1. What is feature engineering, and how does it work? Explain the various aspects of feature engineering in depth.

>>>Feature engineering is the process of creating, selecting, or transforming features from raw data to improve the performance of machine learning models.

Aspects of Feature Engineering:

Creation: Designing new features based on domain knowledge or insights.

Selection: Choosing relevant features and discarding irrelevant ones.

Transformation: Modifying or scaling existing features to improve their representation.

Extraction: Deriving new features from existing ones using mathematical techniques.

2. What is feature selection, and how does it work? What is the aim of it? What are the various methods of function selection?

>>>Feature selection is the process of selecting a subset of relevant features to reduce dimensionality, improve model performance, and enhance interpretability.

Methods of Feature Selection:

Filter Methods: Evaluate features independently of the model using statistical tests or scoring functions.

Wrapper Methods: Use the model's performance on subsets of features to select the best subset.

3. Describe the function selection filter and wrapper approaches. State the pros and cons of each approach?

>>>Filter Approach:

Pros: Computationally efficient, independent of the model, reduces overfitting.

Cons: Ignores feature interactions, may not consider the model's performance.

Wrapper Approach:

Pros: Considers feature interactions, considers the model's performance, can optimize specific model performance.

Cons: Computationally expensive, can lead to overfitting.

4.

i. Overall Process: Starts with a pool of candidate features, evaluates their importance, selects the best subset, and tests model performance.

ii. Key Principle of Feature Extraction: Creating new features by combining or transforming existing ones. Example: Extracting polynomial features (quadratic, cubic) from original numeric features.

Common Feature Extraction Algorithms: Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA).

5. Describe the feature engineering process in the sense of a text categorization issue.

>>>In text categorization, features could include word frequencies, n-grams, or TF-IDF values.

Steps: Tokenization, stop word removal, stemming, vectorization (using techniques like TF-IDF), feature scaling.

6. What makes cosine similarity a good metric for text categorization? A document-term matrix has two rows with values of (2, 3, 2, 0, 2, 3, 3, 0, 1) and (2, 1, 0, 0, 3, 2, 1, 3, 1). Find the resemblance in cosine.

>>>Cosine similarity measures the cosine of the angle between two vectors.

Calculation: (22 + 31 + 20 + 00 + 23 + 32 + 31 + 03 + 1\*1) / (sqrt(2^2 + 3^2 + 2^2 + 0^2 + 2^2 + 3^2 + 3^2 + 0^2 + 1^2) \* sqrt(2^2 + 1^2 + 0^2 + 0^2 + 3^2 + 2^2 + 1^2 + 3^2 + 1^2)) = 0.7325.

7.

i. What is the formula for calculating Hamming distance? Between 10001011 and 11001111, calculate the Hamming gap.

>>>i. Hamming Distance Formula: Count positions where bits are different.

Hamming Distance = 4 (bits that differ between 10001011 and 11001111)

ii. Compare the Jaccard index and similarity matching coefficient of two features with values (1, 1, 0, 0, 1, 0, 1, 1) and (1, 1, 0, 0, 0, 1, 1, 1), respectively (1, 0, 0, 1, 1, 0, 0, 1).

>>>ii. Jaccard Index: Measures the similarity between two sets.

Jaccard Index = 3 / (3 + 2 + 2) = 0.5

8. State what is meant by "high-dimensional data set"? Could you offer a few real-life examples? What are the difficulties in using machine learning techniques on a data set with many dimensions? What can be done about it?

>>>High-dimensional data sets have a large number of features compared to the number of observations.

Real-Life Examples: Gene expression data, images, social network data.

Difficulties: Curse of dimensionality (increased computation, overfitting risk), reduced interpretability, visualization challenges.

9. Make a few quick notes on:

PCA: Stands for "Principal Component Analysis," a dimensionality reduction technique.

Use of Vectors: Vectors are used to represent data points and features in machine learning algorithms.

Embedded Technique: A method where feature selection is integrated with the model training process.

10. Make a comparison between:

Sequential Backward Exclusion vs. Sequential Forward Selection:

Backward Exclusion removes features iteratively.

Forward Selection adds features iteratively.

Filter vs. Wrapper Feature Selection:

Filter methods evaluate features independently.

Wrapper methods consider the model's performance.

3. SMC vs. Jaccard coefficient