**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?**

The target function is basically the goal or objective of a problem, typically it is related to input variables.

In context to the real-life example of the machine learning problem of classification of the images of handwritten digits into categories.

Choice of target heavily depends upon specific problem, different problem having different target variable, like minimizing error, maximizing profit.

**2. 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.**

*Predictive models* are statistical or machine learning model that are design to make predictions. These model analyze historical patterns and relationships within the data and generates predictions. Working of predictive models are as follows:

* Data collections
* Data preprocessing
* Feature Selections/Engineering
* Model training
* Model evaluations
* Model depeloyement

Examples of predictive model: Predicting price of houses and corresponding features such as size, location, and numbers of bedrooms and a linear regression model can be used to predict the price of the house.

*Descriptive Model* are models that are aim to summarize and describe patterns, realationships within the data. These models doesn’t focus on making predictions. Working of these models are as follows:

* Data explorations
* Data visualisations
* Statistical analysis
* Pattern Recognitions

Cluster analysis can be good example

The main difference is that the descriptive model aims to summarize and describe the patterns and relationships within the data, whereas the predictive model aims to predict outcomes based on the data.

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

Assessing the classification model's efficiency involves evaluating performance in accurately classifying instances or observations into their respective categories.

* Accuracy- It measures the proportion of correctly classified objects to the total objects.
* Precision- It is a proporation of true positive out of all objects. It focus on correctness of positive predictions.
* Recall- It measures all positive objects out of all actual positive values
* F1-score-It’s the harmonic mean of precision and recall.

**4.**

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

Underfitting is when a model fails to capture the underlying patterns and relationships in the data, resulting in poor performance on both the training and test datasets. It occurs when the model is too simple. The most common reason for underfitting is using a model that is too basic or has insufficient complexity to learn from the data, such as using a linear model for highly nonlinear data.

* **What does it mean to overfit? When is it going to happen?**

Overfitting occurs when a model performs exceedingly well on the training data but fails to generalize well to new or unseen data. It happens when the model becomes too complex. Overfitting tends to happen when the model has a high variance, meaning it is overly sensitive to the training data and captures both the true patterns and the noise.

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

The bias-variance trade-off in model fitting refers to the relationship between the model's ability to capture the true underlying patterns in the data (bias) and its sensitivity to variations in the training data (variance). A model with high bias tends to oversimplify the data, leading to underfitting, while a model with high variance is overly complex and captures noise, resulting in overfitting.

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

Yes, it is possible, some techniques are:

* Feature engineering
* Hyperparameter
* Ensemble methods
* Cross-Validation

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

When the success of an unsupervised learning model, the evaluation metrics, and indicators can differ from those used in supervised learning. Some common success indicators are: -

* Clustering Quality
* Reconstruction error
* Anomaly Detection
* Visualization and interpretability

**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.**

Reasons are as follows:

Classification models are designed to predict discrete categorical classes. They are not suitable for handling numerical data as the classification model have predefined categories or limited catagories.

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

A predictive model for numerical values is commonly referred to as regression model. Regresssion model are designed to predict a continuous numeric target variable based on input data. Few points that can distinguish the predictive numerical model from categorical predictive model-

* Target variable
* Model output
* Evaluation Metrics
* Model selection
* Handling input varaible

**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.**

Model error rate – (False positive + False Negative)/Total = (7 + 3)/(15+75+7+30) = 0.1% or 10%

Kappa Value –

Sensitivity – TP/(TP+FN) = 83.3%

Precision – TP/(TP + FP) = 68.2%

F-measure – 2\*(Precision \* Sensitivity) / (Precision + Sensitivity) = 75.1%

**10. Make quick notes on:**

**1. The process of holding out**

Practice reserving a portion of the available data as a validation or test set, while using the remaining data for training a model. The held-out data is typically not used during the training process and is used later to assess the model's performance on unseen data. This process helps evaluate the model's generalization ability and detect potential overfitting.

**2. Cross-validation by tenfold**

It is a technique used to assess the performance and generalizability of a model. Tenfold cross-validation involves dividing the data into ten equally sized subsets or folds. The model is trained and evaluated ten times, with each fold serving as a validation set once while the remaining nine folds are used for training. This process is repeated for all folds, and the performance metrics are averaged across the ten iterations. Tenfold cross-validation provides a robust estimate of model performance by utilizing all data for both training and validation.

**3. Adjusting the parameters**

It involves tuning the hyperparameters of a model to optimize its performance. Hyperparameters are settings or configurations that are set before training the model and are not learned from the data. Examples of hyperparameters include learning rate, regularization strength, number of hidden layers in a neural network, or the depth of a decision tree. By adjusting the hyperparameters, such as through grid search or random search, different combinations can be explored to find the best configuration that yields the highest performance or minimizes a chosen evaluation metric.

**11. Define the following terms:**

**1. Purity vs. Silhouette width**

**Purity:** Purity is a measure used to evaluate the quality of clustering in unsupervised learning. It quantifies how well the clusters contain instances of the same class.

**Silhouette Width:** Silhouette width is another metric used to assess the quality of clustering. It measures both the cohesion of data points within their own cluster and the separation between clusters.

**2. Boosting vs. Bagging**

**Boosting:** Boosting is an ensemble learning technique where multiple weak learners, such as decision trees, are combined to create a strong learner. The learners are trained sequentially, with each subsequent learner focusing on the instances that were misclassified or given higher weights in the previous iterations.

**Bagging**: Bagging, short for bootstrap aggregating, is an ensemble learning technique. It involves training multiple independent learners on different random subsets of the training data. Each learner is trained independently, and the final prediction is obtained by aggregating the predictions of all the individual learners. It helps reduce variance and improve the accuracy of the model

**3. The eager learner vs. the lazy learner**

**Eager Learner**: It is a machine learning algorithm that constructs a model during the training phase and eagerly uses it to make predictions on unseen data. Eager learners typically build a general model that captures the relationships between input features and the target variable. Examples include decision trees, support vector machines (SVM), and neural networks.

**Lazy Learner**: Also known as an instance-based learner or a lazy algorithm, defers the model construction until the prediction phase. Instead of building a generalized model during training, lazy learners memorize the training instances and their associated target values. When making predictions, lazy learners compare the new instance to the stored training instances to determine the most similar instances and use their target values to make predictions. Examples of lazy learners include k-nearest neighbors (KNN)