**Walchand College of Engineering, Sangli**

## **Machine Learning Lab (6CS372)**

**TY BTech | AY 2023-2024 | Even Sem**

**Assignment 2**

**Batch 1 -**[**https://www.kaggle.com/datasets/spittman1248/cdc-data-nutrition-physical-activity-obesity**](https://www.kaggle.com/datasets/spittman1248/cdc-data-nutrition-physical-activity-obesityhttps:/www.kaggle.com/spittman1248/cdc-data-nutrition-physical-activity-obesity%20%20)

**Batch 2-** [**https://archive.ics.uci.edu/ml/datasets/Air+Quality**](https://archive.ics.uci.edu/ml/datasets/Air+Quality)

**Batch 3-** [**https://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction**](https://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction)

**Batch 4-** [**https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset**](https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset)

**Batch 5-** [**https://archive.ics.uci.edu/ml/datasets/Demand+Forecasting+for+a+store**](https://archive.ics.uci.edu/ml/datasets/Demand+Forecasting+for+a+store)

**Batch 6-** [**https://archive.ics.uci.edu/ml/datasets/Hungarian+Chickenpox+Cases**](https://archive.ics.uci.edu/ml/datasets/Hungarian+Chickenpox+Cases)

**Batch 7-** [**https://archive.ics.uci.edu/ml/datasets/KDD+Cup+1998+Data**](https://archive.ics.uci.edu/ml/datasets/KDD+Cup+1998+Data)

**Batch 8-** [**https://archive.ics.uci.edu/ml/datasets/Water+Quality+Prediction**](https://archive.ics.uci.edu/ml/datasets/Water+Quality+Prediction)

Download the dataset based on your batch and refer to the description given on the link. Perform following operations:

1. Perform steps 1 to 6 and 12 to 17 from assignment 1 on the given dataset.
2. If your dataset does not contain null values, create a dummy copy of the dataset by introducing null values. Perform following operations:
   1. Determine % of null values column-wise.
   2. Decide and perform imputation method for features having null values and justify choice of method.
3. Plot scatter plot of every numerical attribute with every other numerical attribute and note your observations with respect to correlation of the attributes.
4. Plot histogram of every numerical attribute and mention your observations in terms of normal behavior, skewness and kurtosis.
5. Plot bar graph for categorical attributes.
6. Plot box plot for all numerical attributes. Mention your observations about distribution of the data and anomalies.
7. Perform standardization and normalization of the numerical features and analyze the transformation visually using suitable visualization.
8. Perform dimensionality reduction using PCA and t-SNE in 2 dimensions. Compare visual plots and note your observations.
9. If *m* denotes number of features and *n* denotes number of data points, for n = 10000, create following datasets using randn function in python.
   1. m = 10
   2. m = 50
   3. m = 100
   4. m = 500
   5. m = 1000
   6. m = 1500

Plot m vs execution time for PCA and t-SNE and note your observations. You are free to experiment with many more *m* values. (use timeit module in python for computing execution time.)