**Read the following instructions carefully before attempting the questions.**

* There are 4 sections in this question paper. All Questions are compulsory
* Only write the answer to each question with the question number clearly marked.
* All questions must be solved using Jupyter Notebook or Google Colab.
* Evaluation criteria are mentioned wherever required.

Time Allowed: 2.5 hours

MCQ Assessment: Python Fundamentals, NumPy, Pandas, and Matplotlib

1.Which of the following statements about Python is/are true?

a) Python is a high-level programming language.

b) Python supports multiple programming paradigms.

c) Python has a built-in garbage collector.

d) Python is a statically typed language.

2.Given the NumPy array 'arr', what will be the output of the following code?

import numpy as np

arr = np.array([1, 2, 3, 4, 5])

new\_arr = arr[arr > 2]

print(new\_arr)

a) [1, 2]

b) [3, 4, 5]

c) [1, 2, 3, 4, 5]

d) [2, 3, 4, 5]

3.Which of the following is NOT a valid way to create a DataFrame in Pandas?

a) Using a NumPy array

b) Reading data from a CSV file

c) Converting a dictionary to a DataFrame

d) Using a Python list

4.What does the 'axis' parameter in NumPy do?

a) It specifies the dimension along which an operation is performed.

b) It specifies the size of the array.

c) It indicates the number of dimensions in the array.

d) It determines the data type of the array.

5.Which of the following statements about Pandas is/are true?

a) Pandas is built on top of NumPy.

b) Pandas provides data structures like Series and DataFrame.

c) Pandas is primarily used for numerical computations.

d) Pandas is not compatible with Python 3.

6.Which of the following is used to visualize data using Matplotlib?

a) Bar plots

b) Line plots

c) Scatter plots

d) All of the above

7.What is the correct syntax to save a Matplotlib figure as an image file?

a) save\_figure('output.png')

b) plt.save\_figure('output.png')

c) plt.savefig('output.png')

d) save\_plot('output.png')

8.How can you remove missing values (NaN) from a DataFrame in Pandas?

a) df.dropna()

b) df.fillna(0)

c) df.replace(np.nan, 0)

d) All of the above

9.In NumPy, what does the 'dtype' parameter specify?

a) The shape of the array.

b) The data type of the array.

c) The dimensions of the array.

d) The values in the array.

10.Which of the following statements about indexing and slicing in NumPy is/are true?

a) Slicing creates a view of the original array.

b) Indexing and slicing both create copies of the original array.

c) Slicing can be used to modify the original array.

d) Indexing and slicing are not supported in NumPy

11.Which of the following statements about Python's built-in 'range()' function is/are true?

a) 'range()' returns a list of numbers.

b) 'range()' generates numbers starting from 0 by default.

c) 'range()' can take a step parameter to specify the increment.

d) 'range()' can only be used with integers.

12.What will be the output of the following Pandas code?

import pandas as pd

data = {'Name': ['John', 'Sarah', 'Michael'],

'Age': [25, 30, 35],

'City': ['New York', 'London', 'Paris']}

df = pd.DataFrame(data)

filtered\_df = df[df['Age'] > 30]

print(filtered\_df['Name'])

a) John

b) Sarah

c) Michael

d) John, Sarah, Michael

13.Which of the following operations can be performed using NumPy's 'random' module?

a) Generating random integers

b) Generating random arrays

c) Shuffling an array

d) Calculating mean and standard deviation

14.Which of the following statements about Pandas DataFrame indexing is/are true?

a) Indexing can be done using both labels and integer-based positions.

b) DataFrame columns can be used as an index.

c) Indexing can be performed using boolean conditions.

d) Indexing always creates a copy of the original DataFrame.

15.What does the 'plt.xlabel()' function in Matplotlib do?

a) It sets the x-axis label of the plot.

b) It sets the title of the plot.

c) It creates a legend for the plot.

d) It saves the plot as an image file.

16.How can you rename a column in a Pandas DataFrame?

a) df.rename\_column('old\_column', 'new\_column')

b) df.column\_rename('old\_column', 'new\_column')

c) df.rename(columns={'old\_column': 'new\_column'})

d) df.rename('old\_column', 'new\_column')

17.Which of the following methods can be used to read data from a CSV file into a Pandas DataFrame?

a) pd.read\_csv()

b) pd.read\_excel()

c) pd.from\_csv()

d) pd.load\_csv()

18.What is the purpose of the 'plt.legend()' function in Matplotlib?

a) It adds a legend to the plot.

b) It sets the title of the plot.

c) It saves the plot as an image file.

d) It adjusts the spacing between subplots.

19.In NumPy, how can you find the maximum value in a given array?

a) np.max(arr)

b) np.maximum(arr)

c) np.argmax(arr)

d) np.amax(arr)

20. Which of the following statements about Pandas Series is/are true?

a) Series can only store numeric data.

b) Series can have a custom index.

c) Series is a two-dimensional data structure.

d) Series does not support vectorized operations.

21.What is the difference between mean, median, and mode? How would you decide which one to use in a given situation?

a) Mean represents the most frequently occurring value, while median represents the average value.

b) Mean represents the average value, while median represents the middle value.

c) Mean represents the middle value, while median represents the average value.

d) Mode represents the average value, while median represents the most frequently occurring value.

22.What is the Central Limit Theorem? Explain its significance in statistical analysis.

a) It states that all data points in a sample are normally distributed.

b) It states that the mean of a large sample tends to follow a normal distribution.

c) It states that the variance of a large sample tends to follow a normal distribution.

d) It states that the sample size has no impact on the distribution of the data.

23.What is correlation? How is it different from causation?

a) Correlation measures the strength of a relationship between variables, while causation implies a cause-and-effect relationship.

b) Correlation and causation are the same concepts and can be used interchangeably.

c) Correlation implies a cause-and-effect relationship, while causation measures the strength of a relationship.

d) Correlation and causation have no relationship in statistical analysis.

24.How would you handle missing values in a dataset? Explain different approaches you can take.

a) Ignore the missing values and proceed with the analysis.

b) Replace missing values with the mean of the variable.

c) Drop the rows with missing values from the dataset.

d) All of the above approaches can be used depending on the context.

25.What is the p-value in hypothesis testing? How do you interpret it?

a) The p-value represents the probability of making a Type II error.

b) The p-value represents the probability of rejecting the null hypothesis when it is true.

c) The p-value represents the probability of accepting the null hypothesis when it is false.

d) The p-value represents the probability of obtaining the observed result or more extreme, assuming the null hypothesis is true.

26.What is the difference between Type I and Type II errors in hypothesis testing?

a) Type I error occurs when the null hypothesis is incorrectly rejected, while Type II error occurs when the null hypothesis is incorrectly accepted.

b) Type I error occurs when the null hypothesis is correctly rejected, while Type II error occurs when the null hypothesis is incorrectly rejected.

c) Type I error occurs when the null hypothesis is incorrectly accepted, while Type II error occurs when the null hypothesis is correctly rejected.

d) Type I error occurs when the null hypothesis is incorrectly rejected, while Type II error occurs when the null hypothesis is correctly accepted.

27.How would you detect outliers in a dataset? What techniques or statistical measures would you use?

a) Outliers can be identified by calculating the z-scores and considering values above a certain threshold.

b) Outliers are identified based on the range of values within the dataset.

c) Outliers are detected by comparing the values to the mean of the dataset.

d) Outliers cannot be detected using statistical techniques.

28.Explain the concept of sampling bias. How can you minimize or mitigate its effects in data analysis?

a) Sampling bias occurs when the sample is not representative of the population, leading to inaccurate results.

b) Sampling bias occurs when the sample size is too small to draw meaningful conclusions.

c) Sampling bias occurs when the sample is randomly selected, ensuring accurate results.

d) Sampling bias cannot be minimized or mitigated in data analysis.

29.What is A/B testing? Describe the steps involved in conducting an A/B test.

a) A/B testing is a technique to compare two different datasets and determine statistical significance.

b) A/B testing is a technique to compare the performance of two different versions of a webpage or app.

c) A/B testing is a technique to analyze categorical data using statistical models.

d) A/B testing is a technique to measure the accuracy of machine learning models.

30.How would you assess the statistical significance of a result? What statistical tests or methods can be used for this purpose?

a) Statistical significance is assessed by comparing the mean and median values of the dataset.

b) Statistical significance is determined by calculating the range of values within the dataset.

c) Statistical significance is assessed using hypothesis testing, such as t-tests or chi-square tests.

d) Statistical significance is determined by visualizing the data using graphs or plots.

**Scenario Based Questions**

1. Scenario: You are given a dataset with missing values in multiple columns. How would you handle these missing values in order to prepare the data for analysis and modeling? What techniques or methods would you consider?
2. Scenario: You are tasked with building a churn prediction model for a telecommunications company. Which machine learning algorithms would you consider for this task, and why? How would you evaluate the performance of the model?
3. Scenario: You have been given a dataset containing customer information for a retail company. The task is to segment the customers into different groups based on their purchasing behavior. Describe the approach you would take to perform customer segmentation and explain the potential benefits of this analysis for the company.
4. Scenario: You have been provided with a large dataset that contains both numerical and categorical variables. The goal is to predict a binary outcome based on these variables. How would you handle the categorical variables during the modeling process? Explain the different encoding techniques you would consider.

**Concept Based Questions:**

1. Question: What is multicollinearity, and how does it impact the regression model? Explain the potential challenges it presents and discuss the techniques used to handle multicollinearity in regression analysis.
2. Question: In regression modeling, how do you address outliers in the dataset? Describe the impact of outliers on the regression model and discuss the different techniques to detect and handle outliers effectively.
3. Question: Overfitting is a common challenge in regression modeling. Explain what overfitting is and why it is problematic. Discuss regularization techniques, such as Ridge and Lasso regression, and how they can be used to address the issue of overfitting.
4. Question: When dealing with non-linear relationships between predictors and the target variable, how can regression models be enhanced? Describe the concept of polynomial regression and its role in capturing non-linear patterns. Explain how to select an appropriate degree for the polynomial regression.

**Application Question**

**Problem Statement #1:**

You have been provided with a dataset containing information about the COVID-19 pandemic in various countries. Your task is to perform exploratory data analysis, feature engineering, model selection and tuning to create a model that predicts the number of COVID-19 cases in a given country.

*Data Description:*

The dataset contains the following columns:

Country\_Region: The name of the country or region

Date: The date on which the data was collected

ConfirmedCases: The total number of confirmed COVID-19 cases in the country or region on the given date

Fatalities: The total number of deaths due to COVID-19 in the country or region on the given date

You can download the dataset from this link: <https://www.kaggle.com/c/covid19-global-forecasting-week-5/data>

Task 1: Exploratory Data Analysis

Perform exploratory data analysis on the dataset and answer the following questions:

1. Which country or region has the highest number of confirmed COVID-19 cases?
2. Which country or region has the highest number of deaths due to COVID-19?
3. Which country or region has the highest number of recovered COVID-19 cases?

Task 2: Feature Engineering

Create new features from the existing data that might be useful for predicting the number of COVID-19 cases. You can use any techniques of your choice such as time series decomposition, moving averages, lag features, etc.

Task 3: Model Selection and Tuning

Select an appropriate machine learning algorithm for the problem and train a model on the dataset. You can use any machine learning algorithm of your choice. Your model should take the following inputs:

* Country\_Region: The name of the country or region for which the prediction is to be made
* Date: The date for which the prediction is to be made
* The features you created in Task 2
* Your model should output the predicted number of COVID-19 cases and fatalities for the given country or region and date.

Evaluate your model using appropriate evaluation metrics and tune it for better performance.

Task 4: Deployment

Deploy your model to a web application using streamlit so that it can be used to make predictions on new data.

Evaluation:

Your submission will be evaluated on the following criteria:

Correctness and quality of the exploratory data analysis and feature engineering

Accuracy of the model predictions

Quality of the code and documentation

Clarity and completeness of the model deployment instructions