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

* Feature engineering is the process that takes raw data and transforms it into features that can be used to create a predictive model using machine learning or statistical modeling, such as deep learning.
* Feature engineering in ML consists of four main steps: Feature Creation, Transformations, Feature Extraction, and Feature Selection. Feature engineering consists of creation, transformation, extraction, and selection of features, also known as variables, that are most conducive to creating an accurate Machine Learning algorithm.

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 method of reducing the input variable to your model by using only relevant data and getting rid of noise in data. It is the process of automatically choosing relevant features for your machine learning model based on the type of problem you are trying to solve.
* Feature selection methods are intended to reduce the number of input variables to those that are believed to be most useful to a model in order to predict the target variable. Feature selection is primarily focused on removing non-informative or redundant predictors from the model.
* These are some feature selection methods :
* Chi-square Test.
* Fisher's Score.
* Correlation Coefficient.
* Dispersion ratio
* Backward Feature Elimination.
* Recursive Feature Elimination.
* Random Forest Importance.

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

* 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.
* Filter methods are much faster compared to wrapper methods as they do not involve training the models.

4.

i. **Describe the overall feature selection process**.

* Feature Selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data. It is the process of automatically choosing relevant features for your machine learning model based on the type of problem you are trying to solve

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

* Though PCA is a very useful technique to extract only the important features but should be avoided for supervised algorithms as it completely hampers the data. If we still wish to go for Feature Extraction Technique then we should go for LDA instead.

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

* Text classification is a machine learning technique that assigns a set of predefined categories to open-ended text. Text classifiers can be used to organize, structure, and categorize pretty much any kind of text – from documents, medical studies and files, and all over the web.

6. **What makes cosine similarity a good metric for text categorization?**

* The cosine similarity is advantageous because even if the two similar documents are far apart by the Euclidean distance because of the size (like, the word 'cricket' appeared 50 times in one document and 10 times in another) they could still have a smaller angle between them. Smaller the angle, higher the similarity.

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.

7.

1. **What is the formula for calculating Hamming distance?**

In order to calculate the Hamming distance between two strings, and, we perform their XOR operation, (a⊕ b), and then count the total number of 1s in the resultant string.

**Between 10001011 and 11001111, calculate the Hamming gap.**

* 10001011⊕ 11001111 = 01000100

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).

8**. State what is meant by "high-dimensional data set"?**

* 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), n.

**Could you offer a few real-life examples?**

* Data on health status of patients can be high-dimensional (100+ measured/recorded parameters from blood analysis, immune system status, genetic background, nutrition, alcohol- tobacco- drug-consuption, operations, treatments, diagnosed diseases,)..
* Wafer detection data can be high-dimensional

**What are the difficulties in using machine learning techniques on a data set with many dimensions?**

* 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.

What can be done about it?

* One can solve it by using PCA or feature selection

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

PCA is an acronym for Personal Computer Analysis.

* False, PCA is an acronym for Principal Component Analysis. It is a popular unsupervised learning technique for reducing the dimensionality of data.

2. Use of vectors

* In machine learning, feature vectors are used to represent numeric or symbolic characteristics, called features, of an object in a mathematical, easily analyzable way.

3. **Embedded technique:**

* The word embedding techniques are used to represent words mathematically. One Hot Encoding, TF-IDF, Word2Vec, FastText are frequently used Word Embedding methods. One of these techniques (in some cases several) is preferred and used according to the status, size and purpose of processing the data

10. **Make a comparison between:**

1. **Sequential backward exclusion vs. sequential forward selection**

* Sequential floating forward selection (SFFS) starts from the empty set. After each forward step, SFFS performs backward steps as long as the objective function increases. Sequential floating backward selection (SFBS) starts from the full set.

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

3. **SMC vs. Jaccard coefficient**

* Thus, the SMC counts both mutual presences (when an attribute is present in both sets) and mutual absence (when an attribute is absent in both sets) as matches and compares it to the total number of attributes in the universe, whereas the Jaccard index only counts mutual presence as matches and compares it to the .