1. **Question**: What is the main reason for the standardization of numeric attributes?
    - A) To assign binary values to all attributes
    - B) To ensure each feature contributes equally to the analysis
    - C) To convert all attributes to categorical data
    - D) To increase the range of values for each attribute

    Answer: B) To ensure each feature contributes equally to the analysis

2. **Question**: Which of the following is NOT a property of a metric distance function?
    - A) Non-negativity
    - B) Symmetry
    - C) Triangle Inequality
    - D) Boundedness

    Answer: D) Boundedness

3. **Question**: In a decision tree, what does pruning aim to eliminate?
    - A) Branches with the highest information gain
    - B) Parts of the tree where decisions could be influenced by random effects
    - C) The root of the tree
    - D) All leaf nodes

    Answer: B) Parts of the tree where decisions could be influenced by random effects

4. **Question**: Which of the following measures is used to compute the similarity between two binary vectors according to the Jaccard Coefficient?
    - A) The sum of the matching entries
    - B) The ratio of matching entries to total entries
    - C) The ratio of the intersection to the union of the sets
    - D) The difference between the number of 1s and 0s

    Answer: C) The ratio of the intersection to the union of the sets

5. **Question**: What is the main purpose of the MinMax scaling (also known as “rescaling”) of attributes?
    - A) To map all numeric attributes to the same range without altering distribution
    - B) To increase the variance of the data
    - C) To convert numeric attributes into categorical ones
    - D) To normalize the data distribution

    Answer: A) To map all numeric attributes to the same range without altering distribution

6. **Question**: What does the Gini Index measure in a dataset?
    - A) The accuracy of a classifier.
    - B) The impurity of a dataset.
    - C) The correlation between variables.

- D) The distance between clusters.

Answer: B) The impurity of a dataset.

7. **Question**: Which of the following is an alternative measure to the Information Gain in decision trees?
   - A) Accuracy
   - B) Precision
   - C) Gini Index
   - D) Recall

   Answer: C) Gini Index

8. **Question**: In a decision tree, the number of objects in a node...
   - A) Is always equal to the number of objects in its ancestor.
   - B) Is smaller than the number of objects in its ancestor.
   - C) Increases as we move down the tree.
   - D) Is independent of the number of objects in its ancestor.

   Answer: B) Is smaller than the number of objects in its ancestor.

9. **Question**: What is a base hypothesis for a Bayesian classifier?
   - A) The dataset must be large.
   - B) The attributes must be statistically independent inside each class.
   - C) The target variable should follow a normal distribution.
   - D) All features must be equally important.

   Answer: B) The attributes must be statistically independent inside each class.

10. **Question**: When changing the clustering scheme, how are the total sum of squared errors and separation related?

- A) They are inversely proportional.
- B) They are directly proportional.
- C) If one increases, the other decreases.
- D) They are not related.

Answer: C) If one increases, the other decreases.

11. **Question**: Which of the following statements is true about k-means clustering?

- A) It always converges to the configuration with the minimum distortion for the chosen number of clusters.
- B) It is not efficient for large datasets.
- C) It is sensitive to the initial assignment of the centers.
- D) It guarantees the optimal number of clusters.

Answer: C) It is sensitive to the initial assignment of the centers.

12. **Question**: Which of the following statements is true about DBSCAN clustering?

- A) It cannot handle clusters with concavities.
- B) It always stops at a configuration including at least one cluster.
- C) Increasing the radius of the neighborhood can decrease the number of noise points.
- D) It assigns all points to a cluster, leaving no noise points.

Answer: C) Increasing the radius of the neighborhood can decrease the number of noise points.

13. **Question**: What does the statement “the support is anti-monotone” imply in association rule mining?

- A) The support of an itemset always exceeds the support of its subsets.
- B) The support of an itemset is equal to the support of its subsets.
- C) The support of an itemset never exceeds the support of its subsets.
- D) The support of an itemset is not related to the support of its subsets.

Answer: C) The support of an itemset never exceeds the support of its subsets.

14. **Question**: What does the coefficient of determination (R2) indicate in a linear regression model?

- A) The ratio of the sum of squares of residuals to the total sum of squares.
- B) The percentage of the variance in the dependent variable that is predictable from the independent variables.
- C) The absolute value of the correlation between the independent and dependent variables.
- D) The average distance of the data points from the regression line.

Answer: B) The percentage of the variance in the dependent variable that is predictable from the independent variables.

15. **Question**: What is the primary objective of K-means clustering?

- A) To maximize the average distance between data points in different clusters.
- B) To minimize the sum of the squared distances of each point to its centroid.
- C) To find the optimal number of clusters in a dataset.
- D) To ensure that each cluster has an equal number of data points.

Answer: B) To minimize the sum of the squared distances of each point to its centroid.

16. **Question**: In the CRISP-DM methodology, which of the following activities is part of "Business Understanding"?

- A) Determining the available resources (manpower, hardware, software, etc.)
- B) Conducting the initial data collection
- C) Performing the data modeling
- D) Evaluating the final results

Answer: A) Determining the available resources (manpower, hardware, software, etc.)

17. **Question**: Which of the following statements is true regarding outliers and noise in data?

- A) Outliers can never be a result of noise.
- B) Noise in the data cannot generate outliers.
- C) Outliers can be due to noise, and noise can generate outliers.
- D) Noise and outliers are always unrelated phenomena in a dataset.

Answer: C) Outliers can be due to noise, and noise can generate outliers.

18. **Question**: In which mining activity is Information Gain most useful?

- A) Clustering
- B) Association rule mining
- C) Classification
- D) Regression

Answer: C) Classification

19. **Question**: What is cross-validation?

- A) A technique to combine different models into one
- B) A method to estimate the performance of a model on new, unseen data
- C) The process of dividing the dataset into equal parts for training and testing
- D) A technique to ensure all features contribute equally to the model

Answer: B) A method to estimate the performance of a model on new, unseen data

20. **Question**: Which preprocessing activity is useful for building a Naive Bayes classifier if the independence hypothesis is violated?

- A) Normalization
- B) Dimensionality reduction
- C) Feature selection
- D) Data augmentation

Answer: C) Feature selection

21. **Question**: What is the main reason for MinMax scaling of attributes?

- A) To normalize the distribution of the attributes.
- B) To map all numeric attributes to the same range.
- C) To convert categorical attributes to numeric.
- D) To increase the variance of the attributes.

Answer: B) To map all numeric attributes to the same range.

22. **Question**: What is the primary objective of feature selection in data preprocessing?

- A) To select the features with the highest range and influence.
- B) To reduce the computational complexity of the model.
- C) To increase the accuracy of the model by including more features.
- D) To ensure all features contribute equally to the model's performance.

Answer: B) To reduce the computational complexity of the model.

23. **Question**: Which distance function is best suited for Boolean data?

- A) Euclidean Distance
- B) Manhattan Distance
- C) Jaccard Coefficient
- D) Cosine Distance

Answer: C) Jaccard Coefficient

24. **Question**: What is a common symptom of overfitting when developing a classifier?

- A) The error rate in the test set is much lower than in the training set.
- B) The error rate in the test set is much greater than in the training set.
- C) The model performs equally well on the training and test sets.
- D) The model's error rate is consistently high in both training and test sets.

Answer: B) The error rate in the test set is much greater than in the training set.

25. **Question**: What is true about Hierarchical Agglomerative Clustering?

- A) It does not require the definition of the number of clusters.
- B) It is primarily used for large datasets.
- C) It requires the definition of distance between sets of objects.
- D) It is based on the principle of maximizing the distance between clusters.

Answer: C) It requires the definition of distance between sets of objects.

26. **Question**: Which of the following is not an objective of aggregating attributes in data preprocessing?

- A) To obtain a more detailed description of data.
- B) To reduce the variability of data.
- C) To obtain a less detailed scale.
- D) To reduce the number of attributes or distinct values.

Answer: A) To obtain a more detailed description of data.

27. **Question**: What best describes the strategy of Apriori in finding frequent itemsets?

- A) Evaluating the support of itemsets in random order.
- B) Evaluating all possible itemsets regardless of their support.
- C) Evaluation of the support of the itemsets in an order that prunes uninteresting parts of the search space as soon as possible.
- D) Using a distance-based approach to evaluate itemsets.

Answer: C) Evaluation of the support of the itemsets in an order that prunes uninteresting parts of the search space as soon as possible.

28. **Question**: What is an advantage of Multi-Dimensional Scaling (MDS) over Principal Component Analysis (PCA)?

- A) MDS is faster than PCA.
- B) MDS can be used also with categorical data, while PCA is limited to vector spaces.
- C) MDS always produces more accurate results than PCA.
- D) MDS is easier to interpret than PCA.

Answer: B) MDS can be used also with categorical data, while PCA is limited to vector spaces.

29. **Question**: Which of the following is different from the others? Decision tree, K-means, Expectation Maximization, Apriori.

- A) Decision tree
- B) K-means
- C) Expectation Maximization

- D) Apriori

Answer: D) Apriori (It's an association rule mining algorithm, while others are supervised and unsupervised learning methods)

30. **Question**: Which of the following is different from the others? Decision tree, K-means, Expectation Maximization, DBSCAN.

- A) Decision tree
- B) K-means
- C) Expectation Maximization

- D) DBSCAN

Answer: A) Decision tree (It's a supervised learning method, while others are clustering algorithms)

31. **Question**: Which of the following is different from the others? DBSCAN, SVM, Neural Network, Decision Tree.

- A) DBSCAN
- B) SVM
- C) Neural Network
- D) Decision Tree

Answer: A) DBSCAN (It's a clustering method, while others are classification methods)

32. **Question**: Which of the following is different from the others? Silhouette Index, Misclassification Error, Gini Index, Entropy.

- A) Silhouette Index
- B) Misclassification Error
- C) Gini Index
- D

- D) Entropy

Answer: A) Silhouette Index (It's an index for evaluating clustering performance, while others are measures for classification purity or error)

33. **Question**: What is the effect of pruning in generating frequent itemsets?

- A) It evaluates all possible itemsets for maximum frequency.
- B) It ensures that only the largest itemsets are considered.
- C) If an itemset is not frequent, then none of its supersets can be frequent, therefore the frequencies of the supersets are not evaluated.
- D) It selects itemsets based on their contribution to overall data variance.

Answer: C) If an itemset is not frequent, then none of its supersets can be frequent, therefore the frequencies of the supersets are not evaluated.

34. **Question**: What measure is maximized by the Expectation Maximization algorithm for clustering?

- A) The distance between clusters.
- B) The likelihood of a class label, given the values of the attributes of the example.
- C) The number of clusters.
- D) The homogeneity of each cluster.

Answer: B) The likelihood of a class label, given the values of the attributes of the example.

35. **Question**: What is the primary use of information gain in data mining?

- A) To select the attribute which maximizes the ability to predict the class value for a given training set.
- B) To determine the number of clusters in a dataset.
- C) To evaluate the performance of a regression model.
- D) To normalize the data attributes.

Answer: A) To select the attribute which maximizes the ability to predict the class value

36. **Question**: What is the effect of normalization in data preparation?

- A) To convert all variables to a nominal scale.
- B) To map all numeric attributes to the same range, without altering the distribution.
- C) To increase the variance of numeric attributes.
- D) To convert all attributes to binary values.

Answer: B) To map all numeric attributes to the same range, without altering the distribution.

37. **Question**: Which of the following clustering methods is not based on distances between objects?

- A) K-means
- B) Hierarchical Clustering
- C) DBSCAN
- D) Expectation Maximization

Answer: D) Expectation Maximization (It's based on probability distributions rather than distances)

38. **Question**: In a dataset with D attributes, how many subsets of attributes should be considered for feature selection according to an exhaustive search?

- A) D
- B) D^2
- C) 2^D
- D) D!

Answer: C) 2^D

39. **Question**: What is the effect of the curse of dimensionality?

- A) As the number of dimensions increases, the Euclidean distance becomes more effective.

- B) When the number of dimensions increases, the Euclidean distance becomes less effective to discriminate between points in the space.
- C) Increasing dimensions reduces computational complexity.
- D) High dimensions increase the accuracy of clustering algorithms.

Answer: B) When the number of dimensions increases, the Euclidean distance becomes less effective to discriminate between points in the space.

40. **Question**: What is the main purpose of smoothing in Bayesian classification?

- A) To ensure all features contribute equally to the classification.
- B) To classify an object containing attribute values which are missing from some classes in the training set.
- C) To increase the computational efficiency of the classifier.
- D) To convert categorical attributes into numeric.

Answer: B) To classify an object containing attribute values which are missing from some classes in the training set.

41. **Question**: Which characteristic of data can reduce the effectiveness of DBSCAN?

- A) Uniform density of clusters.
- B) Presence of clusters with different densities.
- C) Low dimensionality of the dataset.
- D) Small dataset size.

Answer: B) Presence of clusters with different densities.

42. **Question**: Which type of data allows the use of the Euclidean distance?

- A) Categorical data.
- B) Textual data.
- C) Points in a vector space.
- D) Time-series data.

Answer: C) Points in a vector space.

43. **Question**: What is the effect of the curse of dimensionality on Euclidean distance?

- A) It becomes more effective in high dimensions.
- B) It becomes less effective in discriminating between points in space.
- C) It becomes faster to compute in high dimensions.
- D) It becomes irrelevant in high dimensions.

Answer: B) It becomes less effective in discriminating between points in space.

44. **Question**: What are the hyperparameters of a Neural Network?

- A) Learning rate, number of layers, types of activation functions, number of epochs.
- B) The weights and biases of the neurons.
- C) The input and output data.
- D) The type of optimization algorithm used.

Answer: A) Learning rate, number of layers, types of activation functions, number of epochs.

45. **Question**: How can we measure the quality of a trained regression model?

- A) By the number of features used.
- B) By the training time.
- C) By a formula evaluating the difference between forecasted values and true ones.
- D) By the type of regression used.

Answer: C) By a formula evaluating the difference between forecasted values and true ones.

46. **Question**: What is the difference between classification and regression?

- A) Classification is unsupervised, while regression is supervised.
- B) Classification has a categorical target, while regression has a numeric target.
- C) Classification can only be used for large datasets.
- D) Regression is used exclusively for time-series data.

Answer: B) Classification has a categorical target, while regression has a numeric target.

47. **Question**: What is Principal Component Analysis (PCA) in feature selection?

- A) A technique to classify data into distinct categories.
- B) A mathematical technique to transform a set of numeric attributes into a smaller set.
- C) A method to predict the outcome of a dependent variable.
- D) A clustering technique.

Answer: B) A mathematical technique to transform a set of numeric attributes into a smaller set.

48. **Question**: What is backpropagation in a Neural Network?

- A) A method for increasing the number of layers in a network.
- B) The technique used to adjust connection weights based on the output error.
- C) A technique to reduce the size of the input data.
- D) The process of adding more neurons to each layer.

Answer: B) The technique used to adjust connection weights based on the output error.

49. **Question**: What is the main reason for normalization of numeric attributes?

- A) To convert them into categorical attributes.
- B) To map all numeric attributes to the same range without altering distribution.
- C) To increase the variance of the attributes.
- D) To remove all outliers from the dataset.

Answer: B) To map all numeric attributes to the same range without altering distribution.

50. **Question**: Which of the following is not an objective of feature selection?

- A) To improve model accuracy by removing irrelevant features.
- B) To reduce overfitting.
- C) To select features with higher range and more influence on computations.
- D) To reduce the computational complexity of the model.

Answer: C) To select features with higher range and more influence on computations.

51. **Question**: For each type of data, choose the best-suited distance function: Vector space with real values.

- A) Euclidean Distance
- B) Jaccard Coefficient
- C) Cosine Distance
- D) Manhattan Distance

Answer: A) Euclidean Distance

52. **Question**: When developing a classifier, a symptom of overfitting is:

- A) The error rate in the test set is much greater than the error rate in the training set.
- B) The classifier performs equally well on both training and test sets.
- C) The classifier underperforms on the training set.
- D) The error rate in the training set is much greater than in the test set.

Answer: A) The error rate in the test set is much greater than the error rate in the training set.

53. **Question**: In a decision tree, an attribute which is used only in nodes near the leaves indicates:

- A) High importance of the attribute.
- B) The attribute is irrelevant.
- C) Little insight with respect to the target.
- D) The attribute is a primary splitting attribute.

Answer: C) Little insight with respect to the target.

54. **Question**: Which of the following statements about Hierarchical Agglomerative Clustering is true?

- A) It does not require the definition of distance between sets of objects.
- B) It is best used for very large datasets.
- C) It requires the definition of distance between sets of objects.
- D) It always produces the most accurate clustering.

Answer: C) It requires the definition of distance between sets of objects.

55. **Question**: Match the rule evaluation formula with its name: !"#(%&'() !"# (%)

- A) Confidence
- B) Lift
- C) Leverage
- D) Conviction

Answer: A) Confidence

56. **Question**: In data preprocessing, the aggregation of attributes aims to:

- A) Obtain a more detailed description of data.
- B) Obtain a less detailed scale.
- C) Increase the variability of data.
- D) Increase the number of attributes.

Answer: B) Obtain a less detailed scale.

57. **Question**: What best describes the strategy of Apriori in finding frequent itemsets?

- A) Evaluates support in a random order.
- B) Evaluates all itemsets, regardless of support.
- C) Evaluates support in an order to prune uninteresting parts as soon as possible.
- D) Uses a distance-based method to evaluate itemsets.

Answer: C) Evaluates support in an order to prune uninteresting parts as soon as possible.

58. **Question**: What is an advantage of Multi-Dimensional Scaling (MDS) over PCA?

- A) MDS is faster than PCA.
- B) MDS can be used with categorical data, while PCA cannot.
- C) MDS is more accurate than PCA.
- D) MDS is easier to interpret than PCA.

Answer: B) MDS can be used with categorical data, while PCA cannot.

59. **Question**: Which of the following is different from the others? Decision tree, K-means, Expectation Maximization, Apriori.

- A) Decision tree
- B) K-means
- C) Expectation Maximization
- D) Apriori

Answer: D) Apriori (It's an association rule mining algorithm, while others are learning methods)

60. **Question**: Which of the following is different from the others? Decision tree, K-means, Expectation Maximization, DBSCAN.

- A) Decision tree
- B) K-means
- C) Expectation Maximization
- D) DBSCAN

Answer: A) Decision tree (It's a supervised learning method, while others are clustering algorithms)

61. **Question**: Which of the following is different from the others? DBSCAN, SVM, Neural Network, Decision Tree.

- A) DBSCAN
- B) SVM
- C) Neural Network
- D) Decision Tree

Answer: A) DBSCAN (It's a clustering method, while the others are classification methods)

62. **Question**: Which is different from the others? Silhouette Index, Misclassification Error, Gini Index, Entropy.

- A) Silhouette Index
- B) Misclassification Error
- C) Gini Index
- D) Entropy

Answer: A) Silhouette Index (It's a measure for evaluating clustering performance, while the others are used for classification)

63. **Question**: How does pruning work when generating frequent itemsets in association rule mining?

- A) If an itemset is frequent, all its supersets are evaluated.
- B) If an itemset is not frequent, then none of its supersets can be frequent.
- C) Pruning is not used in generating frequent itemsets.
- D) All itemsets are evaluated regardless of frequency.

Answer: B) If an itemset is not frequent, then none of its supersets can be frequent.

64. **Question**: What measure is maximized by the Expectation Maximization algorithm for clustering?

- A) The distance between clusters.
- B) The likelihood of a class label, given the attributes of the example.
- C) The number of clusters.
- D) The homogeneity of each cluster.

Answer: B) The likelihood of a class label, given the attributes of the example.

65. **Question**: The information gain in decision trees is used to...

- A) Determine the number of leaves in the tree.
- B) Select the attribute which maximizes the ability to predict the class value for a given training set.
- C) Calculate the depth of the tree.
- D) Estimate the error rate of the tree.

Answer: B) Select the attribute which maximizes the ability to predict the class value for a given training set.

66. **Question**: In data preparation, what is the effect of normalization?

- A) To convert all attributes to a common scale without distorting differences in the ranges.
- B) To remove outliers from the dataset.
- C) To convert numeric attributes into categorical ones.
- D) To increase the variance of the data.

Answer: A) To convert all attributes to a common scale without distorting differences in the ranges.

67. **Question**: Which of the following clustering methods is not based on distances between objects?

- A) K-means
- B) Hierarchical Clustering
- C) DBSCAN
- D) Expectation Maximization

Answer: D) Expectation Maximization (It uses probability distributions rather than distance)

68. **Question**: In a dataset with D attributes, how many subsets of attributes should be considered for feature selection according to an exhaustive search?

- A) D
- B) D^2
- C) 2^D
- D) D!

Answer: C) 2^D

69. **Question**: What is the effect of the curse of dimensionality?

- A) It makes Euclidean distances more effective in high-dimensional spaces.
- B) It reduces the computational complexity of algorithms.
- C) It makes Euclidean distances less effective in discriminating between points.
- D) It increases the accuracy of clustering algorithms.

**

Answer**: C) It makes Euclidean distances less effective in discriminating between points.

70. **Question**: What is the main purpose of smoothing in Bayesian classification?

- A) To ensure equal contribution of all features in the classification.
- B) To classify an object containing attribute values missing from some classes in the training set.
- C) To increase the computational efficiency of the classifier.
- D) To convert categorical attributes into numeric ones.

Answer: B) To classify an object containing attribute values missing from some classes in the training set.

70. **Data Understanding:**
    Question: What is the main focus during the "Data Understanding" phase of the CRISP-DM methodology?
    • A) Reviewing the deployment strategy
    • B) Initial exploration and collection of data
    • C) Identifying project risks and mitigation strategies
    • D) Formulating data-driven hypotheses
    Answer: B) Initial exploration and collection of data

71. **Data Preparation:**
    Question: In the "Data Preparation" phase of CRISP-DM, which activity is typically carried out?
    • A) Normalizing and transforming data
    • B) Setting performance metrics for the model
    • C) Documenting the project scope and objectives
    • D) Selecting tools for data visualization
    Answer: A) Normalizing and transforming data
72. **Modeling:**
    Question: Which activity is commonly undertaken in the "Modeling" phase of CRISP-DM?
    • A) Selection of appropriate modeling techniques
    • B) Outlining the deliverables and timelines
    • C) Conducting a preliminary risk analysis
    • D) Drafting an initial report of findings
    Answer: A) Selection of appropriate modeling techniques
73. **Evaluation:**
    Question: What is a crucial activity during the "Evaluation" phase of CRISP-DM?
    • A) Assessing the scalability of the model
    • B) Ensuring the integrity and quality of data
    • C) Documenting the modeling approach
    • D) Evaluating the model's performance against business objectives
    Answer: D) Evaluating the model's performance against business objectives
74. **Deployment:**
    Question: What is a key task in the "Deployment" phase of the CRISP-DM process?
    • A) Optimization of model parameters
    • B) Integration of the model into an operational environment
    • C) Exploration and integration of new data sources
    • D) Clarification and definition of the problem statement
    Answer: B) Integration of the model into an operational environment