## Q1

Feature Engineering:

Feature engineering is the process of creating new features from existing data or modifying existing features to enhance the performance of machine learning models.

Aspects of feature engineering:

Feature Creation: Generating new features from raw data.

Feature Transformation: Scaling, encoding, or transforming features to meet model requirements.

Feature Extraction: Reducing the dimensionality of data while preserving essential information.

## Q2

Feature Selection:

Feature selection is the process of choosing a subset of relevant features to improve model performance and reduce overfitting.

Methods of feature selection:

Filter Methods: Statistically measure feature importance (e.g., correlation, chi-squared) independently of the learning algorithm.

Wrapper Methods: Use a predictive model to evaluate feature subsets' performance (e.g., forward selection, backward elimination).

## Q3

Filter vs. Wrapper Approaches:

Filter Methods:

Pros: Fast, less prone to overfitting, suitable for high-dimensional data.

Cons: Ignores feature dependencies, may not be optimal for complex models.

Wrapper Methods:

Pros: Considers feature interactions, suitable for complex models, can find optimal subsets.

Cons: Computationally expensive, prone to overfitting.

## Q4

Feature Selection Process:

i. Data Collection

ii. Data Preprocessing (e.g., scaling, encoding)

iii. Feature Selection (filter or wrapper methods)

iv. Model Training

v. Model Evaluation

vi. Iterate if necessary

Feature Extraction Principle: Reduce dimensionality while preserving information. Example: Principal Component Analysis (PCA).

Common Feature Extraction Algorithms: PCA, Linear Discriminant Analysis (LDA), t-Distributed Stochastic Neighbor Embedding (t-SNE).

## Q5

Feature Engineering in Text Categorization:

In text categorization, feature engineering includes tasks like:

Text preprocessing (e.g., tokenization, stemming).

Feature creation (e.g., TF-IDF vectors, word embeddings).

Handling categorical features (e.g., one-hot encoding of words).

## Q6

Cosine Similarity:

Cosine similarity is a metric used to measure the similarity between two non-zero vectors.

Cosine Similarity Calculation:

Dot product of vectors A and B divided by the product of their magnitudes.

Cosine Similarity Example Calculation:

Cosine Similarity = (A · B) / (||A|| \* ||B||)

## Q7

Distance Metrics:

i. Hamming Distance:

- Formula: Number of positions at which two strings of equal length differ.

- Hamming Distance between "10001011" and "11001111" = 2.

ii. Jaccard Index and Similarity Matching Coefficient:

- Jaccard Index: Intersection size divided by union size.

- Similarity Matching Coefficient: Intersection size divided by the smaller set's size.

- Calculation:

- Intersection = {1, 0, 0, 0, 1, 0, 0, 1}

- Jaccard Index = |Intersection| / |Union| = 4 / 12

- Similarity Matching Coefficient = |Intersection| / |Smaller Set| = 4 / 8

## Q8

High-Dimensional Data:

High-dimensional datasets have a large number of features compared to samples.

Real-life examples include gene expression data, text data with many words, and image data with many pixels.

Challenges: Increased computation time, overfitting, difficulty visualizing data.

Solutions: Feature selection, dimensionality reduction (e.g., PCA), regularization.

## Q9

Quick Notes:

PCA (Principal Component Analysis): Dimensionality reduction technique.

Use of Vectors: Vectors represent data points in multidimensional space.

Embedded Technique: Feature selection techniques embedded within the learning algorithm.

## Q10

Comparisons:

i. Sequential Backward Exclusion vs. Sequential Forward Selection:

- Backward Exclusion removes features iteratively, starting with all features.

- Forward Selection adds features iteratively, starting with none.

ii. Filter vs. Wrapper Methods:

- Filter methods evaluate feature importance independently of the learning algorithm.

- Wrapper methods use a predictive model to evaluate feature subsets.

iii. SMC vs. Jaccard Coefficient:

- SMC (Similarity Matching Coefficient) is based on intersection size and the smaller set's size.

- Jaccard Coefficient is based on intersection size and union size.