**NED UNIVERSITY OF ENGINEERING & TECHNOLOGY**

Centre for Multidisciplinary Postgraduate Programmes (CMPP) – NED Academy

Postgraduate Diploma in \_\_\_\_\_\_\_\_Data Science & AI\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Final Examination – Spring-I-2025

Course: **Machine Learning**

Time: 3 Hours Max. Marks: 60

Instructions:

1. Attempt all questions
2. Marks for each question are given.
3. You are required to abide by all rules and regulations set for the examination by the NED Academy.
4. Total time of examination including uploading is 3 hours. No extra time will be provided after the time is over.

|  |  |  |
| --- | --- | --- |
| **S. No.** | **Question** | **Marks** |
| 1. | **Objective:**  Group retail store customers based on their purchase history using the  • Dataset: Attached with paper  **Data Analysis:**  • Explored the first five rows of the dataset and examined its shape for initial insights  Data Cleaning:  • Identified and removed null/missing values  • Detected and eliminated duplicate entries  • Dropped unnecessary columns  • Identified and managed outliers to ensure data quality  **ML Model:**  • Implemented the K-Means Clustering Algorithm for customer grouping  • Calculated WCSS (Within-Cluster Sum of Squares) values for K values ranging  from 1 to 20.  • Applied the elbow method to determine the optimal K value by plotting WCSS  values against different K values  • Identified clusters and their corresponding data points.  Data Visualization:  • Created scatter plots to visualize clusters and their data points for clear insight | 12 |
| 2. | Titanic- the ML challenge- a model build with Scikit learn Pipelines  1. Do some simple Pre-Processing and EDA.  2. Build a simple and reusable machine learning work flow with Scikit learn’s  Pipeline Architecture  3. Pick the best model (a random forest / Decision Tree) and improve it further with hyper parameter tuning. | 12 |
| 3. | Data: Through the diagnosis test I predicted 100 report as COVID positive, but only 45 of those were actually positive. Total 50 people in my sample were actually COVID positive. I have total 500 samples.  Create confusion matrix based on given data mathematically and find  I. Accuracy  II. Precision  III. Recall  IV. F-1 score | 12 |
| 4. | Download heart dataset from following link.  <https://www.kaggle.com/datasets/zhaoyingzhu/heartcsv>  Perform following operation on given dataset.  a) Find Shape of Data  b) Find Missing Values  c) Find data type of each column  d) Finding out Zero's  e) Find Mean age of patients  f) Now extract only Age, Sex, ChestPain, RestBP, Chol. Randomly divide dataset in training (75%)  and testing (25%).  g- Create confusion matrix and classification report. | 12 |
| 5. | Discuss the following (Any 4)   1. Overfitting 2. Under fitting 3. Best fit 4. Bias 5. Variance | 12 |