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

Ans-The three main stages of creating a model in machine learning are-

1. Model building- Here model are prepared using proper algorithm based on requirement and then train it with training dataset. Which build the model.
2. Model Testing- Here the model are tested with test dataset to find the accuracy of the model and tune some hyper parameter if needed.
3. Applying Model- Here model are applied to the real dataset to predict the value.

2. What is the standard approach to supervised learning?

Ans. When we dataset in which we have features or input data and labels or output too, then we use supervised learning to build the model. In supervised learning underlying model get input features and corresponding output , it then compare the output generated by it and correct output to find error and then it modifies the model according to the error . So supervised learning is used where historical data is used to predict future output.

There are several types of supervised learning-

Classification- When we want to classify a data into different classes then classification algorithms are used . For example if we want to classify a product in good or bad product classes based on user reviews then we will use classification algorithms.

Regression-In regression the future value is predicted using linear relationship in input and output. These are used in case of stock price predictions.

Anomaly detection- In the cases where we want to find the points where data are unusual like in fraud detection in case of credit card transactions. It simply learns how the normal data should be and then identify anything beyond normal values.

3. What is Training set and Test set?

Ans- The main aspect of machine learning is making model to understand the data and learn from it ,so when we have a dataset to build a model then from same we need to train and test our model. So we have two subsets training dataset and test dataset

Training Dataset- This the dataset we use to train our model and it is the important dataset because from this set only model understands the relations which it use to predict results in future. So from data set the specific features are selected which are capable of making our model understand the relations. These features are provided model for training and known as training dataset ,usually we provide major portion of our dataset as training data set

Test Dataset- After we have trained our model it is required to test our model so that we can get to know how efficient our model is, for that we provide we provide a subset of main dataset as test dataset. These data are remaining data which are left after taking out training dataset.

4. What is the general principle of an ensemble method and what is bagging and

boosting in ensemble method?

Ans.- The general principle of an ensemble method is that it to combine the prediction of several smaller model to increase performance of the model. The small models are known as base model and there different ways to combine these small models.

Bagging- It is the process in which several estimators or models are built independently using subsets of the main data and the predictions of these are averaged thats result in a better efficiency.

The random forest has uses the bagging process.

Boosting- In boosting the estimators or models are built in a sequence, here the first weak estimator is built and next estimator is built on weakness of the previous one , so these weak models are combined to give a better model. xgboost is an example of boosting

5. How can you avoid overfitting ?

Ans- Overfitting is case where our model is trained too well, i.e. when a huge number of data is supplied which contains lots of noises and outliers. So during training model try to lean from noises too which can impact the accuracy of the model and model with such impact is called over fitted model. So it is very important to avoid it.

The most common ways to avoid overfitting are-

Cross validation- It is one the most important method to remove over fitting .In this the dataset is first divided in different segments suppose D1-Dn ,now in first loop the first segment D1 is kept separately and remaining segments D2-Dn are used to train the model and then separated segment D1 is used to test the model. In next loop the second D2 is kept separately and D1,D3-Dn are used to train the model after that separated segment D2 is used to test the model .This process are repeated n times. Using this we can train data more efficiently.

Early stopping –It states the rules which guide to find the number of repetition which can be performed without causing over fit , so it is the number of repletion after which the model starts overfitting.

Pruning- This is the method where the nodes which are very less capable in predicting the result are dropped or removed.

Regularization- The basic use of this is to decide the correct complexity of model,many times a simple model does not work and at the same time more complex model my fail too so regularization helps to decide the correct model. It adds a cost term to add more features with the function , so it force coefficient of many variables to 0.