	120 (FP)	20 (TN)

Key Metrics:

1. **Accuracy**:

$$Accuracy = \frac{TP + TN}{Total} = 0.4$$

2. **Precision**:

$$Precision = 0.33$$

3. **Recall**:

$$Recall = 1.0$$

4. **F1-Score**:

$$F1 = 0.5$$

Observations:

- High recall ensures no fraud is missed.
- Low precision indicates many false positives.

Question 5: Decision Tree Using ID3 Algorithm

Step-by-Step:

1. **Calculate Entropy of Target Variable**:
    - $H(Target) = 1.5\,\text{bits.}$
2. **Calculate Entropy for Each Feature**:
    - Split data by feature values.
    - Compute weighted entropy.
3. **Calculate Information Gain**:

$$IG = H(Target) - H(Feature)$$

- Feature with the highest IG is the root node.

Question 6: Model Performance Justification

Statement:

"Assessing model performance using only training data is detrimental."

Justification:

1. **Overfitting**:
    - Training accuracy does not reflect generalization.
2. **Solution**:
    - Use validation/testing datasets for evaluation.
    - Apply cross-validation for better insights.