1. Feature Engineering:

- Definition: Feature engineering involves creating new features or modifying existing ones to enhance the performance of machine learning models.

- Aspects of Feature Engineering:

1. Imputation: Handling missing data.

2. Encoding: Converting categorical variables.

3. Scaling: Normalizing numeric features.

4. Creation: Generating new features.

5. Transformation: Applying mathematical functions.

6. Binning: Grouping continuous values.

7. Interaction: Combining multiple features.

8. Aggregation: Summarizing data.

2. Feature Selection:

- Definition: Feature selection involves choosing a subset of relevant features to improve model performance and interpretability.

- Aim:

- Optimization: Improve model accuracy and efficiency.

- Simplification: Enhance model interpretability.

- Methods:

1. Filter Methods: Evaluate features' relevance independently.

2. Wrapper Methods: Use models to evaluate feature subsets.

3. Embedded Methods: Incorporate feature selection within the model.

3. Filter and Wrapper Approaches:

- Filter Approach:

- Pros:

1. Efficiency: Computationally less intensive.

2. Independence: Evaluates features individually.

- Cons:

1. Interaction Neglect: Ignores feature dependencies.

2. Model-Agnostic: May not align with specific model needs.

- Wrapper Approach:

- Pros:

1. Interaction Consideration: Considers feature dependencies.

2. Model Alignment: Tailored to specific model requirements.

- Cons:

1. Computational Intensity: Requires repeated model training.

2. Overfitting Risk: Can overfit to training data.

4. Feature Selection Process:

- Overall Process:

1. Data Collection: Gather relevant dataset.

2. Exploratory Data Analysis (EDA): Understand feature characteristics.

3. Correlation Analysis: Identify feature dependencies.

4. Feature Selection Technique Application: Apply filter, wrapper, or embedded methods.

5. Model Training: Train machine learning models.

6. Evaluation: Assess model performance.

7. Iteration: Refine feature selection based on results.

- Feature Extraction Principle:

- Example: Principal Component Analysis (PCA) extracts new features capturing maximum variance.

5. Feature Engineering in Text Categorization:

- Process:

1. Text Preprocessing: Tokenization, stemming, and stop-word removal.

2. TF-IDF Transformation: Convert text to numerical representations.

3. Feature Creation: N-grams, sentiment analysis, or topic modelling.

4. Word Embeddings: Utilize pre-trained embeddings.

5. Dimensionality Reduction: Apply techniques like PCA.

6. Cosine Similarity Calculation:

7. Hamming Distance and Jaccard Index:

8. High-Dimensional Dataset:

- Definition: A dataset with a large number of features.

- Examples:

1. Genomics Data: Gene expression profiles.

2. Image Data: Pixels in high-resolution images.

- Difficulties:

1. Curse of Dimensionality: Increased computational complexity.

2. Overfitting Risk: Models may perform well on training data but generalize poorly.

- Mitigation:

- Dimensionality Reduction Techniques: PCA, t-SNE.

- Feature Selection: Choose relevant features.

9. Quick Notes:

1. PCA (Principal Component Analysis):

- Definition: Dimensionality reduction technique.

- Application: Transforming data into principal components.

2. Use of Vectors:

- Application: Vectors represent data points or features in machine learning.

3. Embedded Technique:

- Definition: Feature selection methods integrated into the model training process.

10. Comparisons:

1. Sequential Backward Exclusion vs. Sequential Forward Selection:

- Difference:

- Backward Exclusion: Starts with all features and removes them sequentially.

- Forward Selection: Starts with an empty set and adds features sequentially.

2. Filter vs. Wrapper Feature Selection Methods:

- Filter Methods:

- Evaluate features independently.

- Less computationally intensive.

- Wrapper Methods:

- Use models to evaluate feature subsets.

- More computationally intensive.

3. SMC vs. Jaccard Coefficient:

- Difference:

- SMC measures the similarity between two binary features.

- Jaccard Coefficient measures the similarity between sets.