<https://docs.google.com/document/d/1DKihnvCxJDWFZ6eovnFjJ00HOiuWuUYfK2Xg4qTK3do/edit?tab=t.0#heading=h.7qdvksekgyss>

### **1. What is a hypothesis?**

**Options:**

1. A statement that the researcher wants to test through the data collected in a study
2. A research question the results will answer
3. A theory that underpins the study
4. A statistical method for calculating the extent to which the results could have happened by chance  
    **✅ Answer:** 1

### **2. As the number of training examples goes to infinity, your model trained on that data will have:**

**Options:**

1. Lower variance
2. Higher variance
3. Same variance
4. None of the above  
    **✅ Answer:** 1

### **3. Which of the following is/are true regarding an SVM?**

**Options:**

1. For two-dimensional data points, the separating hyperplane learnt by a linear SVM will be a straight line.
2. In theory, a Gaussian kernel SVM cannot model any complex separating hyperplane.
3. Overfitting in an SVM is not a function of the number of support vectors.
4. A and C  
    **✅ Answer:** 4

### **4. High entropy means that the partitions in classification are:**

**Options:**

1. Pure
2. Heterogeneous
3. Homogeneous
4. Useless  
    **✅ Answer:** 2

### **5. A machine learning problem involves four attributes plus a class. The attributes have 3, 2, 2, and 2 possible values each. The class has 3 possible values. How many maximum possible different examples are there?**

**Options:**

1. 12
2. 24
3. 48
4. 72  
    **✅ Answer:** 4  
    → Calculation: 3×2×2×2×3=723 × 2 × 2 × 2 × 3 = 72

### **6. Which of the following is NOT supervised learning?**

**Options:**

1. PCA
2. Decision Tree
3. Linear Regression
4. Naive Bayesian  
    **✅ Answer:** 1

### **7. Suppose we would like to perform clustering on spatial data such as the geometrical locations of houses. We wish to produce clusters of many different sizes and shapes. Which of the following methods is the most appropriate?**

**Options:**

1. Decision Trees
2. Density-based clustering
3. Model-based clustering
4. K-means clustering  
    **✅ Answer:** 2

### **8. Which of the following are the spatial clustering algorithms?**

**Options:**

1. Partitioning based clustering
2. K-means clustering
3. Grid based clustering
4. All of the above  
    **✅ Answer:** 4

### **9. Which of the following tasks can be best solved using Clustering?**

**Options:**

1. Predicting the amount of rainfall based on various cues
2. Detecting fraudulent credit card transactions
3. Training a robot to solve a maze
4. All of the above  
    **✅ Answer:** 2

### **10. Choose the correct option(s) from the following:**

**Options:** a) When working with a small dataset, one should prefer low bias/high variance classifiers over high bias/low variance classifiers.  
 b) When working with a small dataset, one should prefer high bias/low variance classifiers over low bias/high variance classifiers.  
 c) When working with a large dataset, one should prefer high bias/low variance classifiers over low bias/high variance classifiers.  
 d) When working with a large dataset, one should prefer low bias/high variance classifiers over high bias/low variance classifiers.

1. a and b
2. b and c
3. a and c
4. b and d  
    **✅ Answer:** 4

### **11. Predicting the amount of rainfall in a region based on various cues is a \_\_\_\_\_\_ problem.**

**Options:**

1. Supervised learning
2. Unsupervised learning
3. Clustering
4. None of the above  
    **✅ Answer:** 1

### **12. In building a linear regression model for a particular dataset, you observe the coefficient of one of the features having a relatively high negative value. This suggests that:**

**Options:**

1. The feature is not relevant to the target variable.
2. The feature has a weak positive relationship with the target variable.
3. The feature has a strong negative relationship with the target variable.
4. Nothing can be determined.  
    **✅ Answer:** 3

### **13. After applying a regularization penalty in linear regression, you find that some of the coefficients of w are zeroed out. Which of the following penalties might have been used?**

**Options:**

1. L0 norm
2. L1 norm
3. L2 norm
4. Either (a) or (b)  
    **✅ Answer:** 2

### **14. Maximum Likelihood Estimate (MLE) estimates are often undesirable because:**

**Options:**

1. They are biased
2. They have high variance
3. They are not consistent estimators
4. None of the above  
    **✅ Answer:** 2

### **15. Which of the following is a clustering algorithm in machine learning?**

**Options:**

1. Expectation Maximization
2. CART
3. Gaussian Naïve Bayes
4. Apriori  
    **✅ Answer:** 1

### **16. You've just finished training a decision tree for spam classification, and it is getting abnormally bad performance on both your training and test sets. What could be causing the problem?**

**Options:**

1. Your decision trees are too shallow.
2. You need to increase the learning rate.
3. You are overfitting
4. None of the above  
    **✅ Answer:** 1

### **17. \_\_\_\_\_\_\_\_\_\_\_ refers to a model that can neither model the training data nor generalize to new data.**

**Options:**

1. Good fitting
2. Overfitting
3. Underfitting
4. All of the above  
    **✅ Answer:** 3

### **18. Which of the following can only be used when training data are linearly separable?**

**Options:**

1. Linear hard-margin SVM
2. Linear Logistic Regression
3. Linear Soft margin SVM
4. The centroid method  
    **✅ Answer:** 1

### **19. The K-means algorithm:**

**Options:**

1. Requires the dimension of the feature space to be no bigger than the number of samples
2. Has the smallest value of the objective function when K = 1
3. Minimizes the within-class variance for a given number of clusters
4. Option B and C  
    **✅ Answer:** 4

### **20. For polynomial regression, which one of these structural assumptions most affects the trade-off between underfitting and overfitting?**

**Options:**

1. The polynomial degree
2. Whether we learn the weights by matrix inversion or gradient descent
3. The assumed variance of the Gaussian noise
4. The use of a constant-term unit input  
    **✅ Answer:** 1

Let me know if you want these formatted as a printable PDF or converted into a quiz format (Google Forms, Kahoot, etc.).

Here are the **Multiple Choice Questions with Options and Correct Answers** from your list, clearly formatted for study or quiz creation:

### **1. Which among the following prevents overfitting when we perform bagging?**

**Options:**

1. The use of sampling with replacement as the sampling technique
2. The use of weak classifiers
3. The use of classification algorithms which are not prone to overfitting
4. The practice of validation performed on every classifier trained  
    **✅ Answer:** 1

### **2. Averaging the output of multiple decision trees helps \_\_\_\_\_\_\_\_.**

**Options:**

1. Increase bias
2. Decrease bias
3. Increase variance
4. Decrease variance  
    **✅ Answer:** 4

### **3. K-means assumptions (assuming Euclidean distance):**

a) Clusters are spherical  
 b) Clusters are of similar size  
 c) Well-separated clusters  
 d) No wide variation in density  
 **Options:**

1. a and b
2. b and c
3. c and d
4. All of these  
    **✅ Answer:** 4

### **4. Which is more appropriate for feature selection?**

**Options:**

1. Ridge
2. Lasso
3. Both (a) and (b)
4. Neither (a) nor (b)  
    **✅ Answer:** 2

### **5. The number of test examples needed to get statistically significant results should be:**

**Options:**

1. Larger if the error rate is larger
2. Larger if the error rate is smaller
3. Smaller if the error rate is smaller
4. It does not matter  
    **✅ Answer:** 1

### **6. Most important factor in Kernel Regression affecting over/underfitting:**

**Options:**

1. Kernel type (Gaussian, etc.)
2. Distance metric used
3. The kernel width
4. Max height of kernel function  
    **✅ Answer:** 3

### **7. Main reason for pruning a Decision Tree:**

**Options:**

1. To save computing time
2. To save memory
3. To make training error smaller
4. To avoid overfitting  
    **✅ Answer:** 4

### **8. Bias-Variance correct statement(s):**

a) Overfit → High bias  
 b) Overfit → Low bias  
 c) Underfit → High variance  
 d) Underfit → Low variance  
 **Options:**

1. a and b
2. b and c
3. c and d
4. b and d  
    **✅ Answer:** 4

### **9. Learning task: Clustering heart disease patients for treatments**

**Options:**

1. Supervised learning
2. Unsupervised learning
3. Both
4. Neither  
    **✅ Answer:** 2

### **10. Predicting rain tomorrow evening is a type of:**

**Options:**

1. Classification
2. Regression
3. Unsupervised learning
4. All of the above  
    **✅ Answer:** 1

### **11. Proper data preprocessing sequence:**

**Options:**

1. Normalize → PCA → Train
2. PCA → Normalize PCA output → Train
3. Normalize → PCA → Normalize PCA output → Train
4. None of the above  
    **✅ Answer:** 3

### **12. Difference between supervised and unsupervised learning:**

**Options:**

1. Supervised needs labels, unsupervised doesn’t
2. Unsupervised needs labels, supervised doesn’t
3. Supervised doesn’t need data
4. No difference  
    **✅ Answer:** 1

### **13. Example of a classification problem:**

**Options:**

1. Predicting house price
2. Predicting weight from height
3. Predicting customer churn
4. Predicting age from income  
    **✅ Answer:** 3

### **14. Purpose of cross-validation in ML:**

**Options:**

1. Evaluate on held-out test set
2. Evaluate across different data subsets
3. Compare different models
4. Tune hyperparameters  
    **✅ Answer:** 2

### **15. Example of dimensionality reduction:**

**Options:**

1. PCA
2. SVM
3. KNN
4. AdaBoost  
    **✅ Answer:** 1

### **16. Example of a supervised learning problem:**

**Options:**

1. Image classification
2. Market segmentation
3. Fraud detection
4. Social network analysis  
    **✅ Answer:** 1 (or 3 – both are valid supervised learning problems)

### **17. Regularization techniques for linear regression:**

**Options:**

1. L1 (Lasso)
2. L2 (Ridge)
3. Dropout
4. Both a and b  
    **✅ Answer:** 4

### **18. Common dimensionality reduction approaches:**

**Options:**

1. Feature selection
2. Feature extraction
3. Feature scaling
4. Both a and b  
    **✅ Answer:** 4

### **19. Solving classification with imbalanced data:**

**Options:**

1. Resampling the data
2. Use different evaluation metric
3. Adjust class weights
4. All of the above  
    **✅ Answer:** 4

### **20. Common ensemble learning approaches:**

**Options:**

1. Bagging
2. Boosting
3. Stacking
4. All of the above  
    **✅ Answer:** 4

Let me know if you'd like:

* An **Excel sheet** or **PDF version** of these
* A **quiz app format** (like Google Forms, Kahoot, etc.)
* More MCQs in similar style for practice

Here is the list of MCQs along with **options** and the **correct answers**, formatted for quick review:

### **1. What is the purpose of regularization in machine learning?**

**Options:**

1. To reduce the number of features in a model
2. To prevent overfitting and improve generalization
3. To speed up the training process
4. To increase the accuracy of the model  
    **✅ Answer:** 2

### **2. Difference between a validation set and a test set:**

**Options:**

1. Validation tunes hyperparameters, test evaluates final performance
2. Validation evaluates during training, test evaluates after training
3. Validation and test sets are the same
4. Validation is not necessary  
    **✅ Answer:** 1

### **3. Example of a clustering algorithm:**

**Options:**

1. Decision tree
2. Random forest
3. K-means
4. Gradient descent  
    **✅ Answer:** 3

### **4. Purpose of cross-validation in ML:**

**Options:**

1. Evaluate on held-out test set
2. Evaluate on different subsets of data
3. Compare models
4. Tune hyperparameters  
    **✅ Answer:** 2

### **5. Purpose of data augmentation in ML:**

**Options:**

1. To increase the size of a dataset
2. To reduce the size of a dataset
3. To improve the quality of a dataset
4. None of these  
    **✅ Answer:** 1

### **6. Which is a hyperparameter in ML?**

**Options:**

1. Learning rate
2. Training data
3. Test data
4. Validation set  
    **✅ Answer:** 1

### **7. Evaluation metric for binary classification:**

**Options:**

1. Accuracy
2. F1 score
3. a, b, and d
4. Area under the ROC curve (AUC)  
    **✅ Answer:** 3

### **8. Handling missing data in ML:**

**Options:**

1. Remove missing data
2. Fill with constant
3. Fill with mean/median
4. All of the above  
    **✅ Answer:** 4

### **9. Bagging-based ML algorithm:**

**Options:**

1. Decision Tree
2. Regression
3. Classification
4. Random Forest  
    **✅ Answer:** 4

### **10. Correct preprocessing sequence for regression/classification:**

**Options:**

1. Normalize → PCA → Training
2. PCA → Normalize PCA output → Training
3. Normalize → PCA → Normalize PCA output → Training
4. None of the above  
    **✅ Answer:** 3

### **11. Reasonable way to select number of principal components k:**

**Options:**

1. Smallest k for 99% variance
2. k = 0.99 \* m
3. Largest k for 99% variance
4. Elbow method  
    **✅ Answer:** 1 (or 4 – both acceptable depending on context)

### **12. Model fails to capture data trend is called:**

**Options:**

1. Overfitting
2. Underfitting
3. Both A and B
4. None  
    **✅ Answer:** 2

### **13. Average squared difference between predicted & actual values:**

**Options:**

1. Mean relative error
2. Mean squared error
3. Mean absolute error
4. Root mean squared error  
    **✅ Answer:** 2

### **14. Feature F1 = {A, B, C, D, E, F} grades → type of feature:**

**Options:**

1. Nominal
2. Ordinal
3. Categorical
4. Boolean  
    **✅ Answer:** 2 (Assuming grades A-F imply order)

### **15. Which is true about SVM?**

**Options:**

1. Useful only in high-dimensions
2. Requires less memory
3. Doesn’t perform well with large datasets
4. Performs well with large datasets  
    **✅ Answer:** 3

### **16. Term for transforming data before algorithm use:**

**Options:**

1. Problem Identification
2. Identification of Required Data
3. Data Pre-processing
4. Definition of Training Set  
    **✅ Answer:** 3

### **17. Distance between hyperplane and data points is called:**

**Options:**

1. Hyper Plan
2. Margins
3. Error
4. Support Vectors  
    **✅ Answer:** 2

### **18. Measure of cluster quality:**

**Options:**

1. Purity
2. Distance
3. Accuracy
4. All of the above  
    **✅ Answer:** 1 (or 4 depending on interpretation – but typically *Purity* is direct)

### **19. What is Machine Learning?**

**Options:**

1. Autonomous learning via computer programs
2. Autonomous learning via manual programs
3. Selective learning via computer programs
4. Selective learning via manual programs  
    **✅ Answer:** 1

### **20. Language model categories do NOT include:**

**Options:**

1. System Unit
2. Structural Units
3. Data Units
4. Empirical Units  
    **✅ Answer:** 1

Let me know if you'd like this exported as a **CSV**, **PDF**, **Google Form**, or any other format for a quiz or training.

Here is the complete list of MCQs with **correct answers** highlighted:

### **1. Widely used ML algorithm based on bagging:**

**Options:**

1. Decision Tree
2. Regression
3. Classification
4. **✅ Random Forest**

### **2. Handling missing or corrupted data:**

**Options:**

1. Drop missing rows or columns
2. Replace missing values with mean/median/mode
3. Assign a unique category to missing values
4. **✅ All of the above**

### **3. K-means fails in which cases (Outliers, Densities, Nonconvex shapes):**

**Options:**

1. 1 and 2
2. 2 and 3
3. 1 and 3
4. **✅ All of the above**

### **4. Difference between supervised and unsupervised learning:**

**Options:**

1. **✅ Supervised learning requires labeled data while unsupervised does not**
2. Unsupervised learning requires labeled data
3. Supervised learning does not require data
4. No difference

### **5. Purpose of regularization:**

**Options:**

1. Reduce number of features
2. **✅ Prevent overfitting and improve generalization**
3. Speed up training
4. Increase accuracy

### **6. ML algorithm based on bagging (again):**

**✅ Answer:** 2. **Random Forest**

### **7. Disadvantage of decision trees:**

**Options:**

1. Robust to outliers
2. Factor Analysis
3. **✅ Prone to overfit**
4. All of the above

### **8. K-NN difficulties:**

**Options:**

1. Curse of dimensionality
2. Must calculate distance from all training samples
3. **✅ Both A and B**
4. None

### **9. Skills for a Data Scientist:**

**Options:**

1. Probability & Statistics
2. ML / Deep Learning
3. Data Wrangling
4. **✅ All of the above**

### **10. Unstructured data is not organized:**

**Options:**

1. **✅ TRUE**
2. FALSE
3. Can be true or false
4. Cannot say

### **11. Iterative process on ML models builds on:**

**Options:**

1. Mini-batches
2. Optimized parameters
3. **✅ Hyperparameters**
4. Superparameters

### **12. (Duplicate) ML algorithm based on bagging:**

**✅ Answer:** 4. **Random Forest**

### **13. Why set gradient to zero at extrema:**

**Options:**

1. **✅ Gradient is zero at extrema**
2. Depends on problem
3. Both A and B
4. None

### **14. Disadvantage of decision trees (again):**

**✅ Answer:** 3. **Decision trees are prone to be overfit**

### **15. Correct preprocessing flow for regression/classification:**

**Options:**

1. Normalize → PCA → Training
2. PCA → Normalize → Training
3. **✅ Normalize → PCA → Normalize → Training**
4. None

### **16. Incorrect statement about regularization:**

**Options:**

1. Too large lambda → underfit
2. **❌ Too large lambda → overfit**
3. **✅ Very large lambda cannot hurt → Not correct**
4. None

### **17. Not a normalization technique in text mining:**

**Options:**

1. Stemming
2. Lemmatization
3. Stop Word Removal
4. **✅ None of the above** (All are used)

### **18. K-means clustering limitations (repeated):**

**✅ Answer:** 4. **All of the above**

### **19. PCA true statements:**

**Options:**

1. PCA is unsupervised
2. It finds directions of max variance
3. Max components ≤ features
4. Components are orthogonal  
    **✅ Answer:** 4. **All of the above**

### **20. Clustering people by professional connection → Type of ML:**

**Options:**

1. **✅ Unsupervised learning**
2. Binary classification
3. Supervised learning
4. Reinforcement learning

Let me know if you'd like a **PDF, Excel sheet, or Google Form** created from these questions for teaching or testing purposes.

Here is a clean, structured list of the **MCQs with options and the correct answers marked** (✅) for easy reference or quiz preparation:

### **1. You are using the K-nearest neighbor and you have a K of 1. What are you likely to see when you train the model?**

1. ✅ High variance and low bias
2. Low bias and low variance
3. Low variance and high bias
4. High bias and high variance

### **2. \_\_\_ refers to a model that can neither model the training data nor generalize to new data.**

1. Good Fitting
2. Overfitting
3. ✅ Underfitting
4. All of the above

### **3. What is the best definition for bias in your data model?**

1. Bias is when your predicted values are scattered.
2. ✅ Bias is the gap between your predicted value and the outcome.
3. Bias is when your data is wrong for different reasons.
4. Bias is when your values are always off by the same percentage.

### **4. How is machine learning related to artificial intelligence?**

1. AI is about classification, ML is about clustering
2. ✅ ML is a type of AI that learns from data
3. AI is unsupervised ML
4. ML and AI are the same

### **5. ML project that adds most value to insurance company?**

1. ANN for company directory
2. ✅ Use ML to better predict risk
3. Consolidate Excel to data lake
4. Research salary using ML

### **6. Why not use the same data for both training and testing?**

1. You will underfit the model
2. You will pick the wrong algorithm
3. You might not have enough data
4. ✅ You will almost certainly overfit the model

### **7. What is conditional probability?**

1. ✅ Probability that doing one thing affects another
2. Probability certain conditions are met
3. Probability that something is always incorrect
4. Probability something is correct

### **8. PCA is**

1. Backward Feature Selection
2. Forward Feature Selection
3. ✅ Feature Extraction
4. All

### **9. Primary goal of feature engineering in ML?**

1. Increase model complexity
2. ✅ Improve model performance
3. Reduce number of features
4. Enhance data visualization

### **10. Common algorithm for binary classification?**

1. K-Means
2. Naive Bayes
3. Logistic Regression
4. ✅ Both b and c

### **11. In decision trees, what is the purpose of entropy?**

1. ✅ Measure of impurity
2. Speed up training
3. Regularize model
4. Control overfitting

### **12. What does "hyperparameter" refer to in ML?**

1. Parameters learned during training
2. Features of dataset
3. ✅ Parameters set before training
4. Output of model

### **13. Suitable metric for imbalanced classification problems?**

1. Accuracy
2. Precision
3. ✅ F1 Score
4. Mean Squared Error

### **14. Purpose of cross-validation in model evaluation?**

1. Speed up training
2. Increase complexity
3. ✅ Assess model performance on data subsets
4. Reduce overfitting

### **15. Which algorithm is an ensemble method?**

1. Decision Trees
2. K-Means
3. ✅ Gradient Boosting
4. Linear Regression

### **16. Main advantage of using SVM algorithm?**

1. Suitable for large datasets
2. Performs well with imbalanced data
3. ✅ Effective in high-dimensional spaces
4. Both c and d *(Note: d is missing, so correct is 3)*

### **17. What does "bias" refer to in ML?**

1. Error on training data
2. Error on unseen data
3. Difference between predicted and actual
4. ✅ Simplifying assumptions made by a model

### **18. Purpose of "cross-entropy loss" in ML?**

1. Minimizing prediction error
2. Regularizing model
3. ✅ Evaluating classification performance
4. Optimizing feature selection

### **19. Common technique to handle imbalanced classification datasets?**

1. Data augmentation
2. Feature scaling
3. ✅ SMOTE
4. PCA

### **20. Purpose of "grid search" in ML?**

1. Training on subsets
2. ✅ Searching for optimal hyperparameters
3. Evaluating model on cross-validation
4. Scaling input features

Let me know if you'd like this in **tabular format, PDF, CSV, or Google Form** for easy use in quizzes or assignments.

Here is a **well-structured MCQ list** with options and the ✅**correct answers** marked:

### **1. What does "ensemble" refer to in machine learning?**

1. ✅ A collection of trained models combined to improve performance
2. A single, complex model
3. The process of feature engineering
4. The training set of data

### **2. What is the purpose of the precision-recall trade-off?**

1. Balancing model bias and variance
2. Adjusting learning rate
3. ✅ Balancing precision and recall
4. Selecting SVM kernel function

### **3. Primary goal of k-fold cross-validation technique?**

1. Minimizing computational cost
2. Improving model interpretability
3. ✅ Evaluating model performance on data subsets
4. Reducing overfitting

### **4. Which technique reduces the curse of dimensionality?**

1. Feature engineering
2. Regularization
3. ✅ Dimensionality reduction
4. Both b and c

### **5. What does "precision" measure in classification?**

1. Ability to capture true positives
2. ✅ Ability to avoid false positives
3. Balance of true positives and negatives
4. Overall accuracy

### **6. Algorithm for anomaly detection in unsupervised learning?**

1. K-Means
2. Random Forest
3. ✅ Isolation Forest
4. Support Vector Machines

### **7. What does bias-variance trade-off mean?**

1. ✅ Balancing the trade-off between bias and variance
2. Optimizing training iterations
3. Selecting learning rate
4. Minimizing features

### **8. Ensemble method that builds models sequentially correcting errors?**

1. Bagging
2. ✅ Boosting
3. Stacking
4. Random Forest

### **9. reshape vs resize in NumPy:**

1. ✅ reshape returns modified shape; resize modifies in-place
2. resize, reshape
3. reshape2, resize
4. All of the mentioned

### **10. Confusion matrix is used for classification tasks with:**

1. ✅ Two or more
2. More than 2
3. Three
4. Two

### **11. Metric for evaluating continuous output model?**

1. Logloss
2. Accuracy
3. AUC-ROC
4. ✅ Mean-Squared-Error

### **12. Data close to each other and to actual value?**

1. ✅ Both precise and accurate
2. Accurate
3. None of the above
4. Precise

### **13. Residual is:**

1. ✅ Observed value - Predicted value
2. Predicted - Observed
3. Observed + Predicted
4. None

### **14. Three people weigh 2.00g mass and all read 7.32g. The balance is:**

1. Both accurate and precise
2. Accurate but not precise
3. ✅ Precise but not accurate
4. Neither

### **15. Hypothesis of logistic regression:**

1. To limit cost between 0 and +infinity
2. To limit cost between -1 and 1
3. ✅ To limit output between 0 and 1
4. Cost between -∞ and +∞

### **16. Data close to each other but far from actual value:**

1. Accurate
2. None
3. ✅ Precise
4. Both precise and accurate

### **17. ML is a field of AI with learning algorithms that:**

1. Improve performance
2. At executing tasks
3. Over time with experience
4. ✅ All of the above

### **18. Correct way to preprocess data for regression/classification:**

1. ✅ Normalize → PCA → Training
2. PCA → Normalize → Training
3. Normalize → PCA → Normalize → Training
4. None

### **19. Common evaluation metric for classification problems:**

1. MAE
2. RMSE
3. ✅ F1 Score
4. R-squared

### **20. Algorithm that is a type of ensemble learning:**

1. Decision Trees
2. KNN
3. SVM
4. ✅ Random Forest

Would you like this as a **PDF**, **quiz format**, or **CSV table**?

Here are the questions with their correct answers highlighted:

### **1. What is the purpose of feature scaling in machine learning?**

1. To increase model complexity
2. To decrease model complexity
3. **To normalize the range of independent variables**
4. To add more features to the dataset

### **2. Which method is used for handling missing data in a dataset?**

1. Data scaling
2. **Data imputation**
3. Data normalization
4. Data transformation

### **3. In PCA the number of input dimensions is equal to principal components.**

1. TRUE
2. **FALSE**

### **4. Which of the following is a widely used and effective machine learning algorithm based on the idea of bagging?**

1. Decision Tree
2. Adaboost
3. SVM
4. **Random Forest**

### **5. What is the purpose of one-hot encoding in machine learning?**

1. To encode numerical values as binary
2. To impute missing values
3. **To convert categorical variables into a binary matrix**
4. To standardize the data

### **6. What does the term "precision" refer to in a classification context?**

1. The ratio of true positives to the total number of positive predictions
2. The ratio of true positives to the sum of true positives and false negatives
3. **The ratio of true positives to the sum of true positives and false positives**
4. The ratio of true positives to the total number of actual positives in the dataset

### **7. Which method is used for handling imbalanced datasets in classification problems?**

1. Feature scaling
2. Data augmentation
3. **Oversampling and undersampling**
4. Principal Component Analysis (PCA)

### **8. What is the purpose of a ROC curve in binary classification?**

1. To evaluate model performance using precision and recall
2. **To visualize the trade-off between sensitivity and specificity**
3. To calculate the mean squared error
4. To determine the number of iterations in gradient descent

### **9. Which of the following is correct with respect to random forest?**

1. **Random forest are difficult to interpret but often very accurate**
2. Random forest are easy to interpret but often very accurate
3. Random forest are difficult to interpret but very less accurate
4. None of the mentioned

### **10. In linear regression, we try to \_\_\_\_\_\_ the least square errors of the model to identify the line of best fit.**

1. Change
2. Maximize
3. **Minimize**
4. None of the above

### **11. Which evaluation metric can be used to evaluate a model while modeling a continuous output variable?**

1. Logloss
2. Accuracy
3. AUC-ROC
4. **Mean-Squared-Error**

### **12. When performing regression or classification, which is the correct way to preprocess the data?**

1. **Normalize the data -> PCA -> training**
2. PCA -> normalize PCA output -> training
3. Normalize the data -> PCA -> normalize PCA output -> training
4. All of the above

### **13. Identify the false statement regarding regression:**

1. It is used for the prediction
2. It is used for interpretation
3. It relates inputs to outputs
4. **It discovers causal relationships**

### **14. Identify the difficulties with the k-nearest neighbor algorithm:**

1. Curse of dimensionality
2. Calculate the distance of the test case from all training cases
3. **Both A and B**
4. None of the above

### **15. Which of the following is not supervised learning?**

1. **PCA**
2. Naive Bayesian
3. Linear regression
4. Decision tree

### **16. What is the purpose of one-hot encoding in machine learning?**

1. To encode numerical values as binary
2. To impute missing values
3. **To convert categorical variables into a binary matrix**
4. To standardize the data

### **17. What does the term "precision" refer to in a classification context?**

1. The ratio of true positives to the total number of positive predictions
2. The ratio of true positives to the sum of true positives and false negatives
3. **The ratio of true positives to the sum of true positives and false positives**
4. The ratio of true positives to the total number of actual positives in the dataset

### **18. What is the key difference between supervised and unsupervised learning?**

1. **Supervised learning requires labeled data, while unsupervised learning does not.**
2. Supervised learning predicts labels, while unsupervised learning discovers patterns.
3. Supervised learning is used for classification, while unsupervised learning is used for regression.
4. Supervised learning is always more accurate than unsupervised learning.

### **19. Which type of machine learning algorithm falls under the category of 'unsupervised learning'?**

1. Linear Regression
2. **K-means Clustering**
3. Decision Trees
4. Random Forest

If you'd like, I can prepare these as a formatted quiz, flashcards, or another format for easy study!

Here are the questions with their correct answers highlighted:

### **1. Which of the following statements is false about Ensemble learning?**

1. It is a supervised learning algorithm  
    **2. It is an unsupervised learning algorithm**
2. More random algorithms can be used to produce a stronger ensemble
3. Ensembles can be shown to have more flexibility in the functions they can represent

### **2. Which of the following statements is true about AdaBoost?**

1. It is particularly prone to overfitting on noisy datasets
2. Complexity of the weak learner is important in AdaBoost
3. It is generally more prone to overfitting  
    **4. It improves classification accuracy**

### **3. What kind of algorithm is logistic regression?**

1. Cost function minimization
2. Ranking
3. Regression  
    **4. Classification**

### **4. The output in a logistic regression problem is yes (equivalent to 1 or true). What is its possible value?**

1. Greater than 0.5  
    **2. Depends on the algorithm's threshold value**
2. Greater than 0.6
3. Equal to 1

### **5. Which of the following statements is not true about SVM?**

1. It is memory efficient
2. It can address a large number of predictor variables
3. It is versatile  
    **4. It doesn't require feature scaling**

### **6. Which of the following statements is not true about the Decision tree?**

**1. It can be applied on binary classification problems only** 2. It is a predictor that predicts the label associated with an instance by traveling from a root node of a tree to a leaf  
 3. At each node, the successor child is chosen on the basis of a splitting of the input space  
 4. The splitting is based on one of the features or on a predefined set of splitting rules

### **7. Which of the following statements is not supporting in defining k Nearest Neighbor as a lazy learning algorithm?**

1. It defers data processing until it receives a request to classify unlabeled data
2. It replies to a request for information by combining its stored training data  
    **3. It stores all the intermediate results**
3. It discards the constructed answer

### **8. Euclidian distance and Manhattan distance are the same in kNN algorithm to calculate the distance.**

1. True  
    **2. False**

### **9. What type of plot is commonly used in Seaborn to visualize distributions of univariate data?**

**1. Histogram** 2. Bar plot  
 3. Scatter plot  
 4. Line plot

### **10. What type of data does a heatmap typically represent?**

**1. Continuous numerical data** 2. Categorical data  
 3. Text data  
 4. Any data type can be visualized using a heatmap

### **11. What criteria are typically used to select the best attribute for splitting at each node in the decision tree?**

1. Mean squared error
2. Euclidean distance, Manhattan distance, and cosine similarity  
    **3. Information gain, Gini impurity, and gain ratio**
3. Option A and C

### **12. How does L1 (LASSO) regularization work?**

**1. Penalizes the sum of the absolute values of the model coefficients (encourages sparsity).** 2. Penalizes the sum of the squared values of the model coefficients (encourages small coefficients).  
 3. Adds random noise to the training data to improve generalization.  
 4. Selects the most important features based on their correlation with the target variable.

### **13. What evaluation metric is appropriate for a binary classification problem where false positives are costly?**

1. Accuracy  
    **2. Precision**
2. Recall
3. F1-score

### **14. What are support vectors in a Support Vector Machine (SVM)?**

1. Misclassified data points used to update the model during training.
2. All data points in the training set.  
    **3. Data points closest to the decision boundary but classified correctly.**
3. Data points furthest from the decision boundary.

### **15. What are some common feature scaling techniques?**

1. Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA)
2. One-Hot encoding, Label encoding  
    **3. Min-Max scaling, Standardization (Z-score), Log scaling**
3. None of the above

### **16. The sigmoid function in logistic regression:**

**1. Transforms the linear combination of features into a probability between 0 and 1.** 2. Squares the predicted values to improve model accuracy.  
 3. Reduces the dimensionality of the feature space.  
 4. Calculates the distance between data points for classification.

### **17. What is the main objective of EDA in the context of Machine Learning?**

1. To build and train the best performing machine learning model.  
    **2. To understand the underlying structure and characteristics of the data.**
2. To evaluate the accuracy and performance of a trained model
3. To clean and pre-process the data for model training.

### **18. What problem does the kernel trick solve in machine learning?**

1. Overfitting complex models to limited data.
2. Handling imbalanced class distributions.
3. Efficiently computing distances in high-dimensional spaces.  
    **4. Representing non-linear relationships between features.**

### **19. What does a ROC curve visualize in the context of binary classification?**

1. The distribution of predicted probabilities for each class.
2. The confusion matrix elements plotted in two dimensions.
3. The accuracy of the model across different data subsets.  
    **4. The relationship between precision and recall at different classification thresholds.**

### **20. Which distance metric is commonly used in K-Means clustering?**

**1. Euclidean distance** 2. Manhattan distance  
 3. Jaccard distance  
 4. All of the above can be used, depending on the data.

If you'd like, I can help explain any of these answers or concepts!

Here are the questions with their correct answers highlighted:

### **1. The model will be trained with data in one single batch is known as?**

1. Batch learning
2. Offline learning  
    **3. Both A and B**
3. None of the above

### **2. How do you handle missing or corrupted data in a dataset?**

1. Drop missing rows or columns
2. Replace missing values with mean/median/mode
3. Assign a unique category to missing values  
    **4. All of the above**

### **3. When performing regression or classification, which of the following is the correct way to preprocess the data?**

1. Normalize the data -> PCA -> training
2. PCA -> normalize PCA output -> training
3. Normalize the data -> PCA -> normalize PCA output -> training  
    **4. None of the above** (Because generally normalize first, then PCA, and then train without further normalization of PCA output)

### **4. \_\_\_\_\_\_\_\_ is a machine learning technique which helps to detect outliers in data?**

1. Classification
2. Regression
3. Clustering  
    **4. Anomaly detection**

### **5. Which of the following learning algorithms is used to learn 'facial identities for facial expressions'?**

1. Supervised learning
2. Unsupervised learning
3. Reinforcement learning  
    **4. Deep learning**

### **6. Select the metrics and tools that are most frequently used to evaluate classification models.**

**1. Accuracy, precision, recall, F1 score, and confusion matrix** 2. Mean squared error, R-squared, and residual plots  
 3. Silhouette coefficient, Rand index, and Davies-Bouldin index  
 4. Area under the ROC curve, precision-recall curve, and lift curve

### **7. What is bagging in ensemble learning?**

1. Combining multiple models by averaging their predictions  
    **2. Training multiple models independently on different subsets of the training data**
2. Using a single model to make predictions
3. None of the above

### **8. Which of the following statements is true about AdaBoost?**

1. It is particularly prone to overfitting on noisy datasets
2. Complexity of the weak learner is important in AdaBoost
3. It is generally more prone to overfitting  
    **4. It improves classification accuracy**

### **9. Which of the following statements is not true about SVM?**

1. It has regularization capabilities
2. It handles non-linear data efficiently
3. It has much improved stability  
    **4. Choosing an appropriate kernel function is easy**

### **10. Which of the following statements is not true about optimal separating hyperplane?**

1. It correctly classifies the training data
2. It is the one which will generalize better with unseen data
3. Finding the optimal separating hyperplane can be formulated as a convex quadratic programming problem  
    **4. The optimal hyperplane cannot correctly classify all the data while being farthest away from the data points**

### **11. Which of the following is not a supervised machine learning algorithm?**

**1. k means** 2. naive bayes  
 3. svm  
 4. decision tree

### **12. Which of the following statements is true about AdaBoost?**

1. It is particularly prone to overfitting on noisy datasets
2. Complexity of the weak learner is important in AdaBoost
3. It is generally more prone to overfitting  
    **4. It improves classification accuracy**

If you want me to explain any of these answers or concepts further, just ask!