DATASCIENCE

TEST-2

1) What is the purpose of the `print` function in Python? **(Python-Basic)**

A) To take user input.

B) To format text output.

C) To define a function.

D) To import modules.

Correct Answer: B) To format text output.

Explanation: The `print` function in Python is used to display text or variable values in the console.

2) Explain the difference between Python 2 and Python 3**. (Python-Medium)**

A) Python 2 focuses on simplicity, while Python 3 emphasizes compatibility and future-oriented features.

B) Python 3 is the latest version with improvements and new features, while Python 2 is deprecated.

C) Python 2 and Python 3 are interchangeable and have no significant differences.

D) Python 2 is for web development, while Python 3 is for data science.

Correct Answer: B) Python 3 is the latest version with improvements and new features, while Python 2 is deprecated.

Explanation: Python 3 is the latest version of the language, introduced to fix design flaws and add new features. Python 2 has reached its end-of-life and is no longer supported.

3) Explain the concept of a generator in Python, and provide an example of how it can be used to efficiently handle large datasets or sequences of data. **(Python - Hard)**

A) A generator is a function that creates a new instance of an object.

B) A generator is a class in Python used for generating random numbers.

C) A generator is a special type of iterable in Python that allows you to iterate over a potentially infinite sequence of values without creating the entire sequence in memory.

D) A generator is a decorator used to enhance the performance of functions.

Correct Answer: C) A generator is a special type of iterable in Python that allows you to iterate over a potentially infinite sequence of values without creating the entire sequence in memory.

Explanation: In Python, a generator is a special type of iterable that allows you to create an iterator for a potentially infinite sequence of values. It generates values on-the-fly, and unlike lists, it doesn't store the entire sequence in memory. This is particularly useful when dealing with large datasets or infinite sequences. Here's an example of a generator that generates Fibonacci numbers:

```
# Example of a Generator in Python

def fibonacci_generator():
    a, b = 0, 1
    while True:
        yield a
        a, b = b, a + b


# Create a Fibonacci generator
fibonacci_gen = fibonacci_generator()


# Generate the first 5 Fibonacci numbers
for _ in range(5):
    print(next(fibonacci_gen))
```

In this example, the `fibonacci_generator` function is a generator that yields Fibonacci numbers one at a time. The generator can be used in a loop to efficiently generate as many Fibonacci numbers as needed without storing the entire sequence in memory.

4) What is feature engineering in the context of machine learning? **(Feature engineering-Basic)**

A) Engineering physical features for a machine.

B) Enhancing model performance by creating new relevant input features.

C) Designing user interfaces for machine learning applications.

D) Optimizing machine learning algorithms.

Correct Answer: B) Enhancing model performance by creating new relevant input features.

Explanation: Feature engineering involves creating new input features or modifying existing ones to improve the performance of machine learning models.

5) Explain one technique for handling missing data in feature engineering.

(Feature engineering-Medium)

A) Dropping the rows with missing data.

B) Replacing missing values with the mean of the column.

C) Ignoring missing data.

D) Creating a new feature for missing values.

Correct Answer: B) Replacing missing values with the mean of the column.

Explanation: One technique for handling missing data is to replace the missing values with the mean (or median) of the column.

6) What is the purpose of one-hot encoding in feature engineering? **(Feature engineering-Medium)**

A) Reducing the dimensionality of the data.

B) Handling missing values.

C) Transforming categorical variables into binary vectors.

D) Scaling numerical features.

Correct Answer: C) Transforming categorical variables into binary vectors.

Explanation: One-hot encoding is used to represent categorical variables as binary vectors, with each category represented by a binary column.

7) Explain the concept of polynomial features in feature engineering**. (Feature engineering- Hard)**

A) Polynomial features involve creating new features by combining existing features using mathematical operations.

B) Polynomial features are irrelevant in feature engineering.

C) Polynomial features are only applicable to linear regression models.

D) Polynomial features are features with a degree of 2 or higher, used to capture non-linear relationships.

Correct Answer: A) Polynomial features involve creating new features by combining existing features using mathematical operations.

Explanation: Polynomial features involve creating new features by raising existing features to a power or combining them using mathematical operations. This can capture non-linear relationships in the data.

8) Explain the concept of feature scaling in feature engineering**. (Feature engineering- Hard)**

A) Feature scaling involves creating new features by combining existing ones.

B) Feature scaling is the process of transforming numerical features to a standardized range.

C) Feature scaling is only relevant for categorical features.

D) Feature scaling is used to add noise to the dataset.

Correct Answer: B) Feature scaling is the process of transforming numerical features to a standardized range.

Explanation: Feature scaling is a critical step in feature engineering that involves transforming numerical features to a standardized range, ensuring that they have a similar scale. This is particularly important for machine learning algorithms that are sensitive to the scale of features, such as distance-based algorithms.

9) What is the purpose of data cleaning in the context of data analysis? **(Data cleaning-basic)**

A) To delete irrelevant columns.

B) To remove duplicate records.

C) To replace missing values.

D) All of the above.

Correct Answer: D) All of the above.

Explanation: Data cleaning involves various tasks, including removing irrelevant columns, handling missing values, and eliminating duplicate records.

10) Explain one method for detecting outliers in a dataset during data cleaning. **(Data cleaning-Medium)**

A) Ignoring outliers.

B) Z-score method.

C) Deleting the entire dataset.

D) Averaging outlier values.

Correct Answer: B) Z-score method.

Explanation: The Z-score method involves calculating the standard score for each data point to identify values that deviate significantly from the mean.

11) import pandas as pd

```
# Read the dataset
df = pd.read_csv('data.csv')

# Remove duplicates
df = df.drop_duplicates()

# Replace missing values with the mean
df = df.fillna(df.mean())
```

What does this code do? **(Data cleaning-Hard)**

A) Drops duplicate records and replaces missing values with the mean for all columns.

B) Drops duplicate records and replaces missing values with the mean for specific columns.

C) Drops missing values and replaces duplicate records with the mean.

D) Drops missing values and duplicates without any replacements.

Correct Answer: A) Drops duplicate records and replaces missing values with the mean for all columns.

Explanation:The code uses `drop_duplicates` to remove duplicate records and `fillna` with the mean to replace missing values for all columns.

12) What is Seaborn used for in Python data visualization? **(Seaborn- Basic)**

A) For mathematical operations.

B) For creating machine learning models.

C) For statistical data visualization.

D) For handling missing data.

Correct Answer: C) For statistical data visualization.

Explanation: Seaborn is a Python data visualization library based on Matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

13) How does Seaborn differ from Matplotlib in terms of usage? **(Seaborn-Medium)**

A) Seaborn is a subset of Matplotlib.

B) Seaborn is built on top of Matplotlib and provides a higher-level interface for statistical data visualization.

C) Seaborn is designed for 3D plotting.

D) Seaborn is a replacement for Matplotlib.

Correct Answer: B) Seaborn is built on top of Matplotlib and provides a higher-level interface for statistical data visualization.

Explanation: Seaborn is built on top of Matplotlib and provides a higher-level interface for creating statistical graphics.

14) import seaborn as sns

import matplotlib.pyplot as plt

Load the 'tips' dataset

tips = sns.load_dataset('tips')

```
# Create a scatter plot
sns.scatterplot(x='total_bill', y='tip', data=tips, hue='sex', style='time')

# Show the plot
plt.show()
```

What does this code do? **(Seaborn- Hard)**

A) Creates a bar plot for the 'tips' dataset.

B) Creates a scatter plot of 'total_bill' vs 'tip' with points colored by 'sex' and styled by 'time'.

C) Creates a line plot for the 'tips' dataset.

D) Creates a box plot for 'total_bill' and 'tip'.

Correct Answer: B) Creates a scatter plot of 'total_bill' vs 'tip' with points colored by 'sex' and styled by 'time'.

Explanation: The code uses Seaborn to create a scatter plot where 'total_bill' is on the x-axis, 'tip' is on the y-axis, points are colored by 'sex', and styled by 'time'.

15) What is the primary purpose of the Pandas library in Python? **(Pandas-Basic)**

A) To create visualizations.

B) To perform data manipulation and analysis.

C) To generate machine learning models.

D) To handle HTTP requests.

Correct Answer: B) To perform data manipulation and analysis.

Explanation: Pandas is a powerful Python library for data manipulation and analysis. It provides data structures like DataFrame for efficient data manipulation.

16) How can you select specific columns from a Pandas DataFrame? **(Pandas-Medium)**

A) Using the `iloc` method.

B) Using the `loc` method.

C) Using square brackets with column names.

D) Both B and C.

Correct Answer: D) Both B and C.

Explanation: You can select specific columns from a Pandas DataFrame using the `loc` method (label-based) or square brackets with column names.

17) import pandas as pd

```
# Create a DataFrame
data = {'Name': ['Alice', 'Bob', 'Charlie'],
        'Age': [25, 30, 22],
        'City': ['New York', 'San Francisco', 'Los Angeles']}

df = pd.DataFrame(data)

# Sort the DataFrame by Age in descending order
df_sorted = df.sort_values(by='Age', ascending=False)

# Display the sorted DataFrame
```

```
print(df_sorted)
```

What is the output of this code? **(Pandas- Hard)**

A)

```
    Name  Age          City
1    Bob   30  San Francisco
0   Alice   25       New York
2  Charlie   22    Los Angeles
```

B)

```
    Name  Age          City
1    Bob   30  San Francisco
0   Alice   25       New York
2  Charlie   22    Los Angeles
```

C)

```
    Name  Age        City
0   Alice   25     New York
1    Bob   30        SF
2  Charlie   22  Los Angeles
```

D)

```
    Name  Age          City
2  Charlie   22    Los Angeles
0   Alice   25       New York
1     Bob   30  San Francisco
```

Correct Answer: A)

```
    Name  Age          City
1     Bob   30  San Francisco
0   Alice   25       New York
2  Charlie   22    Los Angeles
```

Explanation:The code creates a DataFrame, sorts it by the 'Age' column in descending order, and then prints the sorted DataFrame.

18) What is Matplotlib used for in Python? **(Matplotlib-Basic)**

A) To perform statistical analysis.

B) To create machine learning models.

C) To handle HTTP requests.

D) To create static, animated, and interactive visualizations in Python.

Correct Answer: D) To create static, animated, and interactive visualizations in Python.

Explanation: Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.

19) How can you create a scatter plot using Matplotlib? **(Matplotlib-Medium)**

A) Using the `plot` function.

B) Using the `scatter` function.

C) Using the `bar` function.

D) Using the `hist` function.

Correct Answer: B) Using the `scatter` function.

Explanation: The `scatter` function in Matplotlib is used to create scatter plots.

20) import matplotlib.pyplot as plt

import numpy as np

Generate data

x = np.linspace(0, 2 * np.pi, 100)

y1 = np.sin(x)

y2 = np.cos(x)

Create a plot with two lines

plt.plot(x, y1, label='sin(x)')

plt.plot(x, y2, label='cos(x)')

Add labels and legend

plt.xlabel('x')

plt.ylabel('y')

```
plt.legend()

# Show the plot
plt.show()
```

What does this code do? **(Matplotlib-Hard)**

A) Creates a bar plot for sine and cosine functions.

B) Creates a scatter plot for sine and cosine functions.

C) Creates a line plot for sine and cosine functions with labels and a legend.

D) Creates a histogram for sine and cosine functions.

Correct Answer: C) Creates a line plot for sine and cosine functions with labels and a legend.

Explanation: The code uses Matplotlib to create a line plot for the sine and cosine functions with labeled axes and a legend.

21) What is the primary purpose of the Scikit-Learn library in Python? **(Scikit-Learn- Basic)**

A) To perform data cleaning.

B) To create visualizations.

C) To implement machine learning algorithms.

D) To handle HTTP requests.

Correct Answer: C) To implement machine learning algorithms.

Explanation: Scikit-Learn is a machine learning library in Python used for implementing various machine learning algorithms.

22) How can you split a dataset into training and testing sets using Scikit-Learn? **(Scikit-Learn-Medium)**

A) Using the `train_test_split` function.

B) Using the `split_dataset` method.

C) Using the `divide_data` function.

D) Using the `create_train_test` method.

Correct Answer: A) Using the `train_test_split` function.

Explanation:The `train_test_split` function in Scikit-Learn is commonly used to split a dataset into training and testing sets.

23) What is the purpose of the `fit` method in Scikit-Learn? **(Scikit-Learn-Hard)**

A) To evaluate the model's performance.

B) To make predictions on new data.

C) To train the machine learning model on the training data.

D) To split the dataset into training and testing sets.

Correct Answer: C) To train the machine learning model on the training data.

Explanation: The `fit` method in Scikit-Learn is used to train the machine learning model on the training data.

24) from sklearn.linear_model import LinearRegression

from sklearn.metrics import mean_squared_error

import numpy as np

Generate random data

np.random.seed(42)

X = 2 * np.random.rand(100, 1)

y = 4 + 3 * X + np.random.randn(100, 1)

```
# Create a Linear Regression model
model = LinearRegression()

# Train the model
model.fit(X, y)

# Make predictions on the training set
y_pred = model.predict(X)

# Calculate the mean squared error
mse = mean_squared_error(y, y_pred)
print('Mean Squared Error:', mse)
```

What does this code do? **(Scikit-Learn -Hard)**

A) Creates a support vector machine model.

B) Creates a decision tree model.

C) Creates a linear regression model, trains it on random data, makes predictions, and calculates the mean squared error.

D) Creates a k-nearest neighbours model.

Correct Answer: C) Creates a linear regression model, trains it on random data, makes predictions, and calculates the mean squared error.

Explanation: The code uses Scikit-Learn to create a linear regression model, train it on randomly generated data, make predictions, and calculate the mean squared error.

25) Describe the purpose of the `train_test_split` function in Scikit-Learn, and explain why splitting a dataset into training and testing sets is essential in machine learning. **(Scikit-Learn -Medium)**

A) `train_test_split` is used to train a machine learning model and has no impact on testing.

B) `train_test_split` is used to split the dataset into training and testing sets, allowing the model to be trained on one subset and tested on another. This helps assess the model's generalization to new, unseen data.

C) `train_test_split` is only relevant for classification problems.

D) `train_test_split` is used for feature selection in machine learning.

Correct Answer: B) `train_test_split` is used to split the dataset into training and testing sets, allowing the model to be trained on one subset and tested on another. This helps assess the model's generalization to new, unseen data.

Explanation: The `train_test_split` function in Scikit-Learn is commonly used to divide a dataset into training and testing sets. The training set is used to train the machine learning model, and the testing set is used to evaluate its performance on unseen data. This separation is crucial for assessing how well the model generalizes to new, previously unseen data, helping to identify potential overfitting or underfitting issues.