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 transforming raw data into features that are suitable for machine learning models. In other words, it is the process of selecting, extracting, and transforming the most relevant features from the available data to build more accurate and efficient machine learning models.

The key steps in feature engineering include: Data Exploration and Understanding: Explore and understand the dataset, including the types of features and their distributions.

1. Feature Creation

Feature Creation is the process of generating new features based on domain knowledge or by observing patterns in the data. It is a form of feature engineering that can significantly improve the performance of a machine-learning model.

2. Feature Transformation

[Feature Transformation](https://www.geeksforgeeks.org/feature-transformation-techniques-in-machine-learning/) is the process of transforming the features into a more suitable representation for the machine learning model. This is done to ensure that the model can effectively learn from the data.

3. Feature Extraction

[Feature Extraction](https://www.geeksforgeeks.org/feature-extraction-techniques-nlp/)is the process of creating new features from existing ones to provide more relevant information to the machine learning model. This is done by transforming, combining, or aggregating existing features.

4. Feature Selection

[Feature Selection](https://www.geeksforgeeks.org/feature-selection-techniques-in-machine-learning/) is the process of selecting a subset of relevant features from the dataset to be used in a machine-learning model. It is an important step in the feature engineering process as it can have a significant impact on the model’s performance

5. Feature Scaling

[Feature Scaling](https://www.geeksforgeeks.org/ml-feature-scaling-part-1/) is the process of transforming the features so that they have a similar scale. This is important in machine learning because the scale of the features can affect the performance of the model.

1. **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](https://www.geeksforgeeks.org/feature-selection-techniques-in-machine-learning/) is the process of selecting a subset of relevant features from the dataset to be used in a machine-learning model. It is an important step in the feature engineering process as it can have a significant impact on the model’s performance.

Types of Feature Selection:

**Filter Method:**Based on the statistical measure of the relationship between the feature and the target variable. Features with a high correlation are selected.

**Wrapper Method:**Based on the evaluation of the feature subset using a specific machine learning algorithm. The feature subset that results in the best performance is selected.

**Embedded Method:**Based on the feature selection as part of the training process of the machine learning algorithm.

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

i. Describe the overall feature selection process.  
  
[Feature Selection](https://www.geeksforgeeks.org/feature-selection-techniques-in-machine-learning/) is the process of selecting a subset of relevant features from the dataset to be used in a machine-learning model. It is an important step in the feature engineering process as it can have a significant impact on the model’s performance.

Types of Feature Selection:

**Filter Method:**Based on the statistical measure of the relationship between the feature and the target variable. Features with a high correlation are selected.

**Wrapper Method:**Based on the evaluation of the feature subset using a specific machine learning algorithm. The feature subset that results in the best performance is selected.

**Embedded Method:**Based on the feature selection as part of the training process of the machine learning algorithm.

**ii. Explain the key underlying principle of feature extraction using an example. What are the most widely used function extraction algorithms?**

Feature Extraction offers three methods for supervised classification: K Nearest Neighbor (KNN), Support Vector Machine (SVM), or Principal Components Analysis (PCA)

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

Text classification is the problem of assigning categories to text data according to its content. The most important part of text classification is feature engineering: the process of creating features for a machine learning model from raw text data.

**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 is a metric commonly used in text categorization and natural language processing tasks because it measures the cosine of the angle between two vectors, which represent the text documents in a high-dimensional space. Here are some reasons why cosine similarity is a good metric for text categorization:

Scale Invariance: Cosine similarity is scale-invariant, meaning it is not affected by the magnitude of the vectors but only by their directions. This is useful when dealing with text data, where the length of documents may vary.

Ignores Zero-Valued Dimensions: In text data, many dimensions (terms) have zero values for a given document. Cosine similarity ignores these zero values, focusing only on the non-zero dimensions. This is beneficial because it helps to capture the semantic similarity between documents based on the shared non-zero terms.

Efficient Computation: Cosine similarity can be efficiently computed, especially when representing documents as vectors in a high-dimensional space. This makes it computationally less expensive compared to some other similarity metrics.

Now, let's calculate the cosine similarity for the given document-term matrix:

Document 1: (2, 3, 2, 0, 2, 3, 3, 0, 1)

Document 2: (2, 1, 0, 0, 3, 2, 1, 3, 1)

The cosine similarity (cos θ) between two vectors A and B is calculated using the formula:

cosine similarity= (A . B) /(∥A∥∥B∥)

Where:

A⋅B is the dot product of vectors A and B.

∥A∥ and ∥B∥ are the Euclidean norms of vectors A and B, respectively.

Let's calculate it:

A . B= (2×2)+(3×1)+(2×0)+(0×0)+(2×3)+(3×2)+(3×1)+(0×3)+(1×1)=

∥A∥= √(22+32+22+02+22+32+32+02+12=39∥A∥=22+32+22+02+22+32+32+02+12)​= √

39 = 6.24​

∥B∥=√22+12+02+02+32+22+12+32+12=28∥B∥=22+12+02+02+32+22+12+32+12​=√28​ = 5.29

Now, plug these values into the cosine similarity formula:

cosine similarity=20/(6.24 \* 5.29) = 0.66​

7.

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

The Hamming distance is a measure of the difference between two strings of equal length. It is calculated by counting the number of positions at which the corresponding symbols are different

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).

J(A,B)=6/10​=0.6

J(A,C)=4/12​=31​≈0.33

J(B,C)=5/11​≈0.45

SMC(A,B)=6/8=0.75

SMC(A,C)=4/8=0.5

SMC(B,C)=5/8=0.625

**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 are defined as data in which the number of features (variables observed), p, are close to or larger than the number of observations (or data points).

An example of high-dimensional data in biological sciences may include data collected from hospital patients recording symptoms, blood test results, behaviours, and general health, resulting in datasets with large numbers of features.

Issues that arise with high dimensional data are:

* Running a risk of overfitting the machine learning model.
* Difficulty in clustering similar features.
* Increased space and computational time complexity.

Way to deal with high dimension data:

* Choose to include fewer features. The most obvious way to avoid dealing with high dimensional data is to simply include fewer features in the dataset. ...
* Use a regularization method.

**9. Make a few quick notes on:**

PCA is an acronym for Principal Component Analysis.

2. Use of vectors: The result of running PCA on the set of points in the diagram consist of 2 vectors called eigenvectors which are the principal components of the data set. The size of each eigenvector is encoded in the corresponding eigenvalue and indicates how much the data vary along the principal component.

3. Embedded technique: An embedding layer maps each entity to a unique vector, causing the layer's memory requirement to be proportional to the number of entities. In the recommendation domain, a given category can have hundreds of thousands of entities, and its embedding layer can take gigabytes of memory.

**10. Make a comparison between:**

1. Sequential backward exclusion vs. sequential forward selection

Forward selection starts with a (usually empty) set of variables and adds variables to it, until some stop- ping criterion is met. Similarly, backward selection starts with a (usually complete) set of variables and then excludes variables from that set, again, until some stopping criterion is met.

2. Function selection methods: filter vs. wrapper

The main differences between the filter and wrapper methods for feature selection are: Filter methods measure the relevance of features by their correlation with dependent variable while wrapper methods measure the usefulness of a subset of feature by actually training a model on it

1. SMC vs. Jaccard coefficient

The Simple Matching Coefficient (SMC) and the Jaccard coefficient are both measures of similarity between two binary variables. The SMC counts both mutual presence and mutual absence as a match, while the Jaccard index only counts the presence of each other as a match