FINAL QUESTION PAPER

1. 19. When using sklearn.model_selection.train_test_split(), what is the typical number of arrays returned? (a) Two (X_train, y_train) (b) Three (X_train, X_test, y_train) (c) Four (X_train, X_test, y_train, y_test) (d) Five (X_train, X_test, y_train, y_test, scaler)

2. 32. What is the primary purpose of using the pandas read_csv() function in a machine learning workflow? (a) To visualize data distributions. (b) To save trained models to disk. (c) To load a dataset from a CSV file into a DataFrame. (d) To perform statistical tests on the dataset.

3. 5. What does the acronym "sklearn" primarily refer to in the context of Machine Learning? a) Scikit-learn b) Scala-kernel c) Scientific-kernel d) Script-learn

4. 22. Which of the following scikit-learn classes is an example of a supervised learning estimator? (a) KMeans (b) StandardScaler (c) LogisticRegression (d) PCA

5. 4. Which Python library is specifically designed for numerical operations and array manipulation, often serving as a foundation for other ML libraries? a) Pandas b) Matplotlib c) NumPy d) SciPy

6. 37. How many distinct arrays does the train_test_split() function typically return when splitting a features (X) and target (y) dataset? (a) Two (X_train, y_train) (b) Three (X_train, X_test, y_train) (c) Four (X_train, X_test, y_train, y_test) (d) Five (X_train, X_test, y_train, y_test, random_state)

7. 28. In the context of scikit-learn, what type of estimator can also be referred to as a "Predictor"? (a) Any estimator that has a transform() method. (b) Any estimator that has a fit() method. (c) Any estimator that can make predictions, typically having a predict() method. (d) Only estimators used for unsupervised learning.

8. 7. In the context of building an ML model, what is the purpose of splitting a dataset into training and testing sets? a) To speed up the model training process. b) To prevent the model from accessing real-world data. c) To evaluate the model's performance on unseen data. d) To combine different types of data for training.

9. 27. The "Model API" in scikit-learn primarily refers to a consistent interface for: (a) Data visualization tools. (b) Data loading utilities. (c) Estimator methods like fit(), predict(), and transform(). (d) Advanced hyperparameter optimization techniques.

10. 36. What is the main objective of using the train_test_split() function from sklearn.model_selection? (a) To reduce the dimensionality of the dataset. (b) To preprocess categorical features into numerical ones. (c) To divide a dataset into training and testing subsets for model evaluation. (d) To aggregate multiple datasets into a single one.

11. 43. In a typical supervised machine learning workflow, what does 'y' commonly represent after loading and initial data preparation? (a) The feature matrix or independent variables. (b) The target variable or dependent variable. (c) The training error. (d) The model's accuracy score.

12. 20. A "Transformer" in scikit-learn is primarily used for what type of task? (a) Predicting future outcomes. (b) Evaluating model accuracy. (c) Data preprocessing and feature engineering. (d) Visualizing complex datasets.

13. 31. Which Python library is primarily used for loading tabular data from a CSV file into a DataFrame using the read_csv() function? (a) numpy (b) pandas (c) sklearn (d) matplotlib

14. 41. Why is it a standard practice in machine learning to split a dataset into training and testing sets? (a) To increase the overall size of the dataset for training. (b) To prevent the model from overfitting to the training data and to evaluate its generalization performance. (c) To speed up the training process by using smaller subsets. (d) To simplify feature engineering by isolating subsets of data.

15. 8. Which Scikit-learn function is typically used to split a dataset into training and testing subsets? a) data_split() b) split_data_set() c) train_test_split() d) divide_dataset()

16. 15. What does the 'test_size' parameter in train_test_split() function typically control? a) The number of features in the test set. b) The proportion of the dataset to include in the test split. c) The random seed for reproducibility. d) The maximum size of the training set.

17. 3. Why is Python a preferred language for Machine Learning? a) It is the only language that supports mathematical operations. b) It has a simple syntax and a vast ecosystem of scientific libraries. c) It compiles to machine code faster than C++. d) It only runs on specific proprietary hardware.

18. 33. When loading a dataset using pandas.read_csv('data.csv'), what is the default Python data structure that the function returns? (a) A NumPy array (b) A Pandas Series (c) A Pandas DataFrame (d) A Python dictionary

19. 16. What is an "Estimator" in scikit-learn? (a) A function used for plotting data visualizations. (b) An object that learns from data, often by fitting a model to it. (c) A method for evaluating model performance metrics. (d) A utility for loading built-in datasets.

20. 10. What is the standard method name used by Scikit-learn estimators to train a model on the provided data? a) predict() b) transform() c) fit() d) score()

21. 18. In scikit-learn, what is the primary purpose of the transform() method for a preprocessor or transformer object? (a) To train the model on the input data. (b) To make predictions based on the trained model. (c) To apply learned transformations to new data. (d) To calculate the accuracy of the model.

22. 2. Which of the following is NOT a common type of Machine Learning? a) Supervised Learning b) Unsupervised Learning c) Reinforcement Learning d) Algorithmic Learning

23. 35. To treat the first row of a CSV file as data rather than column headers when loading with read_csv(), which argument would you typically set? (a) header=True (b) header=None (c) skip_rows=1 (d) index_col=False

24. 39. What is the primary reason for setting the random_state parameter in the train_test_split() function? (a) To ensure that the training and testing sets are always sorted. (b) To control the specific algorithm used for splitting. (c) To make the data split reproducible across different runs. (d) To randomly shuffle the dataset before splitting.

25. 26. While not part of scikit-learn itself, what pandas function is commonly used to load tabular datasets into a DataFrame before applying scikit-learn models? (a) pandas.load_data() (b) pandas.read_csv() (c) pandas.import_table() (d) pandas.get_dataframe()

26. 25. What is the primary purpose of the random_state parameter in the train_test_split() function? (a) To ensure the model training process is randomized. (b) To control the number of iterations for the split. (c) To provide reproducibility of the data splitting. (d) To randomly select features for training.

27. 21. Which of the following is NOT typically considered a "key concept" or step in the standard scikit-learn model building workflow? (a) Instantiating an estimator object. (b) Calling the fit() method on training data. (c)

Defining a custom neural network architecture from scratch. (d) Making predictions using the predict() method.

28. 1. What is the primary purpose of Machine Learning? a) To write code that perfectly solves all problems without data. b) To enable systems to learn from data to identify patterns and make predictions. c) To manage large databases efficiently. d) To develop operating systems.

29. 17. Which method is commonly used to train a machine learning model in scikit-learn? (a) predict() (b) transform() (c) fit() (d) score()

30. 24. What is the main function of the predict() method in a trained scikit-learn model? (a) To learn patterns from the input data. (b) To apply data scaling or normalization. (c) To generate new labels or values for unseen data. (d) To calculate the model's error rate on a test set.

31. 34. Which argument of the pandas read_csv() function is used to specify the character that separates values in the file, if it's not a comma? (a) delimiter (b) separator (c) sep (d) splitter

32. 12. The "estimator API" in Scikit-learn typically involves which three core methods for a model? a) load(), save(), plot() b) install(), configure(), run() c) fit(), predict(), transform() d) import(), export(), modify()

33. 38. Which parameter in the train_test_split() function is used to define the proportion of the dataset that should be allocated to the testing set? (a) train_size (b) test_size (c) split_ratio (d) validation_split

34. 6. Which of the following is a key feature of the Scikit-learn library? a) It specializes in deep neural networks. b) It provides a consistent API for various ML algorithms. c) It is primarily used for data visualization. d) It is a low-level assembly language wrapper.

35. 9. Before applying a Machine Learning algorithm from Scikit-learn, what is a common initial step for tabular data using Pandas? a) Model instantiation b) Data visualization c) Loading the dataset, often using read_csv() d) Algorithm selection

36. 23. Which of the following scikit-learn classes is an example of an unsupervised learning estimator? (a) LinearRegression (b) DecisionTreeClassifier (c) KMeans (d) SVC (Support Vector Classifier)

37. 40. When dealing with classification tasks on imbalanced datasets, which parameter in train_test_split() is crucial for ensuring that the proportion of classes is maintained in both the training and testing sets? (a) shuffle (b) stratify (c) equalize_classes (d) balance_labels

38. 30. Which scikit-learn module is most likely to contain classes for data preprocessing steps like scaling and imputation? (a) sklearn.ensemble (b) sklearn.datasets (c) sklearn.preprocessing (d) sklearn.metrics

39. 45. Which of the following steps is typically performed after loading a dataset with read_csv() but before calling train_test_split() in a supervised learning task? (a) Instantiating a machine learning model (e.g., LogisticRegression()). (b) Fitting the model to the training data (model.fit()). (c) Separating the features (X) from the target variable (y). (d) Making predictions on the test set (model.predict()).

40. 13. What is the purpose of the 'predict()' method in a Scikit-learn model? a) To train the model with new data. b) To evaluate the model's accuracy on the training data. c) To make predictions on new, unseen input data. d) To preprocess the input features.

41. 14. When loading a dataset in Python for use with Scikit-learn, which library's read_csv() function is most commonly employed for CSV files? a) NumPy b) Matplotlib c) Pandas d) SciPy

42. 11. Which module within Scikit-learn is commonly used for splitting datasets and cross-validation techniques? a) sklearn.datasets b) sklearn.preprocessing c) sklearn.model_selection d) sklearn.metrics

43. 44. What could be a consequence if the random_state parameter is NOT set when calling train_test_split()? (a) The function will raise an error. (b) The splitting process will always produce identical train/test sets. (c) The splitting will be randomized each time the code is executed, leading to different evaluation results. (d) The training set will be disproportionately larger than the test set.

44. 29. What is the purpose of the X and y parameters passed to the fit() method of a supervised learning estimator in scikit-learn? (a) X represents the target variable, and y represents the features. (b) X represents the features (input data), and y represents the target variable (labels). (c) Both X and y represent the input features for training. (d) Both X and y represent the output labels for evaluation.

45. 42. In a typical supervised machine learning workflow, what does 'X' commonly represent after loading and initial data preparation? (a) The target variable or dependent variable. (b) The feature matrix or independent variables. (c) The predicted output of the model. (d) The model's parameters.

ANSWER KEY

1. (c)
2. (c)
3. (a)
4. (c)
5. (c)
6. (c)
7. (c)
8. (c)
9. (c)
10. (c)
11. (b)
12. (c)
13. (b)
14. (b)
15. (c)
16. (b)
17. (b)
18. (c)
19. (b)
20. (c)
21. (c)
22. (d)
23. (b)
24. (c)
25. (b)
26. (c)
27. (c)

28. (b)

29. (c)

30. (c)

31. (c)

32. (c)

33. (b)

34. (b)

35. (c)

36. (c)

37. (b)

38. (c)

39. (c)

40. (c)

41. (c)

42. (c)

43. (c)

44. (b)

45. (b)