STACK ROUTE | AN NIIT VENTURE

WIPRO VELOCITY CONGNTIVE DATA SCIENCE PROGRAM

Capstone – Model Set

Total Marks – 100

CAPSTONE GUIDELINES

Note: This document is for class discussion only. This document is provided for learners in this Data Science Program and is not meant for circulation outside Wipro.

- The Capstone is for 100 marks which covers the entire syllabus.
- The Capstone Project involves different varieties of questions such as, Long/Short Answers, Descriptive/Practical Implementations, etc. to validate the candidate's understanding and learning of the curriculum covered.
- The descriptive questions need to be answered by the learners elaborately by taking some scenario or example as per the question.
- This test consists of the following subject categories and question pattern

S No	Module	Topics	Questions	Total Marks
1	Maths Foundation	Linear Algebra and Multivariate Calculus	1	5
2	Stats Foundation	Bayesian Probability, Probability Distribution, Hypothesis Testing	1	5
3	MySQL	View, Subquery, Stored Procedure, Functions	2	10
4	Python, NumPy, Pandas, Data Visualization	Loops, Functions, Arrays, Series. DataFrames, Merge, Group and Aggregate, Data Visualization	5	25
5	Data Preprocessing and ML Algorithms	Feature Engineering, Feature Selection, Scaling, Imputing, Encoding, ML Algorithms	1	15
6	Best Practices and Optimization in ML	Gradient Descent, Hyperparameter Tuning, Cross Validation, Performance Metrics	1	15
7	Neural Networks	Neural Networks Foundation, Keras, CNN, RNN, LSTM, GRU	4	25
Total			**15**	**100**

Note: The Model test paper, answers should be prepared in external files (.ipynb, .sql, .text etc) and should compress(zip) and upload into the LMS.

MATHS FOUNDATION

1. A Company has examined its cost structure and revenue structure and has determined that C, the total cost, R, the total revenue and x, the number of units produced are related as: C=100+0.015x^2and R=3x. Find the production level x that will maximize the profits of the company. Find that profit. Also find the profit when x = 120. **(5 marks)**

STATS FOUNDATION

2. Among 157 African-American men, the mean systolic blood pressure was 146 mm Hg with a standard deviation of 27. We wish to know if on the basis of these data, we may conclude that the mean systolic blood pressure for a population of African-American is greater than 140.
   - Setup the null and alternate hypothesis **(1 mark)**
   - Determine the type of the test **(1 mark)**
   - Use α=0.01, conduct the test and accept or reject the hypothesis on the basis of the test. (Given Z_0.99=2.33) **(3 marks)**

SQL QUERIES

Problem Statement

- Accounts payable is a critical business process through which all companies track and manage their payable obligations efficiently and effectively. Tracking of your debts and paying them on time is essential for any business. It also helps in maintaining good relations with the vendors and save money by making early payments.
- Being a Data Scientist, you are supposed to query data inside MySQL OLAP DB as per the business requirement of Account Payable department Sharing ER diagram for Account Payable for better insight.
- Note: Use the schema provided with the test

3. Write a statement that returns that returns vendor_id and the largest unpaid invoice for each vendor. Group the result set by the vendor_id. Write another statement that uses the first SELECT statement that uses the first SELECT statement in its FROM clause. The main query should return a single value that represents the sum of the largest unpaid invoices for each vendor. (5 Marks)
4. Write a statement that returns the name, city, and state of each vendor that's located in a unique city and state. In other words, don't include vendors that have a city and state in common with another vendor and Sort the results by the vendor_state and vendor_city columns. **(5 Marks)**

PYTHON, NUMPY, PANDAS, DATA VISUALIZATION

5. Define a Python function to find the minimum and maximum values from a list. **(5 Marks)**

6. Create a 3x3 NumPy array of ones and find out the total memory used for the array. **(5 Marks)**
7. Load the ODI_data.csv file using Pandas and print first 10 rows of the data. Find out the shape of the data. **(5 Marks)**
8. Visualize boxplot and scatterplot for the above mentioned data.
9. Find out the mean scores by Sachin, Rahul and other top 5 Indian Players. **(5 Marks).**
10. Describe the usage of Aggregate Functions such as groupby. **(5 Marks)**

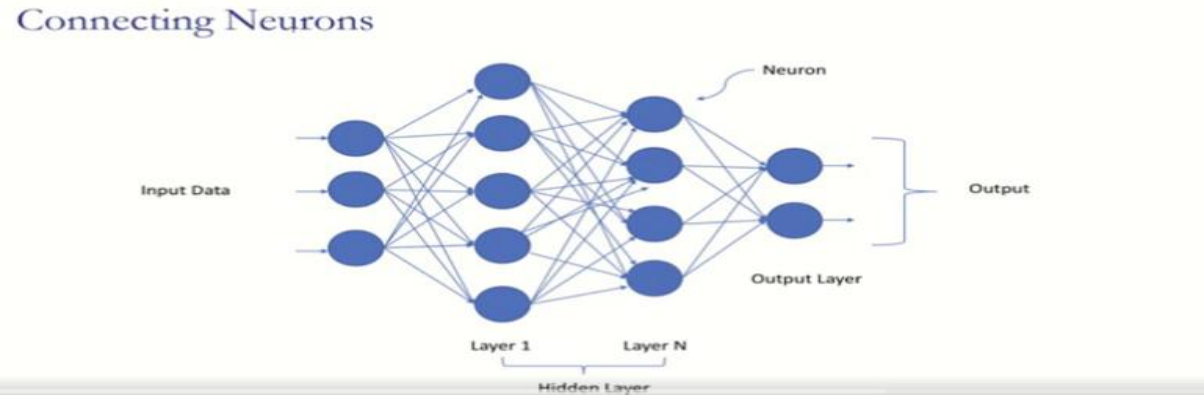
DATA PREPROCESSING AND ML ALGORITHMS

11. **Implement the following preprocessing steps (Data_Train_Cars.xlsx)**
    - Perform loading of the Data_Train_Cars.xlsx with pandas. Print the first and last 5 rows of the data. (5 Marks)
    - Perform complete profiling of the data by Pandas profiling. (5 Marks)
    - Find out if any missing values and impute it with median. **(5 Marks)**

BEST PRACTICES AND OPTIMIZATION IN ML

12. Consider the Customer_Purchases.csv dataset and perform the following operations.
    - Apply Decision Tree Classification technique. **(5 Marks)**
    - Use the criterion as Entropy and Find the Accuracy Score **(5 Marks)**
    - Use the Criterion as gini and find the accuracy Score **(5 Marks).**

NEURAL NETWORKS



13. In the above(Image) neural network compute the Number of parameters or Weights of this Network? . **(10 Marks)**
14. Describe the Importance of Activation Function, list out at least 3 activation functions with their Mathematical Models. **(5 marks)**
15. Describe the usage of Cost Functions with some examples? **(5 marks)**
16. Describe the importance of epochs, Learning Rate and Regularization in Neural Networks. **(5 marks)**