1. What is the definition of a target function? In the sense of a real-life example, express the target function. How is a target function's fitness assessed?

Ans: **Target function is used to predict the results accurately based on the data it was fed. For instance, While playing chess with the opponent, when opponent will play then the machine learning algorithm will decide what be the number of possible legal moves taken in order to get success. If the function predicts the correct result which results in success then it is the target function otherwise not.**

1. What are predictive models, and how do they work? What are descriptive types, and how do you use them? Examples of both types of models should be provided. Distinguish between these two forms of models.

Ans: **predictive models are used to predict the values or class of the data. Regression and classification models are used for predictions depending upon the property of the target variable. Descriptive models use historic report to get insights about the data. We can find what might have caused the event to take place in the past, but predictive models are used to predict the values.**

3. Describe the method of assessing a classification model's efficiency in detail. Describe the various measurement parameters.

Ans: **confusion metric is used in classification for assessing the model efficiency. Other terms like accuracy, recall, precision, F-beta score, ROC and AUC is also used.**

1. In the sense of machine learning models, what is underfitting? What is the most common reason for underfitting?

Ans: **underfitting when the model doesn’t fit the train data. It occurs when the model is too simple for the data.**

1. What does it mean to overfit? When is it going to happen?

Ans: **when the model tries to remember all of the training data values and doesn’t perform well with unseen data. It happens when the model is too complex for the data.**

1. In the sense of model fitting, explain the bias-variance trade-off.

**Bias is training error and variance is testing error. Trade off happens when one increases and the other one decreases. We should choose a model which doesn’t either overfit or underfit the data.**

5. Is it possible to boost the efficiency of a learning model? If so, please clarify how.

Ans: **yes, through feature engineering and feature selection, the model’s efficiency can be boosted.**

**Feature engineering involves data imputation, removal of duplicate data, removal of outliers, deal with imbalanced data etc.**

6. How would you rate an unsupervised learning model's success? What are the most common success indicators for an unsupervised learning model?

Ans: **clustering performance evaluation metrics is the most common evaluation metrics.**

7. Is it possible to use a classification model for numerical data or a regression model for categorical data with a classification model? Explain your answer.

Ans: **using regression model for classification would misclassify the data which in turn, would give less accuracy. Whereas, classification algorithms can be used for numerical data if they are converted into bins to categories/range**

8. Describe the predictive modeling method for numerical values. What distinguishes it from categorical predictive modeling?

Ans: **predictive modelling for numerical values is called regression and it uses to predict the continuous variable only. Classification predictive modelling is used for predicting the category of the data it may belong to. The performance metrics for both is different. For regression, we use R2 and adjusted R2 and for classification we use confusion matrix.**

9. The following data were collected when using a classification model to predict the malignancy of a group of patients' tumors:

i. Accurate estimates – 15 cancerous, 75 benign

ii. Wrong predictions – 3 cancerous, 7 benign

Determine the model's error rate, Kappa value, sensitivity, precision, and F-measure.

**Error rate = 0.1**

**Sensitivity =0.6**

**Precision = 0.8**

**F-measure = 2**

10. Make quick notes on:

1. The process of holding out:

Ans: **the data set is divided into three parts : train, validation and test data set. Then we try to find the best hyperparameters for the model. K fold cross validation is one type of this hold out method.**

2. Cross-validation by tenfold: **in this train dataset is divided into 10 subsets of train and validation set. Where in each iteration, the model is trained with K-1 set and tested on the remaining subset.**

3. Adjusting the parameters: **parameters are adjusted in an attempt to minimise the error in a model. We often use hyperparameters tuning to find the best hyperparameters in order to find parameters.**

11. Define the following terms:

1. Purity vs. Silhouette width:

**Purity in decision tree is calculated in terms of Gini or Entropy value. A node is considered a pure split when it is a leaf node or all of its data belong to only one class. In the case of purity the value of Gini Index and Entropy is 0. Silhouette width In clustering is a measure of how similar an object to its own cluster compared to the neighbouring cluster.**

2. Boosting vs. Bagging

Ans: **Boosting and Bagging are ensemble techniques which use homogenous ML algorithms. Bagging uses a parallel approach to learn from models whereas, Boosting uses sequential approach to learn from the models.**

3. The eager learner vs. the lazy learner

Ans: **eager learner algorithms learn from the data and creates a generalise model which can be used to predict the values for the unseen data. Lazy learner doesn’t learn from the data and rather stores the data for prediction. The examples of lazy learner is KNN where the train data is stored and then the class of the new data is found based on the distance of the new data and stored data.**