1. In the CRISP-DM methodology, what is the primary goal of the "Business Understanding" phase?

a) Assessing the scalability of the model
b) Reviewing the final deployment strategy
c) Defining the project objectives and requirements
d) Preprocessing and cleaning the dataset

Answer: c) Defining the project objectives and requirements
Explanation: The Business Understanding phase focuses on determining the goals and objectives of the project, ensuring alignment with business needs.

---

2. Which preprocessing technique ensures all features contribute equally to the analysis?

a) One-hot encoding
b) MinMax scaling
c) Standardization
d) Aggregation

Answer: c) Standardization
Explanation: Standardization scales numeric features to have a mean of 0 and a standard deviation of 1, ensuring equal contribution to the analysis.

---

3. What is the key assumption in a Naive Bayes classifier?

a) Features are independent given the class label
b) The target variable must be continuous
c) The dataset must be free of noise
d) The features must follow a uniform distribution

Answer: a) Features are independent given the class label
Explanation: Naive Bayes assumes that all features are conditionally independent given the class, which simplifies computation.

---

4. In a Decision Tree, which metric is commonly used to measure impurity?

a) Euclidean Distance
b) Gini Index
c) Jaccard Coefficient
d) Silhouette Score

Answer: b) Gini Index
Explanation: The Gini Index measures the impurity in a node, with lower values indicating purer splits.

---

5. In clustering, what does K-means aim to minimize?

a) The total sum of squared errors within clusters
b) The distance between clusters
c) The number of clusters
d) The variance of the dataset

Answer: a) The total sum of squared errors within clusters
Explanation: K-means minimizes the sum of squared distances between each point and its assigned cluster centroid.

---

6. What does "Schema on read" mean in a data lake architecture?

a) The schema is created at the time of data ingestion
b) Data is stored in raw format and structured when read
c) The schema is defined before storing the data
d) The schema is inferred from external databases

Answer: b) Data is stored in raw format and structured when read
Explanation: "Schema on read" allows raw data to remain unstructured until it is accessed and processed, providing flexibility.

---

7. What does the coefficient of determination R2R2 measure in regression?

a) The absolute error of the predictions
b) The percentage of variance explained by the model
c) The correlation between input and output variables
d) The likelihood of overfitting

nswer: b) The percentage of variance explained by the model
Explanation: indicates how well the independent variables explain the variation in the dependent variable.

---

8. What is the main purpose of feature selection in machine learning?

a) To reduce the number of irrelevant or redundant features
b) To ensure all features are numeric

c) To optimize hyperparameters of the model
d) To create synthetic features

Answer: a) To reduce the number of irrelevant or redundant features
Explanation: Feature selection focuses on selecting the most relevant features to improve model performance and reduce complexity.

---

9. Which of the following clustering methods is density-based?

a) K-means
b) DBSCAN
c) Hierarchical Clustering
d) Expectation Maximization

Answer: b) DBSCAN
Explanation: DBSCAN identifies clusters based on density, allowing it to find clusters of arbitrary shapes.

---

10. In association rule mining, what does the "confidence" of a rule represent?

a) The likelihood that the rule is correct
b) The percentage of transactions where the rule is valid
c) The ratio of the rule's support to the support of its antecedent
d) The difference between the predicted and actual results

Answer: c) The ratio of the rule's support to the support of its antecedent
Explanation: Confidence measures the conditional probability of the consequent given the antecedent.

---

11. What is a common sign of overfitting in a machine learning model?

a) High training accuracy and low test accuracy
b) High accuracy on both training and test sets
c) Low accuracy on the training set
d) Low variance in predictions

Answer: a) High training accuracy and low test accuracy
Explanation: Overfitting occurs when the model performs well on the training set but poorly on unseen data, indicating poor generalization.

---

12. What does the "MinMax scaling" technique achieve in data preprocessing?

a) Removes outliers
b) Maps data to a fixed range without altering distributions
c) Converts numerical data to categorical data
d) Reduces the dimensionality of the dataset

Answer: b) Maps data to a fixed range without altering distributions
Explanation: MinMax scaling ensures that all features are rescaled to a specific range, such as [0, 1].

---

13. Which metric is NOT suitable for evaluating classification models?

a) Silhouette Score
b) F1-Score
c) Accuracy
d) Precision

Answer: a) Silhouette Score

---

14. What does the Apriori algorithm focus on in association rule mining?

a) Identifying outliers in the dataset
b) Pruning infrequent itemsets early in the process
c) Grouping transactions into clusters
d) Calculating confidence for classification

Answer: b) Pruning infrequent itemsets early in the process

---

15. What is the curse of dimensionality in clustering?

a) Increased effectiveness of Euclidean distance in high dimensions
b) Difficulty in distinguishing points as dimensions increase
c) Reduced computational complexity in higher dimensions
d) Higher clustering accuracy in high-dimensional spaces

Answer: b) It makes it harder to distinguish points as dimensions increase
Explanation: The curse of dimensionality refers to the phenomenon where data points become equidistant in high-dimensional spaces, reducing clustering effectiveness.

---

16. Which clustering evaluation metric measures intra-cluster similarity?

a) Silhouette Score
b) Jaccard Coefficient
c) Gini Index
d) Entropy

Answer: a) Silhouette Score

---

17. In CRISP-DM, what activity is part of the "Evaluation" phase?

a) Determining model scalability
b) Deploying the model to production
c) Assessing model performance against business objectives
d) Conducting feature engineering

Answer: c) Assessing model performance against business objectives

---

18. What is the role of Laplace smoothing in Naive Bayes?

a) To avoid zero probabilities for unseen attribute values
b) To increase computational efficiency
c) To normalize data distributions
d) To map attributes to the same scale

Answer: a) To avoid zero probabilities for unseen attribute values
Explanation: Laplace smoothing adjusts probabilities to ensure no attribute value results in zero probability, improving robustness.

---

19. In a dataset with DD attributes, how many subsets must be considered in exhaustive feature selection?

a) D2D2
b) 2D2D
c) D!D!
d) D×log⁡(D)D×log(D)

Answer: b) 2D2D

---

20. What is an advantage of DBSCAN over K-means?

a) Requires fewer parameters
b) Can identify clusters of varying densities
c) Guarantees optimal clustering
d) Uses hierarchical techniques

Answer: b) Can identify clusters of varying densities

---

21. Which of the following is a property of a valid distance metric?

a) Symmetry
b) Non-negativity
c) Triangle inequality
d) All of the above

Answer: d) All of the above
Explanation: A valid distance metric must satisfy all three properties: non-negativity (distance cannot be negative), symmetry (distance between A and B equals distance between B and A), and the triangle inequality (the sum of two sides of a triangle is always greater than or equal to the third side).

22. What is the primary goal of Principal Component Analysis (PCA)?

a) Reduce the dimensionality of the dataset while retaining variance
b) Normalize all features to the same scale
c) Cluster the dataset into predefined groups
d) Predict the target variable

Answer: a) Reduce the dimensionality of the dataset while retaining variance
Explanation: PCA transforms the dataset into a smaller set of uncorrelated components while retaining as much variance from the original data as possible, which is useful for simplifying complex datasets.

23. Which of the following is NOT an advantage of Random Forest over Decision Trees?

a) Reduced risk of overfitting
b) Ability to handle missing values directly
c) Improved interpretability
d) Robustness to noise

Answer: c) Improved interpretability
Explanation: While Random Forest reduces overfitting and is robust to noise, it sacrifices interpretability because it combines multiple trees, making it harder to analyze the model's decision-making process.

24. In regression, which metric measures the average magnitude of prediction errors?

a) Mean Squared Error (MSE)
b) Root Mean Squared Error (RMSE)
c) Coefficient of Determination (R2R2)
d) Mean Absolute Error (MAE)

Answer: d) Mean Absolute Error (MAE)
Explanation: MAE calculates the average of the absolute differences between predicted and actual values, providing a straightforward measure of prediction error.

25. What is a primary drawback of using K-means for clustering?

a) It cannot handle large datasets
b) It requires the number of clusters to be predefined
c) It is computationally expensive
d) It cannot handle numerical data

Answer: b) It requires the number of clusters to be predefined
Explanation: K-means clustering requires the user to specify the number of clusters (kk) before running the algorithm, which can be difficult to determine in advance.

26. Which technique is commonly used to prevent overfitting in machine learning models?

a) Increasing the model complexity
b) Data augmentation
c) Ignoring outliers
d) Reducing the training data size

Answer: b) Data augmentation
Explanation: Data augmentation expands the training dataset by creating modified versions of existing data, helping the model generalize better and reducing the risk of overfitting.

---

27. In classification, what does the ROC curve represent?

a) Precision vs. Recall
b) True Positive Rate vs. False Positive Rate
c) Accuracy vs. False Negative Rate
d) True Negative Rate vs. False Negative Rate

Answer: b) True Positive Rate vs. False Positive Rate
Explanation: The ROC (Receiver Operating Characteristic) curve visualizes the trade-off between the True Positive Rate and False Positive Rate for different classification thresholds, helping evaluate model performance.

---

28. What is the effect of pruning a Decision Tree?

a) It increases the model's training accuracy
b) It simplifies the tree to prevent overfitting
c) It increases the tree's depth
d) It reduces the number of input features

Answer: b) It simplifies the tree to prevent overfitting
Explanation: Pruning removes branches that provide little to no contribution to the model's predictive power, making the tree less complex and reducing overfitting.

---

29. What is the role of information gain in a Decision Tree?

a) To select the attribute that best splits the data
b) To evaluate the accuracy of the tree
c) To identify missing values in the dataset
d) To reduce the dimensionality of the dataset

Answer: a) To select the attribute that best splits the data
Explanation: Information gain measures the reduction in entropy or uncertainty, helping identify the attribute that will create the most homogenous child nodes when used for splitting.

---

30. In DBSCAN, what parameter defines the minimum number of points required to form a cluster?

a) Epsilon (єє)
b) MinPts
c) MaxDepth
d) Support

Answer: b) MinPts
Explanation: MinPts specifies the minimum number of data points that must exist within the $\epsilon\epsilon$-radius of a core point for it to form a cluster.

---

31. Why is cross-validation used in machine learning?

a) To reduce the dimensionality of the dataset
b) To prevent overfitting and evaluate model performance on unseen data
c) To ensure all features are included in the final model
d) To optimize hyperparameters

Answer: b) To prevent overfitting and evaluate model performance on unseen data
Explanation: Cross-validation splits the dataset into training and validation sets multiple times, ensuring the model's performance is assessed on data it hasn't seen during training.

---

32. What does the Gini Index measure in classification?

a) The correlation between features
b) The distance between classes
c) The impurity of a dataset
d) The probability of overfitting

Answer: c) The impurity of a dataset
Explanation: The Gini Index measures how mixed the classes are within a node, with lower values indicating purer nodes in a Decision Tree.

---

33. Which similarity metric is best suited for text data?

a) Cosine Similarity
b) Euclidean Distance
c) Manhattan Distance
d) Jaccard Coefficient

Answer: a) Cosine Similarity
Explanation: Cosine similarity is widely used for text data, as it measures the cosine of the angle between two vectors, focusing on direction rather than magnitude.

---

34. Which metric can be used to evaluate clustering performance?

a) Silhouette Score
b) RMSE

c) F1-Score
d) ROC-AUC

Answer: a) Silhouette Score
Explanation: The Silhouette Score evaluates how well clusters are separated and how close each point is to the centroid of its assigned cluster.

---

35. What does the term "lift" represent in association rule mining?

a) The frequency of an itemset
b) The improvement in prediction accuracy due to a rule
c) The ratio of observed support to expected support under independence
d) The distance between two rules

Answer: c) The ratio of observed support to expected support under independence
Explanation: Lift measures how much more likely two items are to occur together than if they were independent, helping evaluate the significance of association rules.

---

36. What is the key output of feature selection in data preprocessing?

a) A reduced dataset with only relevant features
b) A normalized dataset
c) A dataset with all missing values filled
d) A synthetic dataset with generated features

Answer: a) A reduced dataset with only relevant features
Explanation: Feature selection reduces the dataset to include only the most relevant features, improving model performance and reducing complexity.

---

37. What is "entropy" in the context of Decision Trees?

a) The uncertainty or impurity of a dataset
b) The number of branches in the tree
c) The depth of the tree
d) The measure of model accuracy

Answer: a) The uncertainty or impurity of a dataset
Explanation: Entropy quantifies the level of disorder or impurity in the dataset, with higher entropy indicating greater uncertainty about class labels.

---

38. Which clustering algorithm can detect non-spherical clusters?

a) K-means
b) DBSCAN
c) PCA
d) Agglomerative Clustering

Answer: b) DBSCAN
Explanation: DBSCAN can detect clusters of arbitrary shapes, including non-spherical ones, making it suitable for complex datasets.

---

39. What is the main purpose of normalization in data preprocessing?

a) To ensure all features are on a similar scale
b) To identify and remove duplicate entries
c) To create synthetic features
d) To reduce the dataset's size

Answer: a) To ensure all features are on a similar scale
Explanation: Normalization scales numeric features to a specific range, such as [0,1], ensuring fair contributions during model training.

---

40. Which method is suitable for evaluating a regression model's performance?

a) F1-Score
b) RMSE
c) Silhouette Score
d) Precision

Answer: b) RMSE
Explanation: Root Mean Squared Error (RMSE) measures the average magnitude of prediction errors in regression models, providing an interpretable metric in the same units as the target variable.

41. In which scenario is stratified sampling particularly useful?

a) When the dataset is small
b) When there are missing values in the dataset
c) When the dataset is imbalanced across classes
d) When the features are highly correlated

Answer: c) When the dataset is imbalanced across classes
Explanation: Stratified sampling ensures that the proportion of each class in the training and test sets matches the original dataset, which is especially important for imbalanced datasets.

---

42. What does the term "support" mean in association rule mining?

a) The fraction of transactions that contain a given itemset
b) The difference between observed and expected occurrences of an itemset
c) The confidence level of a rule
d) The statistical significance of a rule

Answer: a) The fraction of transactions that contain a given itemset
Explanation: Support measures how frequently an itemset appears in the dataset, making it a fundamental metric for mining frequent itemsets.

---

43. What is the main assumption of linear regression?

a) The dependent variable is binary
b) There is a linear relationship between the independent and dependent variables
c) The dataset must contain only categorical variables
d) The data must follow a uniform distribution

Answer: b) There is a linear relationship between the independent and dependent variables
Explanation: Linear regression assumes a linear relationship between the predictors and the target variable, allowing it to model and predict outcomes.

---

44. Which evaluation metric is most appropriate for an imbalanced classification problem?

a) Accuracy
b) Precision-Recall AUC
c) Mean Squared Error
d) Entropy

Answer: b) Precision-Recall AUC
Explanation: Precision-Recall AUC focuses on correctly identifying the positive class, which is crucial in imbalanced datasets where accuracy can be misleading.

---

45. What is the role of the epsilon (єє) parameter in DBSCAN?

a) Defines the maximum number of clusters
b) Specifies the radius for neighborhood points to be considered part of a cluster
c) Determines the minimum number of clusters in the dataset
d) Sets the distance metric for calculating centroids

Answer: b) Specifies the radius for neighborhood points to be considered part of a cluster
Explanation: The $\epsilon\epsilon$ parameter in DBSCAN defines the maximum distance within which points are considered neighbors for clustering.

---

46. What does "bagging" in machine learning refer to?

a) Combining weak classifiers to create a strong classifier
b) Resampling the dataset to create multiple models
c) Selecting the best features for a model
d) Reducing dimensionality of the dataset

Answer: b) Resampling the dataset to create multiple models
Explanation: Bagging (Bootstrap Aggregating) involves resampling the dataset to train multiple models and combining their outputs to improve stability and accuracy.

---

47. In CRISP-DM, which phase involves evaluating the final model against business objectives?

a) Deployment
b) Evaluation
c) Data Preparation
d) Modeling

Answer: b) Evaluation
Explanation: In the Evaluation phase, the model's performance is assessed to ensure it meets the business objectives defined during the Business Understanding phase.

---

48. What is a common limitation of hierarchical clustering?

a) It requires a predefined number of clusters
b) It cannot handle non-numerical data
c) It is computationally expensive for large datasets
d) It only works with dense datasets

Answer: c) It is computationally expensive for large datasets
Explanation: Hierarchical clustering's complexity grows with the number of data points, making it unsuitable for very large datasets.

---

49. Which property does the Manhattan distance emphasize compared to Euclidean distance?

a) Squared differences between coordinates
b) Absolute differences between coordinates
c) Weighted differences between coordinates
d) Maximum difference between any two coordinates

Answer: b) Absolute differences between coordinates
Explanation: Manhattan distance calculates the sum of absolute differences between the coordinates of two points, making it robust to outliers.

---

50. What is the primary goal of feature engineering?

a) To create new features that improve model performance
b) To select the most relevant features from the dataset
c) To reduce the size of the dataset
d) To normalize data values

Answer: a) To create new features that improve model performance
Explanation: Feature engineering involves creating new, meaningful features from the existing data to enhance the predictive power of a model.

---

51. What is the main advantage of ensemble methods like Random Forest?

a) They are computationally less expensive
b) They reduce variance and increase stability
c) They are easier to interpret
d) They require no hyperparameter tuning

Answer: b) They reduce variance and increase stability
Explanation: Ensemble methods combine multiple models to average out errors, reducing variance and improving generalization.

---

52. What does the term "lift" signify in association rule mining?

a) The correlation between items in a rule
b) The improvement in prediction accuracy due to the rule
c) The ratio of observed support to expected support
d) The probability of a rule being correct

Answer: c) The ratio of observed support to expected support
Explanation: Lift measures how much more likely two items are to occur together than if they were independent, indicating the strength of an association.

---

53. Why is MinMax scaling applied during data preprocessing?

a) To convert categorical variables into numerical format
b) To map all numeric values into a specified range
c) To reduce the dimensionality of the data
d) To increase the variance of features

Answer: b) To map all numeric values into a specified range
Explanation: MinMax scaling ensures that all numeric features are scaled to the same range, such as [0, 1], without altering their distributions.

---

54. Which algorithm is most suitable for clustering data with arbitrary shapes?

a) K-means
b) DBSCAN
c) Hierarchical Clustering
d) K-Nearest Neighbors

Answer: b) DBSCAN
Explanation: DBSCAN is effective for clustering datasets with arbitrary shapes because it does not assume spherical clusters like K-means.

---

55. What does "Entropy" measure in decision trees?

a) The accuracy of predictions
b) The impurity or uncertainty of a dataset
c) The depth of the tree
d) The number of nodes in the tree

Answer: b) The impurity or uncertainty of a dataset
Explanation: Entropy quantifies the disorder or impurity in a dataset, helping to decide the best splits in a decision tree.

---

56. Which technique is used to reduce dimensionality while retaining variance?

a) PCA
b) Random Forest
c) DBSCAN
d) Bagging

Answer: a) PCA
Explanation: Principal Component Analysis (PCA) reduces the number of dimensions in a dataset while retaining as much variance as possible.

---

57. What is the primary difference between K-means and Hierarchical Clustering?

a) K-means requires predefined clusters; hierarchical clustering does not
b) Hierarchical clustering is faster than K-means
c) K-means is density-based, while hierarchical is centroid-based
d) Hierarchical clustering does not require a distance metric

Answer: a) K-means requires predefined clusters; hierarchical clustering does not
Explanation: K-means needs the number of clusters as input, whereas hierarchical clustering builds a tree structure without requiring predefined clusters.

---

58. What is a leaf node in a Decision Tree?

a) A node that performs a split
b) A node that contains a class label or prediction
c) A node with the highest impurity
d) A node that has multiple child nodes

Answer: b) A node that contains a class label or prediction
Explanation: Leaf nodes are terminal nodes in a Decision Tree that provide the final class label or prediction for a given path.

---

59. Which metric is commonly used for model evaluation in regression?

a) Precision
b) Recall
c) Root Mean Squared Error (RMSE)
d) Gini Index

Answer: c) Root Mean Squared Error (RMSE)
Explanation: RMSE measures the average magnitude of prediction errors in regression models, providing an interpretable evaluation metric.

---

60. What does cross-validation primarily help with?

a) Selecting the most important features
b) Reducing computational cost during training
c) Preventing overfitting and ensuring robust evaluation
d) Reducing variance in predictions

Answer: c) Preventing overfitting and ensuring robust evaluation
Explanation: Cross-validation splits the data into training and validation sets multiple times to ensure the model generalizes well to unseen data.