


## National University of Computer and Emerging Sciences, Lahore Campus

	<b>Course:</b>	DM Lab	<b>Course Code:</b>	DL3002
	<b>Program:</b>	BS (Data Science)	<b>Semester:</b>	Spring 2024
	<b>Duration:</b>	90 Minutes	<b>Total Marks:</b>	40
	<b>Date:</b>	21-March-24	<b>Weight:</b>	20 %
	<b>Section:</b>	6-A	<b>Page(s):</b>	2
	<b>Exam:</b>	Mid	<b>Reg. No.</b>	

### Read below Instructions Carefully:

- Understanding the question statement is also part of the exam, so do not ask for any clarification. In case of any ambiguity, make suitable assumptions.
- You have to complete the exam in 1.5 hrs. No extra time will be given for submission.
- Submit a single **.ipynb file** for each question named as **21L-1122 (Q#)**
- Place all files into the **folder** named as **21L-1122**
  - Submit folder on **cactus** by following path: \\cactus1\xeon\Spring2024\
- Your code should be **intended** and **commented** properly. Use **meaningful variable names**.
- It is your responsibility to save your code from being copied. All matching codes will be considered cheating cases. **PLAGIARISM** will result in forwarding of **case to Disciplinary Committee** and **ZERO** in Midterm.

### Question 1:

20 marks

A major problem in bioinformatics analysis or medical science is in attaining the correct diagnosis of certain important information. For the ultimate diagnosis, normally, many tests generally involve the clustering or classification of large scale data. All of these test procedures are said to be necessary in order to reach the ultimate diagnosis. However, on the other hand, too many tests could complicate the main diagnosis process and lead to the difficulty in obtaining the end results, particularly in the case where many tests are performed. This kind of difficulty could be resolved with the aid of machine learning which could be used directly to obtain the end result with the aid of several artificial intelligent algorithms which perform the role as classifiers.

### Dataset Information

The data used in this investigation is the breast cancer data. For training and testing, only 75% of the overall data is used for training and the rest is used for testing the accuracy of the classification of the selected classification methods.

The ARFF file is provided to load the data into WEKA software for the classification of the Cancer.

- 1) Indicate total number of instances, the number of attributes and number of samples under each class of power quality problems along with a bar graph.
- 2) The data load into WEKA is used to train the data mining algorithms: J48, SVM and Random Forest for classification purpose. After training, the algorithms are tested based on the given training set and as well as using stratified 10-fold cross-validation.
  - a) Indicated the results obtained after testing the algorithms using training set.
  - b) Indicated results obtained after testing the algorithms using stratified 10-fold cross validation
- 3) Compare results of classifiers based on precision, recall, f1 score, and which take less time.
- 4) Talk about any possible challenges with this problem that could arise throughout the classification process.

## Question 2:

20 marks

You are provided with a dataset. Your task is to build a machine learning model to detect its Y variable using various Data Mining techniques within a 1.5-hour time frame.

### 1. Data Exploration and Visualization:

Load the dataset and explore its structure using Pandas.

Visualize key features to gain insights into the data.

### 2. Data Preprocessing:

Handle any missing values and outliers in the dataset.

Perform feature scaling and transformation if necessary.

### 3. Model Building and Evaluation:

Split the dataset into training and testing sets (e.g., 70% training, 30% testing).

Build and train a classification model using any 2 of the following algorithms:

- KNN
- NaiveBayes
- SVM

Evaluate the model's performance using metrics like accuracy, precision, recall, and F1- score on the test set.

Visualize the confusion matrix and ROC curve for model evaluation.

Which model classification accuracy is better ?