1. What is the main purpose of the SMOTE algorithm in machine learning?

a) To reduce the number of samples in the majority class

b) To generate synthetic samples for the minority class to balance imbalanced datasets

c) To improve model performance by feature scaling

d) To perform dimensionality reduction using principal components

Correct answer: b) To generate synthetic samples for the minority class to balance imbalanced datasets

Explanation: SMOTE (Synthetic Minority Over-sampling Technique) creates synthetic minority class samples by interpolating between existing minority samples to address class imbalance.

2. Given the code snippet: `generateData = SMOTE(dataSet[, c(1,2)], dataSet[, c(3)], K=5)`, what does `dataSet[, c(3)]` represent?

a) Features for training

b) Target or class labels

c) Indices for data partitioning

d) Synthetic data generated by SMOTE

Correct answer: b) Target or class labels

Explanation: The third column represents the target variable (labels) used in the SMOTE process to generate synthetic minority samples.

3. Which method is a wrapper approach for feature selection?

a) Pearson correlation

b) Recursive Feature Elimination (RFE)

c) Principal Component Analysis (PCA)

d) Chi-square test

Correct answer: b) Recursive Feature Elimination (RFE)

Explanation: RFE uses the model to recursively remove least important features, evaluating subsets for best performance.

4. Refer to the following table comparing Pseudonymization and Anonymization:

Aspect	Pseudonymization	Anonymization
Linkability	Possible with additional info (key)	Not possible
Data Utility	High	Lower due to irreversible masking
Regulatory Status	Still considered personal data	Not considered personal data

Reversibility	Yes (with key)	No

Which statement correctly differentiates the two?

a) Pseudonymization completely removes all identifiers irreversibly

b) Anonymization allows re-identification with a key

c) Pseudonymization retains high utility and is reversible with a key

d) Both considered non-personal under GDPR

Correct answer: c) Pseudonymization retains high utility and is reversible with a key

Explanation: Pseudonymization replaces identifiers but allows re-identification using a key; anonymization removes identifiers irreversibly.

5. Which of the following is a potential downside of using open data in data science projects?

a) Open data always provides complete and bias-free samples

b) Open data can dominate research due to abundance and availability

c) Open data requires paid subscriptions and strict licensing

d) Open data is always formatted for ease of use

Correct answer: b) Open data can dominate research due to abundance and availability

Explanation: The vast amount of open data can bias research focus toward popular datasets, potentially at the expense of other topics.

6. What does algorithmic bias refer to in machine learning?

a) Hardware errors in data

b) Systematic unfair outputs due to biased data or design

c) Intentional manipulation of results

d) Random noise in predictions

Correct answer: b) Systematic unfair outputs due to biased data or design

Explanation: Algorithmic bias arises when models produce discriminatory or unfair results due to skewed training data or design flaws.

7. In hypothesis testing, what does a p-value less than 0.05 generally indicate?

a) Over 5% probability that the null hypothesis is true

b) Less than 5% probability that results are due to chance if null is true

c) The alternative hypothesis is definitely true

d) Sample size is too small

Correct answer: b) Less than 5% probability that results are due to chance if null is true
Explanation: A p-value < 0.05 indicates the observed data are unlikely under the null hypothesis, leading to its rejection at 5% significance.

8. Refer to the confusion matrix:

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

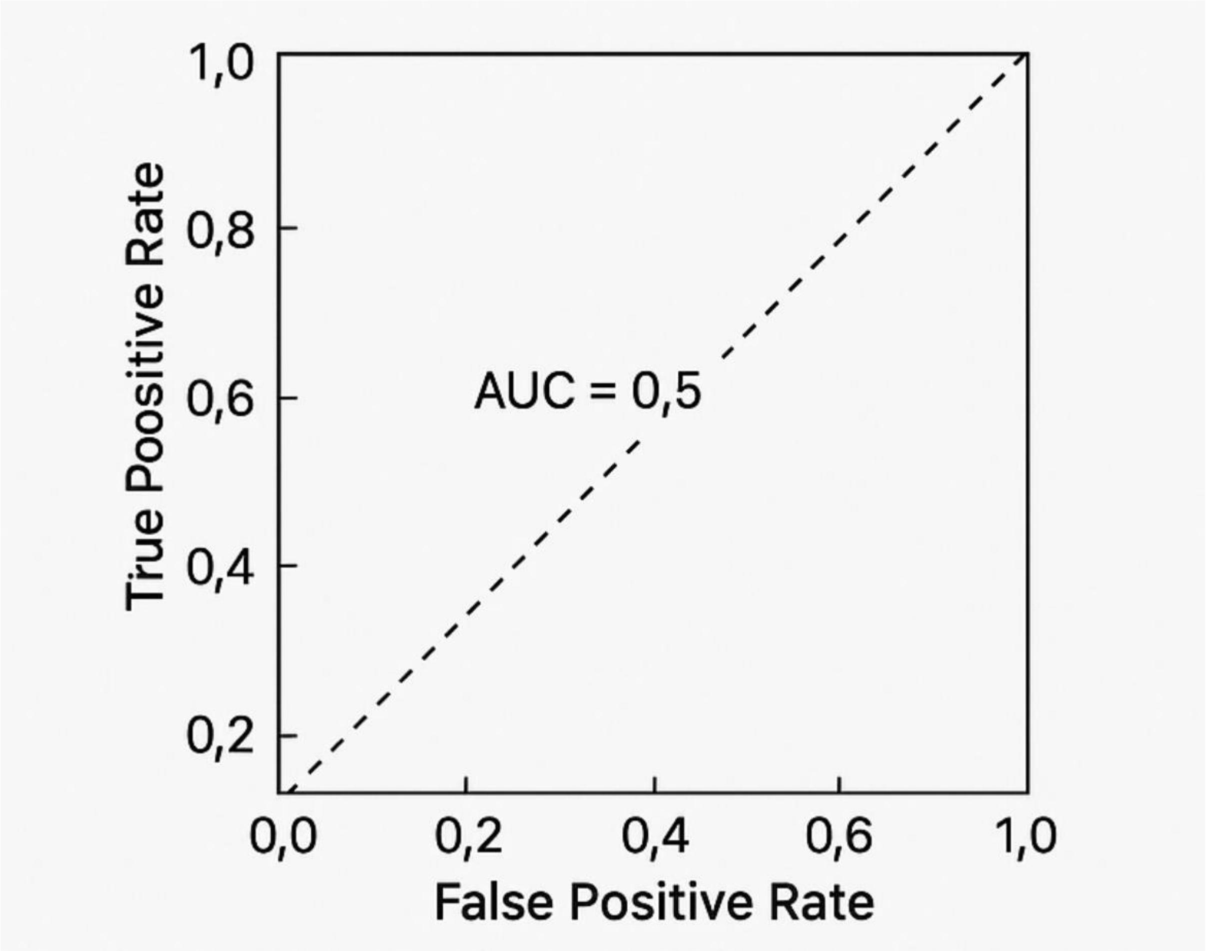
What does “False Negative” mean?
a) Predicting positive when actually negative
b) Predicting negative when actually positive
c) Correct prediction of negative
d) Correct prediction of positive

Correct answer: b) Predicting negative when actually positive
Explanation: A false negative is an instance where the model incorrectly predicts negative but the actual class is positive.

9. Calculate the entropy of a dataset S where 75% belong to class A and 25% to class B. Use Entropy(S) = – ∑ p_i * log2(p_i)
a) 0.81
b) 0.5
c) 1.0
d) 0.0

Correct answer: a) 0.81
Explanation: Entropy = –(0.75*log2(0.75) + 0.25*log2(0.25)) ≈ 0.811 bits.

10. Refer to the ROC curve diagram showing True Positive Rate vs. False Positive Rate. What trade-off is observed when changing the classification threshold?



a) Increasing sensitivity decreases false positives

b) Increasing sensitivity increases false positives

c) False positives remain constant as sensitivity increases

d) Sensitivity does not depend on threshold

Correct answer: b) Increasing sensitivity increases false positives

Explanation: Lowering the threshold typically increases true positives but also raises false positives, showing a trade-off.

11. What is the key difference between Bagging and Boosting methods?

a) Bagging trains sequentially; Boosting trains in parallel

b) Bagging trains in parallel; Boosting trains sequentially focusing on errors

c) Both use random forests identically

d) Boosting works only for regression

Correct answer: b) Bagging trains in parallel; Boosting trains sequentially focusing on errors

Explanation: Bagging reduces variance by training independent models; boosting trains models sequentially focusing on earlier mistakes.

12. Which feature selection approach does the Boruta algorithm use?

a) Filter based on Pearson correlation

b) Wrapper method with random forests and shadow features

c) Dimensionality reduction with PCA

d) Embedded method in logistic regression

Correct answer: b) Wrapper method with random forests and shadow features

Explanation: Boruta compares actual features against permuted shadow features using Random Forest importance to select relevant variables.

13. In R, what does `cor.test(x, y, method = "pearson")` do?

a) Two-sample t-test

b) Calculates Pearson correlation coefficient and tests significance

c) Performs linear regression

d) Computes covariance

Correct answer: b) Calculates Pearson correlation coefficient and tests significance

Explanation: cor.test computes the Pearson correlation and performs hypothesis testing whether correlation differs from zero.

14. What is the purpose of Variance Inflation Factor (VIF) in regression modeling?

a) Detect multicollinearity among predictors

b) Test overall model fit

c) Measure residual errors

d) Predict dependent variable values

Correct answer: a) Detect multicollinearity among predictors

Explanation: VIF quantifies how much a predictor's variance is inflated by correlation with other predictors, warning of multicollinearity.

15. Why is Adjusted R-squared preferred over R-squared in multiple regression?

a) R-squared always decreases with added variables

b) Adjusted R-squared penalizes excess predictors to avoid overfitting

c) Adjusted R-squared is easier to compute

d) They are interchangeable

Correct answer: b) Adjusted R-squared penalizes excess predictors to avoid overfitting

Explanation: Adjusted R-squared adjusts for the number of predictors, improving only if new variables improve model beyond chance.

16. Given a logistic regression coefficient of 0.2 for variable FirstRun, what does the odds ratio indicate?

a) Odds increase by 22% per unit increase in FirstRun

b) Odds decrease by 20%

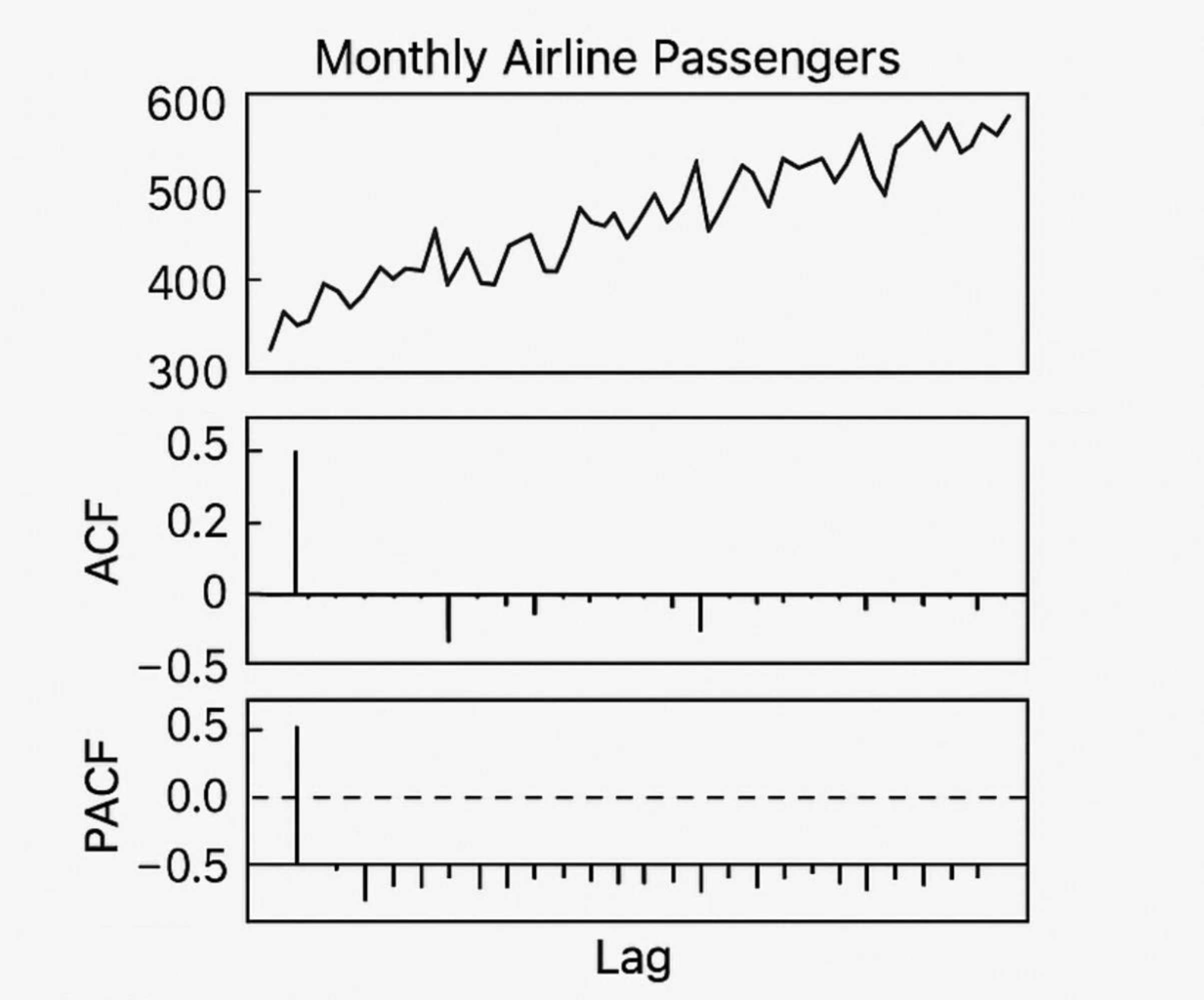
c) Variable is not significant

d) Coefficient is a probability

Correct answer: a) Odds increase by 22% per unit increase in FirstRun

Explanation: Odds ratio = exp(0.2) ≈ 1.22, meaning each unit increase raises odds by ~22%.

17. Refer to the time series correlogram where ACF shows slow decay and PACF cuts off after lag 1. Which model fits best?



a) AR(1)

b) MA(1)

c) ARMA(1,1)

d) White noise

Correct answer: a) AR(1)

Explanation: PACF cutoff after lag 1 with slow ACF decay typically indicates an autoregressive model of order 1.

18. What does the Augmented Dickey-Fuller (ADF) test assess?

a) Stationarity of time series data

b) Seasonality presence

c) Best ARIMA model order

d) Random noise level

Correct answer: a) Stationarity of time series data

Explanation: The ADF test assesses if a time series has a unit root (non-stationary); rejection implies stationarity.

19. What is the purpose of first order differencing in time series?

a) Smooth data by averaging

b) Remove trends to achieve stationarity

c) Remove seasonality entirely

d) Increase data variance

Correct answer: b) Remove trends to achieve stationarity

Explanation: Differencing subtracts consecutive observations, stabilizing the mean and removing trends for modeling.

20. Explain the difference between sensitivity and specificity.

a) Sensitivity measures true negatives; specificity measures true positives

b) Sensitivity: proportion of actual positives correctly identified; specificity: actual negatives correctly identified

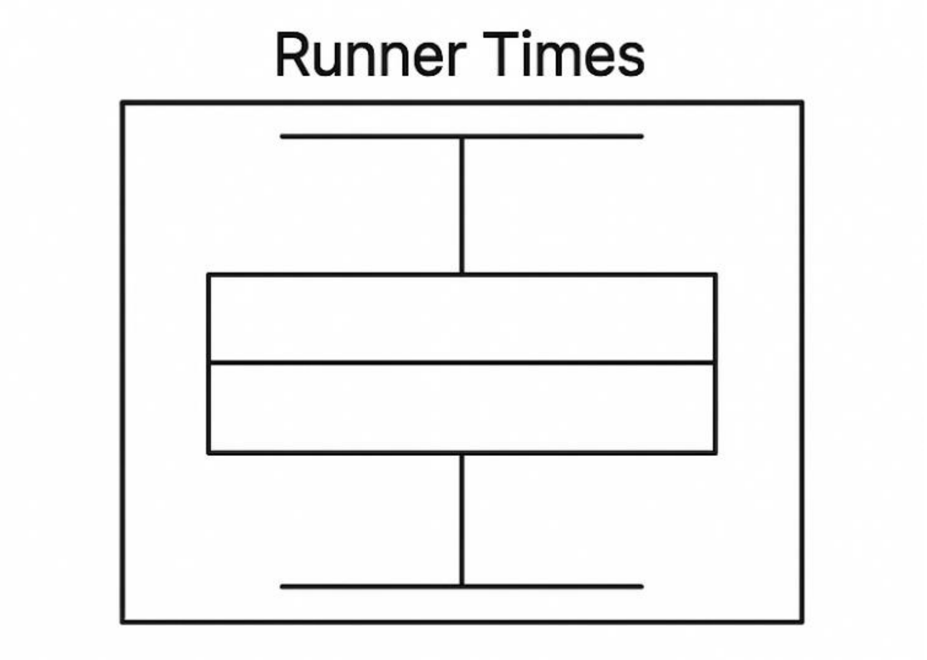
c) Both measure true positives under different thresholds

d) They are inverses

Correct answer: b) Sensitivity: proportion of actual positives correctly identified; specificity: actual negatives correctly identified

Explanation: Sensitivity is also called recall and focuses on detecting positives; specificity measures correctly identified negatives.

21. Refer to the boxplot diagram of runner times with median closer to the lower quartile:
What does this imply about skewness?



a) Symmetric distribution

b) Left skew (negative skew)

c) Right skew (positive skew)

d) Insufficient information

Correct answer: c) Right skew (positive skew)

Explanation: Median near lower edge means a longer tail on the right, indicating positive skewness.

22. In text mining, what is tokenization?

a) Removing punctuation

b) Splitting text into tokens or words

c) Converting words to roots

d) Removing stop words

Correct answer: b) Splitting text into tokens or words

Explanation: Tokenization breaks text into individual words or terms for analysis.

23. What are stop words and why remove them in NLP?

a) Rare words excluded due to insignificance

b) Common words like 'the' and 'and' that add little meaning, removed to reduce noise

c) Misspelled words

d) Proper nouns

Correct answer: b) Common words like 'the' and 'and' that add little meaning, removed to reduce noise

Explanation: Stop words are frequently used but carry little discriminative value, so removing them improves text analysis efficiency.

24. What does the DocumentTermMatrix function in R do?
a) Converts corpus to a matrix showing term frequencies
b) Removes stop words
c) Plots word clouds
d) Performs stemming

Correct answer: a) Converts corpus to a matrix showing term frequencies
Explanation: It creates a sparse matrix with documents as rows and terms as columns indicating term counts.

25. Explain TF-IDF in text analysis.
a) Counts of terms across the corpus
b) Weights terms by frequency in a document, reduced by commonness across corpus
c) Selects only the most frequent terms
d) Assigns equal weights to all terms

Correct answer: b) Weights terms by frequency in a document, reduced by commonness across corpus
Explanation: TF-IDF increases importance of terms frequent in a document but rare globally, highlighting distinctive words.

26. Which statement describes k-Nearest Neighbors (k-NN)?
a) Eager learner building a model during training
b) Lazy learner delaying classification until prediction time
c) Performs dimensionality reduction before classification
d) Only used for regression tasks

Correct answer: b) Lazy learner delaying classification until prediction time
Explanation: k-NN stores all data and classifies new instances using the nearest neighbors at prediction time.

27. Which method is NOT typically used to address class imbalance?
a) Under-sampling majority class
b) Over-sampling minority class
c) Applying SMOTE
d) Increasing the number of features

Correct answer: d) Increasing the number of features

Explanation: Increasing features doesn't solve imbalance and may worsen performance.

28. What ethical principles should guide designing persuasive user interfaces?

a) User benefit and transparency

b) Deceptiveness and manipulation

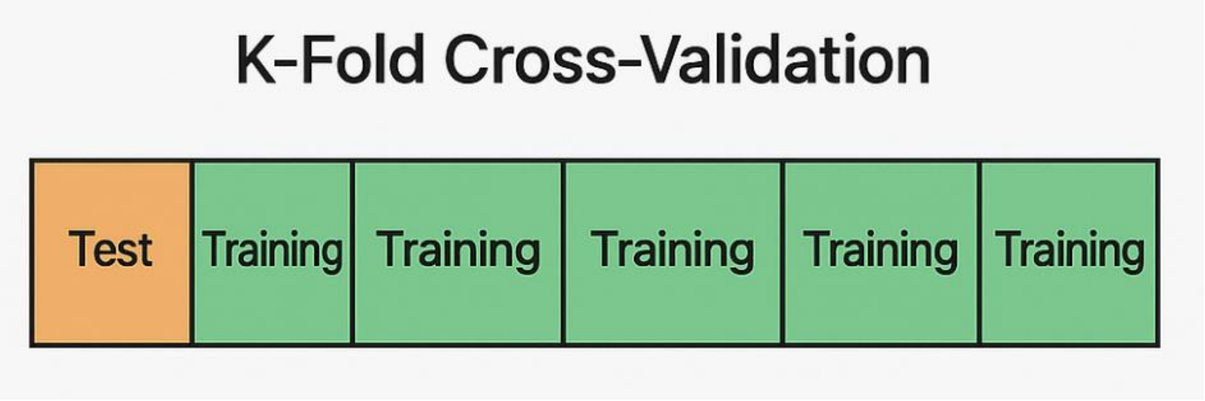
c) Maximizing data collection at all costs

d) Avoiding user consent

Correct answer: a) User benefit and transparency

Explanation: Persuasive design should respect user autonomy and be transparent, avoiding manipulative dark patterns.

29. Refer to the K-fold cross-validation diagram. What is the primary advantage over a simple train-test split?



a) Always uses more data for training

b) Assesses model performance variability by multiple test folds

c) Computationally inexpensive

d) Ignores class imbalance

Correct answer: b) Assesses model performance variability by multiple test folds

Explanation: K-fold CV tests on multiple folds, giving more reliable and stable performance estimates.

30. Write R code to calculate Euclidean distance between vectors `a = c(2,3,5)` and `b = c(1,1,1)`?

a) `sqrt(sum((a - b)^2))`

b) `sum((a - b)^2)`

c) `cor(a, b)`

d) `mean(abs(a - b))`

Correct answer: a) `sqrt(sum((a - b)^2))`
Explanation: Euclidean distance is the square root of sum of squared element-wise differences.

31. Explain the meaning and cause of the 'file drawer effect' in scientific research.
a) Publication bias where only positive or significant results are published, skewing available evidence
b) Data imputation method for missing values
c) The tendency to overfit models with too many features
d) Error in randomness due to small sample sizes

Correct answer: a) Publication bias where only positive or significant results are published, skewing available evidence
Explanation: Studies with non-significant or negative results often remain unpublished ('filed away'), leading to a distorted evidence base.

32. What does the term 'overfitting' mean in machine learning?
a) When a model learns the training data too well, including noise, and performs poorly on new data
b) When a model performs equally well on training and test data
c) Only training on a large dataset
d) Using a model too simple for the problem

Correct answer: a) When a model learns the training data too well, including noise, and performs poorly on new data
Explanation: Overfitting causes the model to memorize noise rather than learn the general pattern; this leads to poor generalization.

33. What is the purpose of the Shapiro-Wilk test in statistics?
a) To test normality of a data sample
b) To test for homoscedasticity
c) To test independence of variables
d) To compare two proportions

Correct answer: a) To test normality of a data sample
Explanation: The Shapiro-Wilk test tests the null hypothesis that a sample comes from a normally distributed population.

34. In the context of machine learning, what is 'concept drift'?
a) When the model forgets training data
b) When the statistical properties of the target variable change over time, reducing model performance

c) Scaling of data incorrectly

d) Using improper loss function

Correct answer: b) When the statistical properties of the target variable change over time, reducing model performance

Explanation: Concept drift means data patterns change over time, requiring model adaptation to maintain accuracy.

35. Describe the difference between 'lazy learning' and 'eager learning' algorithms with examples.

a) Lazy learning builds a model upfront; eager learning delays prediction until data arrives (e.g., KNN is eager, Decision Trees are lazy)

b) Lazy learning waits until prediction time to model (e.g., KNN), while eager learning builds models during training (e.g., Decision Trees, SVM)

c) They are synonyms.

d) Lazy learning requires less computation during prediction.

Correct answer: b) Lazy learning waits until prediction time to model (e.g., KNN), while eager learning builds models during training (e.g., Decision Trees, SVM)

Explanation: Lazy learners store data and classify new instances on the fly, whereas eager learners build a model in the training phase.

36. In survival analysis, what does a Kaplan-Meier plot represent?

a) Probability of survival over time for one or more groups

b) Rate of occurrence of an event

c) Linear regression fitting over time

d) Histogram of survival times

Correct answer: a) Probability of survival over time for one or more groups

Explanation: Kaplan-Meier estimator shows survival probabilities at different time points.

37. When would you prefer a paired t-test over an unpaired t-test?

a) When comparing means of two independent groups

b) When comparing means of the same subjects measured twice

c) When testing for correlation

d) When comparing variances

Correct answer: b) When comparing means of the same subjects measured twice

Explanation: Paired t-test accounts for the fact that observations are related/paired.

38. Explain the difference between precision and recall in classification.

a) Precision measures how many selected items are relevant; recall measures how many relevant

items are selected.
b) Precision measures false negatives; recall measures false positives.
c) They are identical metrics.
d) Precision is always greater than recall.

Correct answer: a) Precision measures how many selected items are relevant; recall measures how many relevant items are selected.
Explanation: Precision quantifies accuracy of positive predictions, while recall measures coverage of actual positives.

39. Provide an example R code snippet to perform a 10-fold cross-validation using the caret package for decision tree training.
a) `control <- trainControl(method = "cv", number = 10)`
`model <- train(target ~ ., data = trainingData, method = "rpart", trControl = control)`
b) `trainControl("cv", folds = 10)`
`train(model, data)`
c) `cv10 <- trainControl(nfold = 10)`
`c50(model, data = trainingData, control = cv10)`
d) No built-in way in caret for cross-validation

Correct answer: a) `control <- trainControl(method = "cv", number = 10)`
`model <- train(target ~ ., data = trainingData, method = "rpart", trControl = control)`
Explanation: caret uses `trainControl` for specifying cross-validation and `train` to build the model.

40. Which of the following best describes the Law of Large Numbers in statistics?
a) Sample mean converges to population mean as sample size increases
b) Sample variance increases with sample size
c) Probability decreases with more trials
d) The larger the sample, the greater the expected error

Correct answer: a) Sample mean converges to population mean as sample size increases
Explanation: The Law of Large Numbers states the sample average approaches the expected value with increasing sample size.

41. Refer to the diagram showing a decision tree structure with root, branches, and leaves. What main criterion is used by decision trees for splitting nodes?
a) Information Gain, based on entropy reduction
b) Random selection of features

c) Euclidean distance minimization

d) Gradient descent

Correct answer: a) Information Gain, based on entropy reduction

Explanation: Decision trees choose splits that maximize information gain by reducing entropy.

42. What is the meaning of 'pseudo-replication' in statistical analyses and why is it problematic?

a) Treating non-independent data points as independent, inflating sample size and risking false significance

b) Repeating analyses multiple times with different data

c) Using bootstrapping to estimate statistics

d) Pooling data from multiple studies

Correct answer: a) Treating non-independent data points as independent, inflating sample size and risking false significance

Explanation: Pseudo-replication falsely inflates sample size, increasing type I error risk.

43. What kind of features are called 'independent' when using Pearson correlation for feature selection?

a) Features highly correlated with each other

b) Features uncorrelated with each other but correlated with the target variable

c) Features not correlated with the target variable

d) Features that are categorical

Correct answer: b) Features uncorrelated with each other but correlated with the target variable

Explanation: Independent features are not mutually correlated but have a predictive relationship with the target.

44. In the R code for standardizing numerical features: `IrisData$SepalLengthCm <- scale(IrisData$SepalLengthCm)`, what does the `scale` function do?

a) Normalizes data to range

b) Subtracts mean and divides by standard deviation, producing standardized features

c) Converts data to factors

d) Performs Principal Component Analysis

Correct answer: b) Subtracts mean and divides by standard deviation, producing standardized features

Explanation: `scale()` centers and scales variables to mean zero and unit variance.

45. Explain the term 'confounding' in the context of data analysis.

a) A variable that distorts the apparent effect of an explanatory variable on the outcome

b) Random noise in data

c) Missing values that reduce power

d) Interaction effects between variables

Correct answer: a) A variable that distorts the apparent effect of an explanatory variable on the outcome

Explanation: Confounders bias estimates if not controlled, making causal inferences misleading.

46. Explain why the Balanced Accuracy metric is preferred over simple Accuracy in imbalanced classification problems.

a) Balanced Accuracy averages sensitivity and specificity, giving fair evaluation even when classes are uneven

b) Balanced Accuracy ignores true negatives

c) Accuracy always overestimates model quality

d) Balanced Accuracy is simpler to compute

Correct answer: a) Balanced Accuracy averages sensitivity and specificity, giving fair evaluation even when classes are uneven

Explanation: Balanced accuracy accounts for both classes, avoiding inflated scores due to majority class dominance.

47. Which R package is commonly used for survival analysis including functions like `survfit` and Kaplan-Meier plots?

a) survival

b) caret

c) tm

d) randomForest

Correct answer: a) survival

Explanation: The 'survival' package provides tools for handling survival data, including Kaplan-Meier estimation.

48. What is the ethical concern with 'dark patterns' in User Interface (UI) design?

a) They improve user experience transparently

b) They intentionally manipulate users for business gains at user's expense without consent

c) They mandate accessibility standards

d) They simplify user decisions

Correct answer: b) They intentionally manipulate users for business gains at user's expense without consent

Explanation: Dark patterns deceive users into unintended actions, violating ethical UI practices.

49. What is the difference between 'discretization' and 'quantization' in data preprocessing?
a) Discretization creates continuous variables from categorical; quantization does the opposite
b) Both refer to dividing continuous variables into intervals or bins, often used interchangeably
c) Discretization applies only to images, quantization only to time series
d) Quantization increases variable range

Correct answer: b) Both refer to dividing continuous variables into intervals or bins, often used interchangeably
Explanation: Both terms describe the binning of continuous data into discrete buckets.

50. Describe the function of AUC (Area Under the Curve) in ROC analysis.
a) Quantifies aggregate model performance across classification thresholds
b) Measures model training time
c) Identifies optimal cutoff value directly
d) Measures correlation between features

Correct answer: a) Quantifies aggregate model performance across classification thresholds
Explanation: AUC summarizes the trade-off between sensitivity and specificity, indicating overall discriminative ability.

51. What is the Synthetic Minority Over-sampling Technique (SMOTE) primarily used for in machine learning?
a) Reducing dimensionality of features
b) Balancing class distribution by generating synthetic minority class samples
c) Encoding categorical variables
d) Improving model interpretability

Correct answer: b) Balancing class distribution by generating synthetic minority class samples
Explanation: SMOTE creates synthetic minority samples by interpolating between existing minority instances to address class imbalance.

52. Given the R code snippet: `generateData = SMOTE(dataSet[, c(1, 2)], dataSet[, c(3)], K=5)`, what is the role of parameter K=5?
a) Number of nearest neighbors to use for synthetic sample generation
b) Number of features included
c) Number of classes in the dataset
d) Number of times SMOTE runs

Correct answer: a) Number of nearest neighbors to use for synthetic sample generation
Explanation: K defines how many neighbors SMOTE uses to create synthetic examples in the minority class.

53. Refer to the following simplified decision tree split diagram on a continuous variable:

[Insert diagram showing root node splitting feature X ≤ 5]

What criterion is commonly used to select the splitting point in decision trees?

a) Correlation coefficient

b) Information gain or reduction in entropy

c) Euclidean distance

d) Logistic regression coefficient

Correct answer: b) Information gain or reduction in entropy

Explanation: Decision trees select splits by maximizing information gain, measured by decrease in entropy.

54. What does 'pseudo-replication' mean in statistical analyses and why is it problematic?

a) Treating non-independent data points as independent, inflating sample size and risking false positives

b) Repeated measures design

c) Using bootstrap methods

d) Pooling datasets from multiple studies

Correct answer: a) Treating non-independent data points as independent, inflating sample size and risking false positives

Explanation: Ignoring dependencies leads to false conclusions due to underestimating variability.

55. Using the data comparison table for Pseudonymization and Anonymization (previously provided), which statement is correct?

a) Anonymization allows reversible linking to original data

b) Pseudonymization is irreversible

c) Pseudonymization retains data utility and reversibility with a key

d) Both are considered personal data under GDPR

Correct answer: c) Pseudonymization retains data utility and reversibility with a key

Explanation: Pseudonymization maintains linkability with a key and higher utility; anonymization is irreversible.

56. Refer to this confusion matrix:

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

What type of error is False Positive?

a) Correctly predicting positive cases

b) Predicting negative when actual is positive

c) Predicting positive when actual is negative

d) Correctly predicting negative cases

Correct answer: c) Predicting positive when actual is negative

Explanation: False positives occur when negatives are incorrectly classified as positive.

57. Calculate the entropy of a dataset with two classes where 90% are class A and 10% are class B.

a) 0.469

b) 0.325

c) 0.9

d) 1.0

Correct answer: a) 0.469

Explanation: Entropy = - (0.9*log2(0.9) + 0.1*log2(0.1)) ≈ 0.469 bits.

58. What does the term 'reproducibility' mean in data science research?

a) Ability to obtain consistent results using the same methods and data

b) Using different data to confirm findings

c) Sharing all data publicly

d) Using multiple models

Correct answer: a) Ability to obtain consistent results using the same methods and data

Explanation: Reproducibility ensures scientific results can be validated and trusted.

59. Which feature selection approach uses a model to recursively remove least important features?

a) Filter method

b) Wrapper method - Recursive Feature Elimination (RFE)

c) Embedded method - Lasso regression

d) Dimensionality reduction - PCA

Correct answer: b) Wrapper method - Recursive Feature Elimination (RFE)

Explanation: RFE iteratively trains models and removes less important features.

60. Refer to the time series PACF diagram where PACF cuts off sharply after lag 1 and ACF decays slowly. Which model is most appropriate?

a) MA(1) model

b) AR(1) model

c) White noise

d) Seasonal ARIMA

Correct answer: b) AR(1) model

Explanation: Sharp PACF cutoff at lag 1 with slow ACF decay indicates autoregressive model order 1.

61. What is the purpose of the Variance Inflation Factor (VIF) in multiple regression analysis?

a) To detect heteroscedasticity

b) To identify predictors that are highly collinear

c) To assess model fit

d) To measure residual variance

Correct answer: b) To identify predictors that are highly collinear

Explanation: VIF measures the extent to which the variance of a coefficient is inflated due to multicollinearity among predictors.

62. What does a Shapiro-Wilk test p-value > 0.05 imply about a dataset?

a) Data significantly deviates from normality

b) Data is consistent with normal distribution

c) Test is invalid

d) Data contains outliers

Correct answer: b) Data is consistent with normal distribution

Explanation: A high p-value indicates no evidence to reject the null hypothesis that data is normally distributed.

63. Explain the difference between sensitivity and specificity in a binary classification problem.

a) Sensitivity is the proportion of actual negatives correctly identified; specificity is actual positives

b) Sensitivity measures how well positives are identified; specificity how well negatives are identified

c) Both measure false positive rate

d) They are unrelated metrics

Correct answer: b) Sensitivity measures how well positives are identified; specificity how well negatives are identified

Explanation: Sensitivity is also known as recall and specificity identifies how well the model detects true negatives.

64. In R, what does the function call `cor.test(x, y, method = "pearson")` compute?

a) Linear regression of y on x

b) Correlation coefficient and significance test between x and y

c) Mean difference between x and y

d) Covariance matrix

Correct answer: b) Correlation coefficient and significance test between x and y

Explanation: This function computes Pearson's correlation and tests whether it's significantly different from zero.

65. Describe the primary difference between bagging and boosting ensemble methods.

a) Bagging trains models independently in parallel; boosting trains sequentially focusing on correcting errors

b) Boosting models run in parallel; bagging uses sequential training

c) Both reduce bias equally

d) They are the same

Correct answer: a) Bagging trains models independently in parallel; boosting trains sequentially focusing on correcting errors

Explanation: Bagging reduces variance with multiple independent models; boosting reduces bias by learning from previous errors.

66. What is tokenization in text mining?

a) Combining several words into a phrase

b) Splitting text into discrete tokens or words

c) Removing stop words

d) Converting text to uppercase

Correct answer: b) Splitting text into discrete tokens or words

Explanation: Tokenization breaks text into its tokens—usually words—for further processing.

67. Explain what stop words are and why they are removed in natural language processing (NLP).

a) Rare words removed for simplicity

b) Common words that add little meaning, removed to reduce noise

c) Misspelled words

d) Proper nouns

Correct answer: b) Common words that add little meaning, removed to reduce noise

Explanation: Stop words like "the" and "and" occur frequently but provide little semantic information.

68. In text analysis, what does the DocumentTermMatrix function in R do?

a) Removes stop words from text

b) Creates a matrix of term frequencies per document

c) Generates word clouds

d) Performs stemming

Correct answer: b) Creates a matrix of term frequencies per document

Explanation: This function transforms text into a sparse matrix indicating frequency of each term in each document.

69. What does the TF-IDF weighting scheme in text mining represent?

a) Raw count of word occurrences

b) Term frequency scaled down by inverse document frequency to highlight important words

c) Binary presence of words

d) Normalization of word lengths

Correct answer: b) Term frequency scaled down by inverse document frequency to highlight important words

Explanation: TF-IDF gives higher weights to words common in a document but rare across the corpus.

70. How does the k-Nearest Neighbors (k-NN) algorithm classify a new instance?

a) Using a pre-built model fitted during training

b) By majority vote of k nearest neighbors in feature space

c) By calculating Euclidean distance to class centroids

d) Using decision trees

Correct answer: b) By majority vote of k nearest neighbors in feature space

Explanation: k-NN classifies based on the classes of the closest training examples to the new data point.

71. Which of these methods is NOT a typical way to handle class imbalance?

a) Under-sampling the majority class

b) Over-sampling the minority class

c) Using SMOTE to generate synthetic samples

d) Increasing the number of features

Correct answer: d) Increasing the number of features

Explanation: Simply adding more features doesn’t address class imbalance and may worsen performance.

72. What is 'algorithmic bias' in machine learning?

a) Random errors caused by data noise

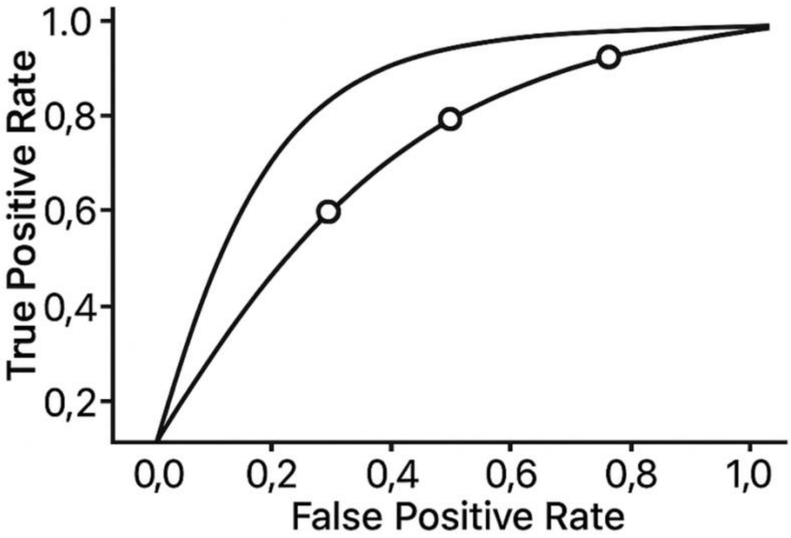
b) Systematic unfair or discriminatory model predictions due to biased data or design
c) Intentional algorithm manipulation
d) Loss of model accuracy during training

Correct answer: b) Systematic unfair or discriminatory model predictions due to biased data or design
Explanation: Algorithmic bias results when models produce unfair outcomes against certain groups.

73. What is the 'file drawer effect' in scientific research?
a) Loss of experimental data during transfers
b) Bias where negative or non-significant results remain unpublished
c) Systematic error from mislabeling data
d) Outdated methodologies discarded unfairly

Correct answer: b) Bias where negative or non-significant results remain unpublished
Explanation: Positive results are published more often, skewing the apparent evidence.

74. Refer to this ROC curve plot with threshold points marked:
What is the effect of lowering the classification threshold on metrics?



a) Sensitivity decreases, false positive rate decreases
b) Sensitivity increases, false positive rate increases
c) Both sensitivity and specificity increase
d) No effect on sensitivity

Correct answer: b) Sensitivity increases, false positive rate increases
Explanation: Lower thresholds classify more positives, catching more true positives but also more false positives.

75. What is the main disadvantage of using a very high value of K in k-NN classifiers?

a) Overfitting the training data

b) Increased sensitivity to noise

c) Underfitting by oversmoothing decision boundaries

d) High computational cost during training

Correct answer: c) Underfitting by oversmoothing decision boundaries

Explanation: Large K values generalize too much, masking patterns and reducing accuracy.

76. Explain the concept of 'concept drift' in machine learning.

a) Model losing training data

b) Changing statistical properties of the target variable over time causing degraded model performance

c) Scaling data inconsistently

d) Non-random missing data

Correct answer: b) Changing statistical properties of the target variable over time causing degraded model performance

Explanation: When the data distribution changes over time, static models may become inaccurate.

77. Describe the key difference between bagging and boosting ensemble methods.

a) Bagging uses sequential training focusing on misclassified samples; boosting trains models independently

b) Bagging trains models independently in parallel; boosting trains sequentially focusing on correcting errors

c) Both train models sequentially

d) Boosting only applies to regression

Correct answer: b) Bagging trains models independently in parallel; boosting trains sequentially focusing on correcting errors

Explanation: Bagging reduces variance; boosting reduces bias by focusing on errors from previous models.

78. How does the Boruta algorithm select important features?

a) Uses Pearson correlation to remove correlated features

b) Creates shadow features and compares their importance using Random Forests

c) Applies PCA for dimensionality reduction

d) Embedded selection during logistic regression

Correct answer: b) Creates shadow features and compares their importance using Random Forests

Explanation: Boruta compares real features with random shadows to retain relevant features.

79. Given a logistic regression coefficient of 0.405 for a variable, what is the interpretation of the odds ratio?

a) Odds increase by 40.5% for one unit increase

b) Odds decrease by 40.5%

c) Coefficient has no meaningful interpretation

d) Odds ratio is 0.405

Correct answer: a) Odds increase by 40.5% for one unit increase

Explanation: Odds ratio = exp(0.405) ≈ 1.5, so odds increase about 50% per unit increase.

80. In time series analysis, what does differencing do?

a) Adds seasonality

b) Removes trend to make series stationary

c) Normalizes data

d) Increases variance

Correct answer: b) Removes trend to make series stationary

Explanation: Differencing subtracts consecutive values to remove underlying trends.

81. What is the general purpose of hyperparameter tuning in machine learning?

a) To find the best configuration of model parameters to optimize performance

b) To clean the dataset

c) To create features

d) None of the above

Correct answer: a) To find the best configuration of model parameters to optimize performance

Explanation: Tuning improves model accuracy by optimizing parameters like learning rate or tree depth.

82. What is the Law of Large Numbers in statistics?

a) Sample mean tends to population mean as sample size increases

b) Variance grows with larger samples

c) Probability decreases as trials increase

d) Sample error grows with sample size

Correct answer: a) Sample mean tends to population mean as sample size increases

Explanation: Larger samples improve the estimate stability of the population mean.

83. Explain the difference between Adjusted R-squared and R-squared in regression analysis.

a) R-squared decreases with more variables, Adjusted R-squared increases

b) Adjusted R-squared accounts for number of predictors, penalizing overly complex models

c) No difference

d) Adjusted R-squared measures residuals only

Correct answer: b) Adjusted R-squared accounts for number of predictors, penalizing overly complex models

Explanation: Adjusted R-squared only increases if new variables improve the model beyond chance.

84. What ethical issue do 'dark patterns' in user interface design raise?

a) Transparent design benefiting users

b) Manipulative design tricking users into unintended actions

c) Mandatory accessibility standards

d) Open source software sharing

Correct answer: b) Manipulative design tricking users into unintended actions

Explanation: Dark patterns deceive users for business gains, violating ethical use principles.

85. In text mining, explain the concept of stemming.

a) Removing noise from data

b) Reducing words to their root form by cutting suffixes

c) Converting documents to matrices

d) Removing stopwords

Correct answer: b) Reducing words to their root form by cutting suffixes

Explanation: Stemming simplifies words to their base for better grouping.

86. What is the key difference between lazy learning and eager learning algorithms?

a) Lazy learning builds models ahead; eager learning waits to predict

b) Lazy learning waits until prediction time to compute; eager learning builds model during training

c) Both build models during training

d) Lazy learning is only unsupervised methods

Correct answer: b) Lazy learning waits until prediction time to compute; eager learning builds model during training

Explanation: Lazy learning defers generalization until needed; eager learning generalizes upfront.

87. What R package provides robust survival analysis including Kaplan-Meier plots?

a) caret

b) survival

c) tm

d) randomForest

Correct answer: b) survival

Explanation: The survival package offers functions for survival modeling and plotting.

88. Explain the difference between precision and recall with formulas.

a) Precision = TP/(TP+FP), Recall = TP/(TP+FN)

b) Precision = TP/(TP+FN), Recall = TP/(TP+FP)

c) Precision and recall are the same

d) Precision measures negatives, recall measures positives

Correct answer: a) Precision = TP/(TP+FP), Recall = TP/(TP+FN)

Explanation: Precision measures accuracy of positives predicted; recall measures coverage of actual positives.

89. Provide an R code snippet to perform a 10-fold cross-validation on a decision tree using the caret package.

a)
```
control <- trainControl(method = "cv", number = 10)
model <- train(target ~ ., data = trainingData, method = "rpart", trControl = control)
```

b)
```
trainControl("cv", folds = 10)
train(model, data)
```

c)
```
cv10 <- trainControl(nfold = 10)
c50(model, data = trainingData, control = cv10)
```

d)
```
crossval <- cv.tree(model, K=10)
```

Correct answer: a)
```
control <- trainControl(method = "cv", number = 10)
model <- train(target ~ ., data = trainingData, method = "rpart", trControl = control)
```

Explanation: caret supports cross-validation via trainControl and train functions.

90. What is the main assumption behind the Central Limit Theorem (CLT)?

a) Samples must be normally distributed

b) Samples must be independent and identically distributed with finite variance

c) Sample size is less than 30

d) Population mean is unknown

Correct answer: b) Samples must be independent and identically distributed with finite variance

Explanation: CLT states sample means approximate normal distribution as sample size grows, assuming i.i.d and finite variance.

91. Describe the concept of 'multicollinearity' in regression analysis.

a) Predictor variables are completely independent

b) Predictors are highly correlated, causing unstable coefficient estimates
c) Outcome variable is binary
d) Residuals follow a normal distribution

Correct answer: b) Predictors are highly correlated, causing unstable coefficient estimates
Explanation: Multicollinearity inflates variance of parameter estimates, making coefficients difficult to interpret reliably.

92. Why should feature selection be performed only on training data?
a) To prevent data leakage and ensure unbiased model evaluation
b) Because test data features are irrelevant
c) To simplify computation
d) No reason; it can be done on entire dataset

Correct answer: a) To prevent data leakage and ensure unbiased model evaluation
Explanation: Using test data during feature selection leaks information influencing training, leading to overoptimistic performance.

93. What is the primary advantage of ensemble learning methods like Random Forests?
a) They reduce variance by averaging many decision trees trained on bootstrapped data
b) They always perform worse than single decision trees
c) Only useful for regression
d) Require no parameter tuning

Correct answer: a) They reduce variance by averaging many decision trees trained on bootstrapped data
Explanation: Ensembles improve prediction stability and accuracy by combining multiple models.

94. What is the impact of class imbalance on machine learning classifiers?
a) Models perform equally well on all classes
b) Models may bias predictions toward the majority class, ignoring minority
c) Improves model precision
d) Not a concern

Correct answer: b) Models may bias predictions toward the majority class, ignoring minority
Explanation: Class imbalance often results in poor detection of minority class due to skewed training data.

95. In survival analysis, what is the primary information depicted in a Kaplan-Meier curve?
a) Hazard rate over time
b) Probability of survival as a function of time

c) Regression coefficient estimates

d) Cumulative incidence of an event

Correct answer: b) Probability of survival as a function of time

Explanation: Kaplan-Meier curves estimate the survival function showing probability individuals survive beyond certain time.

96. In hypothesis testing, what does a p-value less than 0.05 generally signify?

a) Accept the null hypothesis

b) Reject the null hypothesis; results are statistically significant

c) Sample size is too small

d) Data is normally distributed

Correct answer: b) Reject the null hypothesis; results are statistically significant

Explanation: A low p-value suggests observed results unlikely due to chance assuming null is true.

97. What is the typical effect of outliers on correlation analysis?

a) No effect

b) May severely distort correlation coefficients

c) Always increase correlation

d) Always decrease correlation

Correct answer: b) May severely distort correlation coefficients

Explanation: Outliers can inflate or deflate correlations, misrepresenting true relationships.

98. What is the difference between supervised and unsupervised learning?

a) Supervised learning uses labeled data; unsupervised uses unlabeled data to find structure

b) Unsupervised uses labeled data; supervised uses unlabeled data

c) Both always require labeled data

d) No difference

Correct answer: a) Supervised learning uses labeled data; unsupervised uses unlabeled data to find structure

Explanation: Supervised learning predicts target labels; unsupervised discovers hidden patterns without labels.

99. What does the R function `lm()` perform?

a) Principal component analysis

b) Linear regression modeling

c) Logistic regression

d) k-NN classification

Correct answer: b) Linear regression modeling
Explanation: The `lm()` function fits linear models predicting continuous outcomes.

100. Explain the concept of the 'black box' problem in machine learning models.
a) Models whose inner workings are transparent and easily interpretable
b) Models difficult to interpret or explain despite making predictions
c) Simple algorithms with few parameters
d) Open source models

Correct answer: b) Models difficult to interpret or explain despite making predictions
Explanation: Complex models like deep neural networks often lack transparency, complicating trust and debugging.

101. Which of the following best describes the main stages of the data science lifecycle?
a) Data collection, data cleaning, model deployment
b) Data collection, data cleaning, exploratory data analysis, model building, and deployment
c) Only data cleaning and model building
d) Data engineering, feature engineering, testing

Correct answer: b) Data collection, data cleaning, exploratory data analysis, model building, and deployment
Explanation: The full lifecycle includes data gathering, preparation, exploration, modeling, evaluation, and deployment.

102. In R, which function is used to check the structure and data types of a dataset 'df'?
a) summary(df)
b) str(df)
c) head(df)
d) dim(df)

Correct answer: b) str(df)
Explanation: The `str()` function displays the structure and types of variables in the data frame.

103. Which of the following is a continuous random variable?
a) Number of students in a classroom
b) Weight of a person
c) Number of cars in a parking lot
d) Number of heads in coin flips

Correct answer: b) Weight of a person
Explanation: Weight can take any value in a range (continuous), others are count/discrete variables.

104. What does the Shapiro-Wilk test evaluate in a given dataset?
a) Whether the data follows a normal distribution
b) Linearity between two variables
c) Independence of observations
d) Equality of variances

Correct answer: a) Whether the data follows a normal distribution
Explanation: Shapiro-Wilk tests the null hypothesis that data is drawn from a normal distribution.

105. What is 'concept drift' and why is it important in predictive modeling?
a) When data distribution changes over time, causing models to degrade if not updated
b) Loss of training data due to errors
c) Feature scaling issue
d) Overfitting due to large feature sets

Correct answer: a) When data distribution changes over time, causing models to degrade if not updated
Explanation: Recognizing and adapting to concept drift is critical to maintaining model accuracy in changing environments.

106. In hypothesis testing, what does a p-value < 0.05 signify?
a) The null hypothesis is accepted
b) The null hypothesis can be rejected with 95% confidence
c) The alternative hypothesis is false
d) There is a 5% chance of observing the data

Correct answer: b) The null hypothesis can be rejected with 95% confidence
Explanation: p-value < 0.05 means the data provides strong evidence against the null hypothesis.

107. What is the main difference between paired and independent t-tests?
a) Paired t-test compares related samples; independent t-test compares unrelated groups
b) Paired t-test compares more than two groups; independent compares two
c) Paired t-test for large samples; independent for small
d) No difference

Correct answer: a) Paired t-test compares related samples; independent t-test compares unrelated groups

Explanation: Paired test controls for within-subject effects; independent tests compare separate groups.

108. What does the Pearson correlation coefficient quantify?
a) Strength of a linear relationship between two variables
b) Difference between means
c) Causation between variables
d) Variance of one variable

Correct answer: a) Strength of a linear relationship between two variables
Explanation: Pearson correlation measures linear dependence ranging from -1 to 1.

109. What is the purpose of the Variance Inflation Factor (VIF) in regression?
a) Detect multicollinearity among predictors
b) Measure overall model accuracy
c) Test normality of residuals
d) Detect heteroscedasticity

Correct answer: a) Detect multicollinearity among predictors
Explanation: VIF identifies predictors highly correlated with others, which may affect coefficient reliability.

110. Refer to this plotted autocorrelation function (ACF) diagram:
If the ACF decays slowly and the Partial ACF (PACF) cuts off after lag 1, which time series model is appropriate?
a) AR(1) model
b) MA(1) model
c) White noise
d) Seasonal ARIMA

Correct answer: a) AR(1) model
Explanation: AR(1) models produce slow ACF decay and PACF cutoff at lag 1.

111. In logistic regression, a coefficient of 0.4 corresponds approximately to an odds ratio of what?
a) 1.5
b) 0.4
c) 2.5
d) 0.9

Correct answer: a) 1.5
Explanation: Odds ratio = exp(0.4) ≈ 1.5, meaning a 50% increase in odds with a unit increase.

112. In R, which function can be used to compute the Pearson correlation coefficient and perform a significance test?

a) cor.test()

b) lm()

c) t.test()

d) summary()

Correct answer: a) cor.test()

Explanation: `cor.test()` computes the correlation coefficient and its p-value.

113. In R, which package is commonly used for survival analysis?

a) survival

b) caret

c) tm

d) randomForest

Correct answer: a) survival

Explanation: The survival package offers Kaplan-Meier and Cox model functions.

114. Explain the difference between sensitivity and specificity in classification tasks.

a) Sensitivity measures true positive rate; specificity measures true negative rate

b) Both measure false positive rate

c) Sensitivity measures precision; specificity measures recall

d) They are the same

Correct answer: a) Sensitivity measures true positive rate; specificity measures true negative rate

Explanation: Sensitivity measures ability to detect positives; specificity measures ability to detect negatives.

115. Refer to the confusion matrix below:

Actual \ Predicted	Positive	Negative
Positive	TP	FN
Negative	FP	TN

What does 'False Negative' mean in this context?

a) Incorrectly predicting negative when true is positive

b) Correctly predicting negative

c) Incorrectly predicting positive when true is negative

d) Correctly predicting positive

Correct answer: a) Incorrectly predicting negative when true is positive

Explanation: False negatives are missed positive cases predicted as negative.

116. What is SMOTE primarily used for in supervised machine learning?

a) Dimensionality reduction

b) Balancing imbalanced datasets by generating synthetic minority samples

c) Feature scaling

d) Improving feature interpretability

Correct answer: b) Balancing imbalanced datasets by generating synthetic minority samples

Explanation: SMOTE synthesizes new data points for minority classes to reduce imbalance.

117. Which R function standardizes variables by subtracting the mean and dividing by standard deviation?

a) scale()

b) normalize()

c) center()

d) standardize()

Correct answer: a) scale()

Explanation: `scale()` centers and scales features to have zero mean and unit variance.

118. In feature selection, what does the Boruta algorithm do?

a) Filters features based on p-values only

b) Uses Random Forest to compare actual features with shadow features to pick important ones

c) Performs PCA to reduce dimensions

d) Uses manual selection of variables

Correct answer: b) Uses Random Forest to compare actual features with shadow features to pick important ones

Explanation: Boruta iteratively tests whether features perform better than randomized shadows.

119. Which of the following ethical issues arises from ‘dark patterns’ in user interface design?

a) Improving user understanding

b) Manipulation of users to take unintended actions

c) Supporting data privacy

d) Enhancing accessibility

Correct answer: b) Manipulation of users to take unintended actions
Explanation: Dark patterns intentionally deceive or coerce users for business advantage.

120. What does the General Data Protection Regulation (GDPR) principle of 'data minimization' refer to?
a) Collecting and processing only necessary personal data
b) Storing data indefinitely
c) Sharing data without consent
d) Deleting all data immediately

Correct answer: a) Collecting and processing only necessary personal data
Explanation: Data minimization limits personal data collected and retained to what is strictly needed.

121. In topic modeling, what does Latent Dirichlet Allocation (LDA) do?
a) Clusters documents based on explicit labels
b) Discovers hidden topics by analyzing word co-occurrences
c) Removes stopwords from documents
d) Performs sentiment analysis

Correct answer: b) Discovers hidden topics by analyzing word co-occurrences
Explanation: LDA is a probabilistic model that identifies latent topics in a text corpus by grouping words that frequently co-occur.

122. What does the term 'tokenization' mean in text mining?
a) Splitting text into words or tokens
b) Removing punctuation
c) Converting text to lowercase
d) Lemmatizing words

Correct answer: a) Splitting text into words or tokens
Explanation: Tokenization breaks text into meaningful units (tokens), typically words, for further natural language processing.

123. What is the benefit of using term frequency-inverse document frequency (TF-IDF) in text analysis?
a) Emphasizes common words in all documents
b) Highlights terms important to individual documents relative to the corpus
c) Ignores rare words
d) Removes punctuation

Correct answer: b) Highlights terms important to individual documents relative to the corpus
Explanation: TF-IDF weights terms higher when they appear frequently in a document but rarely in the corpus, thus emphasizing distinctive terms.

124. What is the primary goal when tuning hyperparameters in machine learning models?
a) Maximize model interpretability
b) Improve predictive performance on unseen data
c) Reduce dataset size
d) Increase training time

Correct answer: b) Improve predictive performance on unseen data
Explanation: Optimizing hyperparameters helps better generalize the model to new data and avoid overfitting or underfitting.

125. Which of the following R packages is commonly used for ML model training and parameter tuning?
a) caret
b) ggplot2
c) tm
d) survival

Correct answer: a) caret
Explanation: The `caret` package provides tools for building, tuning, and evaluating machine learning models.

126. Explain the 'file drawer effect' in the context of scientific research.
a) Publishing all studies regardless of outcome
b) Non-significant results often remain unpublished, biasing literature
c) Reproducibility of experiments
d) Data sharing among researchers

Correct answer: b) Non-significant results often remain unpublished, biasing literature
Explanation: This publication bias leads to an overrepresentation of positive findings in literature.

127. What R function would you use to perform a one-sample t-test?
a) t.test()
b) lm()
c) cor.test()
d) anova()

Correct answer: a) t.test()

Explanation: `t.test()` can perform one-sample, two-sample, paired, and unpaired t-tests.

128. In R, how would you check for missing values in a vector `x`?

a) is.na(x)

b) is.nan(x)

c) anyNA(x)

d) allNA(x)

Correct answer: a) is.na(x)

Explanation: `is.na()` identifies elements that are `NA` or missing in R.

129. What does the term 'confounding' mean in statistical analyses?

a) A factor that distorts the apparent association between predictor and outcome

b) An outlier value

c) A missing value in the dataset

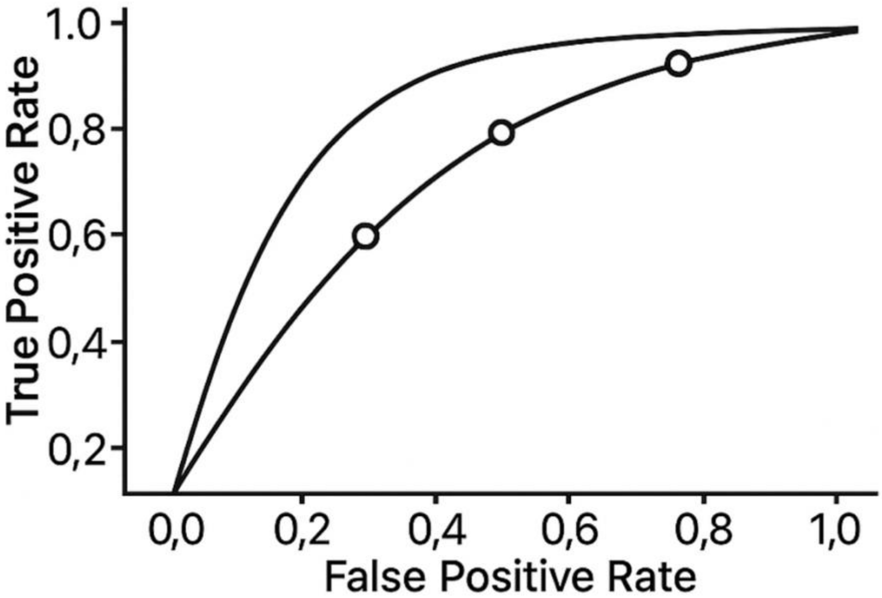
d) A variable with a linear relationship

Correct answer: a) A factor that distorts the apparent association between predictor and outcome

Explanation: Confounders bias the relationship if they are associated with both the predictor and outcome.

130. Refer to the diagram of Receiver Operating Characteristic (ROC) curve:

What does the curve illustrate?



a) Trade-off between sensitivity and specificity at different thresholds

b) Time series trend

c) Confusion matrix outputs

d) Feature importance ranking

Correct answer: a) Trade-off between sensitivity and specificity at different thresholds

Explanation: ROC plots True Positive Rate vs. False Positive Rate as classification threshold varies.

131. What is the primary purpose of Exploratory Data Analysis (EDA)?

a) To build predictive models

b) To summarize main characteristics of data often using visual methods

c) To clean data automatically

d) To conduct hypothesis testing

Correct answer: b) To summarize main characteristics of data often using visual methods

Explanation: EDA helps detect patterns, anomalies, and insights before formal modeling.

132. What is a limitation of using R-squared as a sole metric for regression model fit?

a) It always increases when predictors are added, regardless of usefulness

b) It is difficult to compute

c) It is only valid for logistic regression

d) It measures residual errors poorly

Correct answer: a) It always increases when predictors are added, regardless of usefulness

Explanation: R-squared can increase by adding variables without improving model quality, which may cause overfitting.

133. What is the ethical concern behind algorithmic bias in AI and ML?

a) It results in unfair or discriminatory outcomes for certain groups

b) It increases computational cost

c) It leads to faster model training

d) It improves model interpretability

Correct answer: a) It results in unfair or discriminatory outcomes for certain groups

Explanation: Bias in data or algorithm design can reinforce social inequalities.

134. What type of learning is k-Nearest Neighbors (k-NN) considered?

a) Lazy learning

b) Eager learning

c) Supervised deep learning

d) Unsupervised clustering

Correct answer: a) Lazy learning

Explanation: k-NN defers modeling until query time, using stored data for classification.

135. In R, which function creates a Document-Term Matrix useful for text mining?

a) DocumentTermMatrix()

b) corpus()

c) wordcloud()

d) list.files()

Correct answer: a) DocumentTermMatrix()

Explanation: It converts a corpus to a matrix of term frequencies per document.

136. What is one key benefit of synthetic data like that generated by SMOTE in data science?

a) Helps balance datasets to improve model training on minority classes

b) Removes all personal data from datasets

c) Reduces dataset size

d) Guarantees no data bias

Correct answer: a) Helps balance datasets to improve model training on minority classes

Explanation: Synthetic data supplements minority classes, enhancing classifier learning.

137. Explain the Law of Large Numbers in statistics.

a) As sample size increases, sample mean converges to true population mean

b) Sample variance always increases with sample size

c) Probability of errors increases with samples

d) Population mean decreases with large samples

Correct answer: a) As sample size increases, sample mean converges to true population mean

Explanation: Larger samples yield more accurate estimates of the population mean.

138. What does a p-value represent in hypothesis testing?

a) Probability of observing the data or more extreme, assuming null hypothesis is true

b) Probability the null hypothesis is true

c) Sample size required

d) The alternative hypothesis

Correct answer: a) Probability of observing the data or more extreme, assuming null hypothesis is true

Explanation: p-value measures evidence against the null hypothesis.

139. When performing feature selection, why is it important to only use the training data?

a) To avoid information leakage and preserve unbiased evaluation

b) Because test data contains missing values

c) To reduce computation time

d) It's not important

Correct answer: a) To avoid information leakage and preserve unbiased evaluation

Explanation: Using test data during selection leaks knowledge and falsely inflates performance.

140. In survival analysis, what does a Kaplan-Meier curve show?

a) Estimated probability of survival over time

b) Hazard function

c) Regression coefficients

d) Confusion matrix

Correct answer: a) Estimated probability of survival over time

Explanation: Kaplan-Meier plots the survival function as a step plot of survival probabilities.

141. What is the main purpose of outlier detection in data analysis?

a) To improve data quality and prevent skewed model results

b) To find missing values

c) To increase dataset size

d) To remove all large values

Correct answer: a) To improve data quality and prevent skewed model results

Explanation: Outliers can distort statistical metrics and reduce model robustness.

142. Refer to this decision tree diagram:

What criterion is typically used by decision trees to decide on splits?

a) Information Gain based on entropy reduction

b) Random selection

c) Euclidean distance minimization

d) Manual feature engineering

Correct answer: a) Information Gain based on entropy reduction

Explanation: Splits that maximize the reduction in entropy or impurity are preferred.

143. Which of the following is NOT a principle of GDPR?

a) Purpose limitation

b) Data minimization

c) Indefinite data storage

d) Integrity and confidentiality

Correct answer: c) Indefinite data storage

Explanation: GDPR mandates data retention only as long as necessary.

144. What best describes the concept of ‘data provenance’?

a) History of the data lifecycle, including collection, processing, and transformations

b) Data corruption

c) Encryption

d) User privacy settings

Correct answer: a) History of the data lifecycle, including collection, processing, and transformations

Explanation: Provenance ensures traceability and reproducibility of data.

145. Which of these metrics is most appropriate for assessing model performance on imbalanced classes?

a) Balanced accuracy

b) Simple accuracy

c) Sum of residuals

d) Mean squared error

Correct answer: a) Balanced accuracy

Explanation: Balanced accuracy averages sensitivity and specificity, mitigating bias from class imbalance.