1. **What are the three stages to build the hypotheses or model in machine learning?**

The three stages to build the hypotheses or model in machine learning are:

1. **Model building** – Select the model on this basis of the target variable.
2. **Model testing, Evaluation and Optimization**– Test the model for minimizing loss function (bias-variance error) by selection or deletion or modification of features(columns), Train-Test split and modifying parameters (or hyper-parameters). This is done through various performance metrics based on type of ML algorithm(model) used.
3. **Deploying the model**
4. **What is the standard approach to supervised learning?**

The standard approach to supervised learning is to Identify predictors and target variables, split the dataset in consideration, into the training set and the test set. It is done because the dataset provides the facility to test one’s model with the data present in the dataset. Build the model using train set data and evaluate the model using test set data.

1. **What is Training set and Test set?**

**Training set**: The set of data that is used to discover the potentially predictive relationship between input and outputs of a model is known as ‘Training Set’. It is a set of known inputs and outputs used to train the machine. It lets the model(algorithm) identify patterns in the data and create rules for predicting the output.

**Test set**: The set of data that is used to check how well the model works after it is trained with the training set is known as ‘Test set’. It is used to check the accuracy of the hypotheses generated by the model as well as other performance metrics after learning from the training set data.

Training set must be distinct from the Test set. Test set should a similar distribution as that of training set.

1. **What is the general Principle of an ensemble Method and What is bagging and boosting in ensemble method?**

**Ensemble method** is combining of two or more machine learning algorithms or models to achieve (or predict) a better result than what each model could achieve when used individually. The models can be based on different algorithms or same algorithms used again with different parameters.

**The general principle of an ensemble method** is to combine the predictions of several models built with a given learning algorithm to improve robustness over a single model. This is achieved by bagging and boosting.

**Bagging** is an abbreviation for "bootstrap aggregating". It's meta-algorithm, which takes M subsamples (with replacement) from the initial dataset and trains the predictive model on those subsamples. The final model is obtained by averaging the "bootstrapped" models and usually yields better results.

**Boosting** is a method that builds weak learners into strong ensembles by focusing iterations on the error terms of the previous iteration. Boosting is used sequentially to reduce the bias of the combined model.

1. **How can you avoid overfitting?**

Overfitting can be avoided by using the following methods:

1. **Using a lot of training data**: By using a dataset containing a lot of training data, overfitting can be avoided. Overfitting happens when you have a small dataset, and you try to learn from it.
2. **Cross Validation:** Ifyou have a small dataset and you are forced to come with a model based on that. In such situation, one can use cross validationto avoid overfitting. There are different methods of cross validation:
3. **Validation Set**: The dataset is split into, testing, validation and training datasets, the testing dataset will only test the model while, in training dataset, the datapoints will come up with the model and validation set is used to optimize the model.
4. **K-Fold Cross validation**: Validates the model by creating k-folds and training the model on k-1 folds and testing it on the kth fold and calculating the score for it. And then it repeats the same process for different folds in the training dataset. The standard deviation of the scores should be as minimal as possible.
5. **Stratified k-fold cross validation**: Same as k-fold CV but takes into consideration the high variation and discrepancy in the training dataset to reduce overfitting.
6. **Adversarial cross validation**
7. **Regularization:**  Using models such as Lasso, Ridge and Elasticnet
8. **Using Feature Engineering and Feature Selection methods**