Register Number:…………….

M.TECH DEGREE EXAMINATIONS: NOV/DEC 2023

(Regulation 2018)

First Semester

DATA SCIENCE

P18ITI1204 : Data Science and Analytics with Python

COURSE OUTCOMES

CO1: Explain the roles and stages of data science projects.

CO2: Describe the data structures provided by numpy library for arrays and vectorized computation.

CO3: Explain data structures provided by pandas library for data analysis.

CO4: Perform data wrangling, cleaning and transformation using python.

CO5: Use matplot lib for plotting and visualizing the datasets.

CO6: Demonstrate data aggregation and time series analysis using python programming Language

Time: Three Hours **Maximum Marks: 100**

Answer all the Questions:-

PART A (10 x 1 = 10 Marks)

1. What is the primary purpose of NumPy in data science? CO2 [K_2]

a) Data visualization b) Machine learning algorithms

c) Text processing d) Database management

2. What is the purpose of the NumPy function np.random.randn()? CO2 [K_2]

a) Reading data from files b) Generating random integers

c) Generating an array of random numbers from a standard normal distribution d) Reshaping and pivoting data

3. In data science, what is the role of Matplotlib? CO5 [K_2]

a) Statistical analysis b) Linear algebra operations

c) Data visualization d) Database querying

4. In the context of data wrangling with pandas, what does the fillna() function do? CO4 [K_2]

a) Removes duplicate values from a DataFrame b) Fills missing values in a DataFrame with a specified value or method

c) Computes descriptive statistics of a DataFrame   d) Reshapes and pivots data

5. What is a primary use case for the GroupBy functionality in pandas? CO3 [K_1]

a) String manipulation   b) Data aggregation and transformations

c) Handling missing data   d) File input and output

6. Which library is commonly used for handling time series data in Python? CO6 [K_1]

a) NumPy   b) Matplotlib

c) pandas   d) SciPy

7. What does the acronym "API" stand for in the context of data science? CO1 [K_1]

a) Advanced Programming Interface   b) Application Programming Interface

c) Algorithmic Processing Interface   d) Automated Python Integration

8. What does the term "NaN" represent in pandas? CO3 [K_2]

a) Not a Name   b) No Available Number

c) Null or missing value   d) Numeric Array Notation

9. What does the term "DataFrame" refer to in the context of pandas? CO3 [K_2]

a) A linear algebra structure   b) A two-dimensional tabular data structure

c) A random number generator   d) A data cleaning function

10. In the context of NumPy, what is a universal function (ufunc)? CO2 [K_2]

a) A function that works on arrays element-wise   b) A function for file input and output

c) A function for random number generation   d) A function for data cleaning

PART B (10 x 2 = 20 Marks)

11. What is data science? CO1 [K_1]

12. What is a NumPy ndarray? CO2 [K_2]

13. Mention the usage of Pandas. CO3 [K_2]

14. Define data wrangling? CO4 [K_2]

15. What does ETL stand for in data science? CO1 [K_2]

16. What is a pivot table in pandas? CO3 [K_1]

17. Mention the use of Matplotlib. CO5 [K_2]

18. What are time series in data science? CO6 [K_1]

19. How does Pandas handle missing data? CO3 [K_2]

20. What is the significance of Random Walks in data science? CO1 [K_2]

PART C (6 x 5 = 30 Marks)

21. In a particular pain clinic, 10% of patients are prescribed narcotic pain killers. Overall, five percent of the clinic's patients are addicted to narcotics (including pain killers and illegal substances). Out of all the people prescribed pain pills, 8% are addicts. 5 CO4 [K_3]

    If a patient is an addict, what is the probability that they will be prescribed pain pills?

22. Assume that the test scores of a college entrance exam fits a normal distribution. Furthermore, the mean test score is 72, and the standard deviation is 15.2. What is the percentage of students scoring 84 or more in the exam? 5 CO4 [K_3]

23. Consider a set of patients coming for treatment in a certain clinic. Let A denote the event that a "Patient has liver disease" and B the event that a "Patient is an alcoholic." It is known from experience that 10% of the patients entering the clinic have liver disease and 5% of the patients are alcoholics. Also, among those patients diagnosed with liver disease, 7% are alcoholics. Given that a patient is alcoholic, what is the probability that he will have liver disease? 5 CO3 [K_3]

24. a. Explain the purpose of the 'rolling' function in time series analysis. 5 CO6 [K_3]

    b. What is the difference between upsampling and downsampling in time series analysis?

25. a. Explain the concept of broadcasting in NumPy. 5 CO2 [K_3]

    b. Explain the concept of vectorized operations in NumPy.

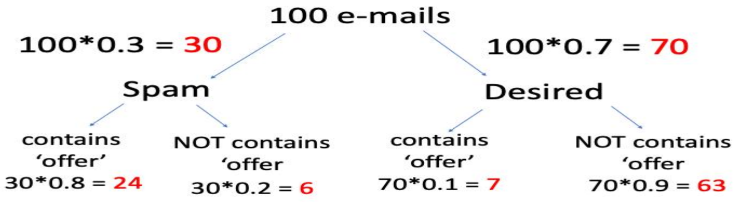
26. a. Explain the concept of merging in Pandas. 5 CO3 [K_3]

    b. Explain the concept of index in a Pandas DataFrame?

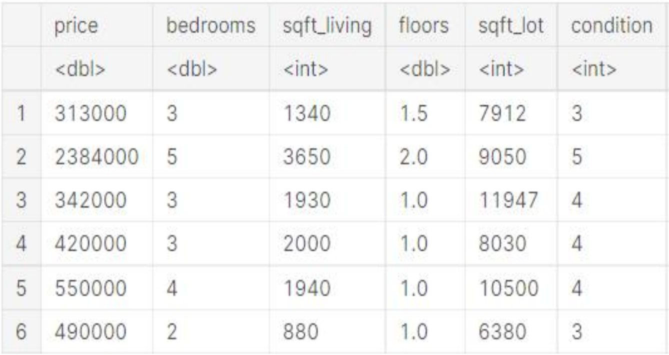
Answer any FOUR Questions

PART D (4 x 10 = 40 Marks)

27. Let's work on a simple NLP problem with Bayes Theorem. By using NLP, I can detect spam e-mails in my inbox. Assume that the word 'offer' occurs in 80% of the spam messages in my account. Also, let's assume 'offer' occurs in 10% of my desired e-mails. If 30% of the received e-mails are considered as a scam, and I will receive a new message which contains 'offer', what is the probability that it is spam? (Bayes rule) 10 CO3 [K_3]


100 e-mails
100*0.3 = 30
100*0.7 = 70
Spam
Desired
contains 'offer' 30*0.8 = 24
NOT contains 'offer 30*0.2 = 6
contains 'offer' 70*0.1 = 7
NOT contains 'offer 70*0.9 = 63

28. Import the house price data set. (Sample data set shown below). Crete a model and predict the house price based on the given features using Python. 10 CO5 [K_3]



	price	bedrooms	sqft_living	floors	sqft_lot	condition
	<dbl>	<dbl>	<int>	<dbl>	<int>	<int>
1	313000	3	1340	1.5	7912	3
2	2384000	5	3650	2.0	9050	5
3	342000	3	1930	1.0	11947	4
4	420000	3	2000	1.0	8030	4
5	550000	4	1940	1.0	10500	4
6	490000	2	880	1.0	6380	3

29. Import the average height and weight for American Women dataset of 15 observations. Using R, Create a model and predict the measure whether heights are positively correlated with weights and explain. 10 CO5 [K_3]

Dataset: women.csv

```
"","height","weight"
"1",58,115
"2",59,117
“3”,45,120
… … ….
"14",71,159
"15",72,164
```

30. Mention NumPy data types and its descriptions. Give example for all the data types. 10 CO2 [K_2]

31. Explain time series analysis using suitable example. 10 CO5 [K_2]

****