IBM Machine Learning Professional Certificate

Quiz Answers 2026

92 solved MCQs with explanations, organised by course: EDA, regression, classification, unsupervised learning, deep learning and reinforcement learning, and the recommender-systems capstone.

Free PDF from mcqstop.com

Updated 16 November 2026. Online version with new questions: mcqstop.com/ibm-machine-learning-professional-certificate-answers/

These are practice questions covering the course topics, written for revision. They are not copied from IBM's graded assessments.

Course 1: Exploratory Data Analysis for Machine Learning

1. What is the main goal of Exploratory Data Analysis (EDA)?

A. To deploy the model
B. To understand the data's structure, patterns and problems before modelling
C. To write the final report
D. To tune hyperparameters

Answer: B. To understand the data's structure, patterns and problems before modelling
EDA uses summaries and visualisations to guide cleaning and feature choices.

2. Which pandas method shows summary statistics for numeric columns?

A. .info()
B. .describe()
C. .head()
D. .shape

Answer: B. .describe()
describe() returns count, mean, std, min, quartiles and max.

3. How can missing values be handled?

A. Always delete the column
B. Remove rows, impute with mean/median/mode, or mask and model them
C. Ignore them
D. Replace with zero only

Answer: B. Remove rows, impute with mean/median/mode, or mask and model them
The right choice depends on how much is missing and why.

4. What is an outlier?

A. A missing value
B. An observation far from the rest of the data
C. A duplicated row
D. A categorical variable

Answer: B. An observation far from the rest of the data
Detect with box plots, z-scores or the IQR rule.

5. The IQR rule flags a point as an outlier if it lies:

A. Within Q1 and Q3
B. More than 1.5 x IQR below Q1 or above Q3
C. Above the mean
D. Below the median

Answer: B. More than 1.5 x IQR below Q1 or above Q3
IQR = Q3 - Q1.

6. What is feature scaling?

A. Adding features
B. Transforming features to a comparable range or distribution
C. Removing features
D. Encoding labels

Answer: B. Transforming features to a comparable range or distribution
Needed for distance-based and gradient-based models.

7. Standard scaling transforms a feature to have:

A. Min 0 and max 1
B. Mean 0 and standard deviation 1
C. Median 0
D. Values between -1 and 1 only

Answer: B. Mean 0 and standard deviation 1
StandardScaler subtracts the mean and divides by the standard deviation.

8. Min-max scaling transforms a feature to:

A. Mean 0, std 1
B. A fixed range, usually 0 to 1
C. A normal distribution
D. Integers

Answer: B. A fixed range, usually 0 to 1
It is sensitive to outliers.

9. One-hot encoding is used to:

A. Scale numeric features
B. Convert a categorical variable into binary columns
C. Remove outliers
D. Reduce dimensions

Answer: B. Convert a categorical variable into binary columns
Each category becomes its own 0/1 column.

10. Ordinal encoding is appropriate when categories:

A. Have no order
B. Have a natural order, such as low/medium/high
C. Are numeric
D. Are missing

Answer: B. Have a natural order, such as low/medium/high
It maps ordered categories to integers.

11. A log transformation is commonly applied to:

A. Normally distributed data
B. Right-skewed data to make it more symmetric
C. Categorical data
D. Binary data

Answer: B. Right-skewed data to make it more symmetric
It compresses large values.

12. What does a correlation coefficient of -0.9 indicate?

A. No relationship
B. A strong negative linear relationship
C. A weak positive relationship
D. Causation

Answer: B. A strong negative linear relationship
Correlation ranges from -1 to 1 and does not imply causation.

13. What is a null hypothesis?

A. The hypothesis you want to prove
B. The default assumption of no effect or no difference
C. A proven fact
D. A model parameter

Answer: B. The default assumption of no effect or no difference
Tests assess evidence against it.

14. A p-value of 0.01 with a significance level of 0.05 means:

A. Fail to reject the null
B. Reject the null hypothesis
C. The null is proven true
D. The test is invalid

Answer: B. Reject the null hypothesis
The result is statistically significant.

15. A Type I error is:

A. Failing to reject a false null
B. Rejecting a true null hypothesis (false positive)
C. A data entry mistake
D. Overfitting

Answer: B. Rejecting a true null hypothesis (false positive)
Its probability is the significance level alpha.

16. Which plot is best for viewing the distribution of a single numeric variable?

A. Scatter plot
B. Histogram
C. Pair plot
D. Heat map

Answer: B. Histogram
Box plots and KDE plots also work.

Course 2: Supervised Machine Learning - Regression

17. In linear regression, the cost function usually minimised is:

A. Cross-entropy
B. Mean squared error
C. Hinge loss
D. Gini impurity

Answer: B. Mean squared error
OLS minimises the sum of squared residuals.

18. Polynomial regression is used to:

A. Reduce features
B. Model non-linear relationships by adding powers of features
C. Classify data
D. Cluster data

Answer: B. Model non-linear relationships by adding powers of features
It is still linear in the coefficients.

19. What is regularisation?

A. Removing data
B. Adding a penalty on coefficient size to reduce overfitting
C. Scaling features
D. Shuffling data

Answer: B. Adding a penalty on coefficient size to reduce overfitting
It trades a little bias for lower variance.

20. Ridge regression uses which penalty?

A. L1 (absolute values)
B. L2 (squared values)
C. No penalty
D. Elastic

Answer: B. L2 (squared values)
Ridge shrinks coefficients toward zero but rarely to exactly zero.

21. Lasso regression uses which penalty?

A. L2
B. L1
C. L0
D. None

Answer: B. L1
Lasso can set coefficients to exactly zero, performing feature selection.

22. Elastic Net combines:

A. Ridge and decision trees
B. L1 and L2 penalties
C. Lasso and KNN
D. PCA and Ridge

Answer: B. L1 and L2 penalties
A mixing parameter balances the two.

23. What is the bias-variance trade-off?

A. More data always helps
B. Simple models have high bias; complex models have high variance; the goal is the balance with lowest total error
C. Bias equals variance
D. Variance is always bad

Answer: B. Simple models have high bias; complex models have high variance; the goal is the balance with lowest total error
Underfitting is high bias; overfitting is high variance.

24. Why must features be scaled before Ridge or Lasso?

A. To speed up pandas
B. Because the penalty depends on coefficient magnitude, which depends on feature scale
C. It is not needed
D. To remove outliers

Answer: B. Because the penalty depends on coefficient magnitude, which depends on feature scale
Unscaled features are penalised unevenly.

25. K-fold cross-validation with K=5 trains the model:

A. Once
B. Five times, each time holding out a different fold
C. Ten times
D. On the test set

Answer: B. Five times, each time holding out a different fold
Scores are averaged across folds.

26. What does a learning curve show?

A. Loss vs epoch only
B. Training and validation error as training set size increases
C. Feature importance
D. Correlation

Answer: B. Training and validation error as training set size increases
It helps diagnose bias and variance.

27. What is the purpose of a validation set in hyperparameter tuning?

A. Final evaluation
B. Choosing hyperparameters without touching the test set
C. Training
D. Scaling

Answer: B. Choosing hyperparameters without touching the test set
The test set is used only once at the end.

28. In scikit-learn, GridSearchCV is used to:

A. Scale data
B. Search over hyperparameter combinations with cross-validation
C. Plot results
D. Encode categories

Answer: B. Search over hyperparameter combinations with cross-validation
It returns the best estimator and parameters.

29. R-squared can be misleading because:

A. It is always 1
B. It never decreases when more features are added
C. It is negative
D. It ignores the target

Answer: B. It never decreases when more features are added
Adjusted R-squared penalises extra features.

30. What does a residual plot help check?

A. Class balance
B. Whether errors are randomly scattered (linearity and constant variance)
C. Cluster shape
D. Missing values

Answer: B. Whether errors are randomly scattered (linearity and constant variance)
Patterns suggest a mis-specified model.

31. What does 'alpha' control in Ridge and Lasso?

A. Learning rate
B. Strength of the regularisation penalty
C. Number of folds
D. Polynomial degree

Answer: B. Strength of the regularisation penalty
Higher alpha means stronger shrinkage.

32. Which is an example of a regression error metric robust to outliers?

A. MSE
B. Mean Absolute Error (MAE)
C. R-squared
D. RMSE

Answer: B. Mean Absolute Error (MAE)
MAE does not square the errors.

Course 3: Supervised Machine Learning - Classification

33. Logistic regression predicts:

A. A continuous value
B. The probability of class membership
C. Cluster labels
D. Principal components

Answer: B. The probability of class membership
A threshold (usually 0.5) converts it to a class.

34. What does the ROC curve plot?

A. Precision vs recall
B. True positive rate vs false positive rate at different thresholds
C. Loss vs epochs
D. Accuracy vs depth

Answer: B. True positive rate vs false positive rate at different thresholds
AUC summarises it; 0.5 is random, 1 is perfect.

35. The F1 score is the:

A. Arithmetic mean of precision and recall
B. Harmonic mean of precision and recall
C. Product of precision and recall
D. Difference between them

Answer: B. Harmonic mean of precision and recall
It balances the two when classes are imbalanced.

36. Why is accuracy misleading for imbalanced classes?

A. It is hard to compute
B. A model can score high by always predicting the majority class
C. It ignores true positives
D. It needs probabilities

Answer: B. A model can score high by always predicting the majority class
Use precision, recall, F1 or AUC instead.

37. What does SMOTE do?

A. Removes outliers
B. Creates synthetic minority-class samples to balance classes
C. Scales features
D. Selects features

Answer: B. Creates synthetic minority-class samples to balance classes
Synthetic Minority Over-sampling Technique.

38. Which techniques address class imbalance?

A. Oversampling, undersampling and class weights
B. Scaling only
C. PCA only
D. Dropping the target

Answer: A. Oversampling, undersampling and class weights
Stratified sampling keeps class ratios in splits.

39. In KNN, the 'K' is:

A. The number of clusters
B. The number of nearest neighbours considered
C. The number of features
D. The number of trees

Answer: B. The number of nearest neighbours considered
Chosen by cross-validation.

40. In SVM, the regularisation parameter C controls:

A. Kernel type
B. The trade-off between a wide margin and misclassification
C. Number of support vectors directly
D. Learning rate

Answer: B. The trade-off between a wide margin and misclassification
Small C gives a wider margin and more tolerance of errors.

41. A decision tree splits to maximise:

A. Entropy
B. Information gain / reduction in impurity
C. Depth
D. Number of leaves

Answer: B. Information gain / reduction in impurity
Gini impurity and entropy are common criteria.

42. How do you reduce overfitting in a decision tree?

A. Grow it deeper
B. Prune it or limit max depth and minimum samples per leaf
C. Add more features
D. Remove the root

Answer: B. Prune it or limit max depth and minimum samples per leaf
Unrestricted trees memorise the training data.

43. What is bagging?

A. Training models sequentially on errors
B. Training many models on bootstrap samples and averaging their predictions
C. Using one large tree
D. Scaling data

Answer: B. Training many models on bootstrap samples and averaging their predictions
Bootstrap aggregating reduces variance.

44. A random forest improves on bagged trees by:

A. Using deeper trees
B. Also choosing a random subset of features at each split
C. Using fewer trees
D. Removing bootstrapping

Answer: B. Also choosing a random subset of features at each split
This decorrelates the trees.

45. What is boosting?

A. Training trees independently
B. Training models sequentially, each correcting the errors of the previous ones
C. Averaging random trees
D. Removing weak learners

Answer: B. Training models sequentially, each correcting the errors of the previous ones
Examples: AdaBoost, Gradient Boosting, XGBoost.

46. In AdaBoost, misclassified samples are:

A. Removed
B. Given higher weight in the next round
C. Ignored
D. Duplicated in the test set

Answer: B. Given higher weight in the next round
Later learners focus on hard cases.

47. What is stacking?

A. Deep trees
B. Combining different models by training a meta-model on their predictions
C. Bagging with one model
D. Feature scaling

Answer: B. Combining different models by training a meta-model on their predictions
The meta-learner learns how to weigh the base models.

48. What does 'out-of-bag' error estimate in a random forest?

A. Training error
B. Generalisation error using samples not in each tree's bootstrap
C. Feature count
D. Tree depth

Answer: B. Generalisation error using samples not in each tree's bootstrap
About one-third of samples are out-of-bag for each tree.

49. What is model interpretability important for?

A. Speed
B. Trust, debugging and regulatory compliance
C. Storage
D. Scaling

Answer: B. Trust, debugging and regulatory compliance
Tools include feature importance, partial dependence, LIME and SHAP.

50. Which model is generally the most interpretable?

A. Gradient boosting
B. A shallow decision tree or logistic regression
C. Deep neural network
D. Random forest of 500 trees

Answer: B. A shallow decision tree or logistic regression
Simple models are easier to explain.

Course 4: Unsupervised Machine Learning

51. K-means requires you to specify:

A. The labels
B. The number of clusters K
C. The tree depth
D. The learning rate

Answer: B. The number of clusters K
K is chosen with the elbow method or silhouette score.

52. What is 'inertia' in K-means?

A. The number of iterations
B. The sum of squared distances from points to their cluster centroid
C. The number of clusters
D. The data size

Answer: B. The sum of squared distances from points to their cluster centroid
Lower inertia means tighter clusters.

53. The elbow method helps choose:

A. The learning rate
B. The number of clusters where adding more gives little improvement
C. The kernel
D. The depth

Answer: B. The number of clusters where adding more gives little improvement
Plot inertia against K and look for the bend.

54. K-means++ improves K-means by:

A. Using more clusters
B. Choosing initial centroids that are far apart
C. Removing outliers
D. Scaling data

Answer: B. Choosing initial centroids that are far apart
It leads to faster and better convergence.

55. Hierarchical agglomerative clustering starts with:

A. One big cluster
B. Each point as its own cluster and merges the closest pairs
C. K centroids
D. A density threshold

Answer: B. Each point as its own cluster and merges the closest pairs
The result is shown as a dendrogram.

56. Which linkage uses the minimum distance between points of two clusters?

A. Complete
B. Single
C. Average
D. Ward

Answer: B. Single
Ward linkage minimises the increase in within-cluster variance.

57. DBSCAN groups points based on:

A. Distance to centroids
B. Density, using epsilon and minimum samples
C. A tree
D. Probabilities

Answer: B. Density, using epsilon and minimum samples
It finds arbitrary shapes and labels noise points.

58. An advantage of DBSCAN over K-means is that it:

A. Needs K specified
B. Can find non-spherical clusters and identify outliers
C. Is always faster
D. Works without parameters

Answer: B. Can find non-spherical clusters and identify outliers
It does not require the number of clusters.

59. Mean Shift clustering finds clusters by:

A. Splitting trees
B. Shifting points toward the densest region (mode)
C. Random assignment
D. Linkage

Answer: B. Shifting points toward the densest region (mode)
The bandwidth parameter controls the result.

60. What is the 'curse of dimensionality'?

A. Too little data
B. As dimensions grow, data becomes sparse and distances less meaningful
C. Too many rows
D. Slow computers

Answer: B. As dimensions grow, data becomes sparse and distances less meaningful
It motivates dimensionality reduction.

61. PCA reduces dimensions by:

A. Removing rows
B. Projecting data onto directions of maximum variance
C. Clustering features
D. Encoding categories

Answer: B. Projecting data onto directions of maximum variance
Principal components are orthogonal.

62. In PCA, the 'explained variance ratio' tells you:

A. The number of clusters
B. How much of the total variance each component captures
C. The error rate
D. The feature names

Answer: B. How much of the total variance each component captures
Keep enough components to explain, say, 95%.

63. Why scale data before PCA?

A. It is optional
B. Features with larger scales would dominate the components
C. To add variance
D. To cluster

Answer: B. Features with larger scales would dominate the components
PCA is variance-based.

64. Kernel PCA is used when:

A. Data is linear
B. The structure in the data is non-linear
C. There are no features
D. Data is categorical

Answer: B. The structure in the data is non-linear
It applies the kernel trick before PCA.

65. Non-negative Matrix Factorization (NMF) requires that:

A. Data be normalised to mean 0
B. All data values be non-negative
C. K equals 2
D. Data be categorical

Answer: B. All data values be non-negative
Used for topic modelling and image decomposition.

66. Which distance metric suits text data represented as word counts?

A. Euclidean
B. Cosine distance
C. Manhattan
D. Chebyshev

Answer: B. Cosine distance
It compares direction rather than magnitude.

Course 5: Deep Learning and Reinforcement Learning

67. What is a perceptron?

A. A clustering model
B. The simplest neural unit: weighted sum plus a step activation
C. An optimiser
D. A loss

Answer: B. The simplest neural unit: weighted sum plus a step activation
Multilayer perceptrons stack them with non-linear activations.

68. What does the learning rate control?

A. Number of layers
B. The size of each weight update step
C. Batch size
D. Number of epochs

Answer: B. The size of each weight update step
Too high diverges; too low is slow.

69. Stochastic gradient descent updates weights using:

A. The whole dataset
B. One sample (or a small batch) at a time
C. No data
D. The test set

Answer: B. One sample (or a small batch) at a time
It is noisier but faster per update.

70. Which is a regularisation technique for neural networks?

A. Adding layers
B. Dropout, weight decay and early stopping
C. Larger learning rate
D. More epochs

Answer: B. Dropout, weight decay and early stopping
They reduce overfitting.

71. A CNN's convolution layer learns:

A. Class labels
B. Filters that detect local patterns such as edges
C. Word order
D. Rewards

Answer: B. Filters that detect local patterns such as edges
Weights are shared across the image.

72. Transfer learning for images typically reuses:

A. The output layer
B. Early convolutional layers trained on a large dataset
C. The labels
D. The optimiser

Answer: B. Early convolutional layers trained on a large dataset
Later layers are retrained for the new task.

73. An RNN suffers from which problem on long sequences?

A. Too many filters
B. Vanishing or exploding gradients
C. Too few parameters
D. No memory

Answer: B. Vanishing or exploding gradients
LSTM and GRU cells address it.

74. A GRU differs from an LSTM in that it:

A. Has more gates
B. Has fewer gates and no separate cell state
C. Cannot handle sequences
D. Is not recurrent

Answer: B. Has fewer gates and no separate cell state
It is simpler and often trains faster.

75. An autoencoder's bottleneck layer:

A. Is the output
B. Holds the compressed representation of the input
C. Holds labels
D. Is unused

Answer: B. Holds the compressed representation of the input
The decoder reconstructs the input from it.

76. A variational autoencoder (VAE) differs from a standard autoencoder by:

A. Having no decoder
B. Learning a probability distribution in the latent space so it can generate new samples
C. Using labels
D. Being a classifier

Answer: B. Learning a probability distribution in the latent space so it can generate new samples
VAEs are generative models.

77. In a GAN, the generator's goal is to:

A. Classify images
B. Produce samples the discriminator cannot tell from real data
C. Reduce dimensions
D. Label data

Answer: B. Produce samples the discriminator cannot tell from real data
The two networks are trained adversarially.

78. In reinforcement learning, the 'policy' is:

A. The reward
B. The agent's strategy mapping states to actions
C. The environment
D. The dataset

Answer: B. The agent's strategy mapping states to actions
The aim is to find the policy that maximises expected return.

79. What is the 'exploration vs exploitation' trade-off?

A. Training vs testing
B. Trying new actions vs using actions known to give reward
C. Bias vs variance
D. Speed vs memory

Answer: B. Trying new actions vs using actions known to give reward
Epsilon-greedy is a simple strategy.

80. Q-learning learns:

A. A classifier
B. The value of taking each action in each state
C. A cluster map
D. An embedding

Answer: B. The value of taking each action in each state
It is model-free and off-policy.

81. The discount factor gamma in reinforcement learning determines:

A. Learning rate
B. How much future rewards are valued relative to immediate ones
C. The number of states
D. Exploration rate

Answer: B. How much future rewards are valued relative to immediate ones
Gamma near 0 is short-sighted; near 1 is far-sighted.

82. Which library does the certificate use for deep learning labs?

A. scikit-learn only
B. Keras (TensorFlow)
C. OpenCV
D. NLTK

Answer: B. Keras (TensorFlow)
Models are built with the Keras API.

Course 6: Machine Learning Capstone

83. What is a recommender system?

A. A classifier of images
B. A system that suggests items to users based on data about preferences
C. A clustering tool
D. A search engine

Answer: B. A system that suggests items to users based on data about preferences
The capstone builds course recommenders.

84. Content-based filtering recommends items:

A. Liked by similar users
B. Similar to items the user liked, using item features
C. At random
D. Most popular overall

Answer: B. Similar to items the user liked, using item features
It uses item profiles such as genres or keywords.

85. Collaborative filtering recommends items based on:

A. Item descriptions
B. Behaviour of similar users or similar items
C. Random choice
D. Price

Answer: B. Behaviour of similar users or similar items
It needs interaction data, not item features.

86. What is the 'cold start' problem?

A. Slow servers
B. Difficulty recommending for new users or items with no history
C. Too much data
D. Overfitting

Answer: B. Difficulty recommending for new users or items with no history
Content-based methods help with new items.

87. A user-item interaction matrix is usually:

A. Dense
B. Sparse, because each user rates few items
C. Square
D. Binary only

Answer: B. Sparse, because each user rates few items
Matrix factorisation fills in the gaps.

88. Matrix factorisation for recommendations decomposes the interaction matrix into:

A. Clusters
B. User and item latent-factor matrices
C. Trees
D. Principal labels

Answer: B. User and item latent-factor matrices
NMF and SVD-based methods are common.

89. Which similarity measure is common in collaborative filtering?

A. Hamming
B. Cosine similarity or Pearson correlation
C. Edit distance
D. Jaccard only

Answer: B. Cosine similarity or Pearson correlation
They compare user or item rating vectors.

90. How can a neural network be used for recommendation?

A. It cannot
B. By learning user and item embeddings and predicting ratings from them
C. Only for images
D. Only for clustering

Answer: B. By learning user and item embeddings and predicting ratings from them
The capstone trains an embedding-based model in Keras.

91. Which metric evaluates predicted ratings?

A. Accuracy
B. RMSE
C. Silhouette
D. Inertia

Answer: B. RMSE
Lower RMSE means closer predictions.

92. Which tool does the capstone use to build an interactive app for the recommender?

A. Excel
B. Streamlit
C. PowerPoint
D. Tableau

Answer: B. Streamlit
Streamlit turns Python scripts into web apps.