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, also known as the hypothesis or model, represents the relationship between input variables (features) and the desired output (target) in a machine learning problem.

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: These models make predictions about future outcomes based on patterns learned from historical data. Example: Predicting customer churn based on past behavior.

Descriptive Models: These models describe patterns and relationships in data but don't necessarily make predictions. Example: Clustering customers into segments based on purchasing behavior.

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

>>>Accuracy: Percentage of correct predictions.

Precision: Proportion of true positive predictions among all positive predictions.

Recall (Sensitivity): Proportion of true positive predictions among all actual positive instances.

F1-Score: Harmonic mean of precision and recall.

4.

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

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

>>>Underfitting: When a model is too simple to capture the underlying patterns in the data, leading to poor performance on both training and test data.

Overfitting: When a model is too complex and fits the training data's noise, leading to excellent performance on training data but poor performance on new, unseen data.

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

>>>Bias-Variance Trade-off: Balancing the model's bias (assumptions) and variance (sensitivity to noise) to achieve the best generalization.

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

>>>Boosting involves combining weak models (e.g., decision trees) into a strong model.

Weak models are iteratively trained, giving more weight to misclassified instances.

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

>>>Success indicators include:

Clustering quality: How well data points within clusters are similar, and different clusters are distinct.

Visualization: Effective visualization of data clusters or patterns.

Interpretability: Clear interpretation of cluster characteristics.

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.

>>>It's not advisable to use a classification model for numerical data or a regression model for categorical data.

Classification models predict discrete classes, while regression models predict continuous values.

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

>>>Predictive modeling for numerical values (regression) aims to predict a continuous output.

For categorical values, predictive modeling involves classification to assign instances to predefined categories.

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

>>>>>>>>Error Rate: (3 + 7) / (15 + 75) = 0.1

Kappa Value: Calculate using the confusion matrix.

Sensitivity (Recall): 15 / (15 + 3) = 0.8333

Precision: 15 / (15 + 7) = 0.6818

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

10. Make quick notes on:

Holding Out: Withholding part of the data for validation while training a model on the rest.

Tenfold Cross-Validation: Splitting data into ten parts and using each part as a validation set while training on the other nine.

Adjusting Parameters: Tuning hyperparameters of a model to achieve optimal performance.

11. Define the following terms:

1. Purity vs. Silhouette width

>>>Purity measures the proportion of instances in a cluster that belong to the majority class.

Silhouette Width measures how similar an object is to its own cluster compared to other clusters.

2. Boosting vs. Bagging

>>>Boosting combines weak models into a strong model by giving more weight to misclassified instances.

Bagging creates multiple bootstrapped training sets and trains models on each, then combines their predictions.

3. The eager learner vs. the lazy learner

>>>Eager Learner precomputes a general model from the training data before receiving new instances.

Lazy Learner defers computation until new instances are presented, making predictions based on stored instances.