Basic Concepts

1. **What is Machine Learning?**
   - **Answer:** Machine Learning is a subset of artificial intelligence that allows systems to learn and improve from experience without being explicitly programmed.
2. **What are the types of Machine Learning?**
   - **Answer:** The three main types are supervised learning, unsupervised learning, and reinforcement learning.
3. **Define supervised learning.**
   - **Answer:** Supervised learning involves training a model on a labeled dataset, where the desired output is known.
4. **What is unsupervised learning?**
   - **Answer:** Unsupervised learning involves training a model on a dataset without labeled responses, aiming to find hidden patterns or intrinsic structures.
5. **Explain reinforcement learning.**
   - **Answer:** Reinforcement learning is a type of learning where an agent learns to make decisions by taking actions in an environment to maximize cumulative reward.
6. **What is overfitting?**
   - **Answer:** Overfitting occurs when a model learns the training data too well, including noise and outliers, leading to poor performance on new, unseen data.
7. **What is underfitting?**
   - **Answer:** Underfitting occurs when a model is too simple to capture the underlying patterns in the data, resulting in poor performance on both training and test data.
8. **Explain the bias-variance tradeoff.**
   - **Answer:** The bias-variance tradeoff is the balance between the error introduced by bias (error from erroneous assumptions) and variance (error from sensitivity to small fluctuations in the training set).
9. **What is a confusion matrix?**
   - **Answer:** A confusion matrix is a table used to evaluate the performance of a classification model, displaying true positives, false positives, true negatives, and false negatives.
10. **What is the purpose of cross-validation?**
    - **Answer:** Cross-validation is used to assess how well a model generalizes to an independent dataset, helping to prevent overfitting.

Data Preprocessing

11. **What is feature scaling and why is it important?**

- **Answer:** Feature scaling is the process of normalizing or standardizing features to bring them to a similar scale, which improves the performance and convergence speed of many machine learning algorithms.

12. **What is one-hot encoding?**
    - **Answer:** One-hot encoding is a process of converting categorical variables into a binary matrix, where each category is represented by a separate binary column.
13. **What is the purpose of data splitting in machine learning?**
    - **Answer:** Data splitting involves dividing the dataset into training and testing sets to evaluate the model's performance on unseen data.
14. **Explain the concept of dimensionality reduction.**
    - **Answer:** Dimensionality reduction involves reducing the number of input variables in a dataset to simplify the model, reduce computation time, and avoid overfitting.
15. **What is PCA (Principal Component Analysis)?**
    - **Answer:** PCA is a dimensionality reduction technique that transforms data into a set of orthogonal components, capturing the maximum variance in the data.

Specific Algorithms

16. **Describe the k-nearest neighbors (k-NN) algorithm.**
    - **Answer:** k-NN is a non-parametric algorithm used for classification and regression, where the output is determined by the k-nearest neighbors to a given data point.
17. **What is linear regression?**
    - **Answer:** Linear regression is a regression algorithm that models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data.
18. **Explain logistic regression.**
    - **Answer:** Logistic regression is a classification algorithm that models the probability of a binary outcome using a logistic function.
19. **What is a decision tree?**
    - **Answer:** A decision tree is a model that makes decisions based on a series of questions or conditions, represented as a tree structure, with nodes representing tests on features and branches representing the outcomes.
20. **Describe the random forest algorithm.**
    - **Answer:** Random forest is an ensemble learning algorithm that creates multiple decision trees and merges their predictions to improve accuracy and prevent overfitting.
21. **What is gradient boosting?**

- **Answer:** Gradient boosting is an ensemble technique that builds models sequentially, with each new model correcting errors made by the previous models, to improve overall prediction accuracy.

22. **Explain support vector machines (SVM).**
    - **Answer:** SVM is a supervised learning algorithm that finds the optimal hyperplane that maximally separates data points of different classes in a high-dimensional space.
23. **What is the kernel trick in SVM?**
    - **Answer:** The kernel trick allows SVM to perform non-linear classification by implicitly mapping input features into a higher-dimensional space using kernel functions.
24. **What are neural networks?**
    - **Answer:** Neural networks are computational models inspired by the human brain, consisting of layers of interconnected nodes (neurons) that learn to recognize patterns and make predictions.
25. **What is deep learning?**
    - **Answer:** Deep learning is a subset of machine learning that uses neural networks with many layers (deep networks) to model complex patterns and representations in data.

Evaluation Metrics

26. **What is accuracy?**
    - **Answer:** Accuracy is the ratio of correctly predicted instances to the total instances in a dataset, often used as a performance metric for classification models.
27. **Define precision and recall.**
    - **Answer:** Precision is the ratio of true positive predictions to the total predicted positives, while recall is the ratio of true positive predictions to the total actual positives.
28. **What is F1 score?**
    - **Answer:** F1 score is the harmonic mean of precision and recall, providing a balance between the two metrics, especially useful for imbalanced datasets.
29. **Explain ROC curve and AUC.**
    - **Answer:** The ROC curve plots the true positive rate against the false positive rate for different threshold values, while AUC (Area Under the Curve) measures the overall performance of the classifier.
30. **What is mean squared error (MSE)?**
    - **Answer:** MSE is a regression metric that calculates the average of the squared differences between predicted and actual values, indicating the model's prediction accuracy.
31. **What is Root Mean Squared Error (RMSE)?**

- **Answer:** RMSE is the square root of the mean squared error and provides a measure of the standard deviation of the prediction errors.

32. **What is Mean Absolute Error (MAE)?**
    - **Answer:** MAE is the average of the absolute differences between predicted and actual values, measuring the average magnitude of the errors.
33. **What is R-squared?**
    - **Answer:** R-squared is a statistical measure that represents the proportion of variance in the dependent variable explained by the independent variables in a regression model.
34. **What is Adjusted R-squared?**
    - **Answer:** Adjusted R-squared adjusts the R-squared value based on the number of predictors in the model, providing a more accurate measure of model performance.
35. **What is Log Loss?**
    - **Answer:** Log Loss, or logarithmic loss, measures the performance of a classification model where the prediction is a probability value between 0 and 1, penalizing false classifications.

Model Selection and Tuning

36. **What is hyperparameter tuning?**
    - **Answer:** Hyperparameter tuning involves selecting the best set of hyperparameters for a machine learning model to improve its performance.
37. **What is GridSearchCV?**
    - **Answer:** GridSearchCV is a technique that performs an exhaustive search over a specified parameter grid for the best hyperparameters using cross-validation.
38. **What is RandomSearchCV?**
    - **Answer:** RandomSearchCV is a technique that randomly samples a subset of the hyperparameter space and evaluates model performance to find the best parameters.
39. **Explain the concept of model validation.**
    - **Answer:** Model validation involves assessing a model's performance on an independent dataset to ensure it generalizes well to new, unseen data.
40. **What is a learning curve?**
    - **Answer:** A learning curve is a plot that shows the model's performance on the training and validation datasets over different sizes of the training set, helping to diagnose bias and variance issues.
41. **What is a validation curve?**

- **Answer:** A validation curve is a plot that shows the model's performance on the training and validation datasets over different values of a hyperparameter.

42. **What is k-fold cross-validation?**
    - **Answer:** k-fold cross-validation involves dividing the dataset into k equally-sized subsets, training the model k times, each time using a different subset as the test set and the remaining k-1 subsets as the training set.
43. **What is leave-one-out cross-validation?**
    - **Answer:** Leave-one-out cross-validation is a special case of k-fold cross-validation where k equals the number of data points, meaning each data point is used once as the test set while the remaining points are used as the training set.
44. **What is stratified cross-validation?**
    - **Answer:** Stratified cross-validation is a variation of k-fold cross-validation where the folds are made by preserving the percentage of samples for each class, ensuring that each fold is representative of the overall class distribution.
45. **What is the purpose of a validation set?**
    - **Answer:** A validation set is used to tune model hyperparameters and prevent overfitting during the training process by providing an independent dataset for model evaluation.
46. **What is early stopping in machine learning?**
    - **Answer:** Early stopping is a regularization technique used to prevent overfitting by stopping the training process when the model's performance on the validation set starts to degrade.
47. **What is model ensembling?**
    - **Answer:** Model ensembling involves combining multiple models to improve overall performance by reducing variance, bias, or improving predictions.
48. **Explain bagging and boosting.**
    - **Answer:** Bagging (Bootstrap Aggregating) involves training multiple models on different subsets of the data and combining their predictions, while boosting involves training models sequentially, each focusing on the errors of the previous model.
49. **What is stacking?**
    - **Answer:** Stacking is an ensemble technique where multiple models are trained, and their predictions are combined using another model (meta-learner) to improve overall performance.
50. **What is a blending ensemble?**
    - **Answer:** Blending is similar to stacking, but it involves using a holdout set to train the meta-learner instead of cross-validation.

Feature Engineering

51. **What is feature engineering?**
    - **Answer:** Feature engineering involves creating new features or transforming existing features to improve model performance.
52. **What is feature selection?**
    - **Answer:** Feature selection is the process of selecting a subset of relevant features for use in model training, reducing dimensionality and improving model performance.
53. **Explain the difference between filter, wrapper, and embedded methods for feature selection.**
    - **Answer:** Filter methods use statistical techniques to select features, wrapper methods use a predictive model to evaluate feature subsets, and embedded methods perform feature selection during the model training process.
54. **What is principal component analysis (PCA)?**
    - **Answer:** PCA is a dimensionality reduction technique that transforms data into a set of orthogonal components, capturing the maximum variance in the data.
55. **What is linear discriminant analysis (LDA)?**
    - **Answer:** LDA is a dimensionality reduction technique that projects data onto a lower-dimensional space to maximize class separability.
56. **What is mutual information?**
    - **Answer:** Mutual information measures the amount of information gained about one variable by knowing another variable, often used for feature selection.
57. **What is the difference between univariate and multivariate feature selection?**
    - **Answer:** Univariate feature selection evaluates features individually based on their relationship with the target variable, while multivariate feature selection considers interactions between features.
58. **What is feature importance?**
    - **Answer:** Feature importance is a measure of the impact of each feature on the model's predictions, often used for feature selection and model interpretation.
59. **What is feature extraction?**
    - **Answer:** Feature extraction involves creating new features from existing data using techniques like PCA, LDA, and autoencoders.
60. **What is a polynomial feature?**
    - **Answer:** Polynomial features are created by raising existing features to a power, enabling linear models to capture non-linear relationships.