**FINAL LAB EXAM – S1 MTECH RAU**

**MACHINE LEARNING (10m)**

**Date: 10/01/2022 Time: 2 Hours**

Instructions:

1. For the 5th question, where certain theory/comment questions are asked (they are highlighted in bold), write all your answers as comments/text in **one line only** in the “.ipynb” file itself in Google Colab.
2. Write the answers for questions 1-4 in your Google Colab notebook itself like your mid-sem exam.
3. You have a total of 2 hours to attempt the exam and upload your “.ipynb” file. Other than this, no extra time will be given. Use your time carefully.

Questions:

**1.** When we perform Principal Component Analysis (PCA), our direction of interest is along which of the following? (0.5m)

A. Along the direction with the largest average

B. Along the direction with the largest variance

C. Along the direction with the smallest variance

D. Along the direction of one of the naive basis vectors

**2.** Having a labelled dataset is not a necessity when performing dimensionality reduction algorithms like PCA. (0.5m)

A. True

B. False

**3.** In PCA, PC-1 and PC-2 are --------- to each other. (0.5m)

**4.** In PCA, the maximum number of principal components is --------- the number of features in the dataset. (0.5m)

**5.** For this question, use the “**kidney\_disease.csv**” file which is attached in Teams. Do the following: (8m)

a) Convert the categorical columns in the dataset into numerical data.

b) Check for null values in your dataset. If there are null values, use any method to fill them/drop them. **Mention which method you have used and why**.

c) Perform PCA on your dataset with the number of components = 2. Find the explained variance ratio for your principal components. **Comment what is the total variance explained by your PCA**.

d) Plot your PCA graph and from the graph, see and **comment, for PC1 = -1 and PC2 = 0, does the patient have a kidney disease or not?**

e) Now, to retain 95% of the data (to explain 95% total variance), find out how many principal components are required in your dataset and **comment the answer**.

--------------------------------------------------The End----------------------------------------------------