**Healthcare**

Project 1

DESCRIPTION

**Problem Statement:**

ICMR wants to analyze different types of cancers, such as breast cancer, renal cancer, colon cancer, lung cancer, and prostate cancer becoming a cause of worry in recent years. They would like to identify the probable cause of these cancers in terms of genes responsible for each cancer type. This would lead us to early identification of each type of cancer reducing the fatality rate.

**Dataset Details:**

The input dataset contains 802 samples for the corresponding 802 people who have been detected with different types of cancer. Each sample contains expression values of more than 20K genes. Samples have one of the types of tumors: BRCA, KIRC, COAD, LUAD, and PRAD.

**Project Task: Week 1**

**Exploratory Data Analysis:**

1. Merge both the datasets.
2. Plot the merged dataset as a hierarchically-clustered heatmap.
3. Perform Null-hypothesis testing.

**Dimensionality Reduction:**

1. Each sample has expression values for around 20K genes. However, it may not be necessary to include all 20K genes expression values to analyze each cancer type. Therefore, we will identify a smaller set of attributes which will then be used to fit multiclass classification models. So, the first task targets the dimensionality reduction using various techniques such as,  
   PCA, LDA, and t-SNE.
2. Input: Complete dataset including all genes (20531)
3. Output: Selected Genes from each dimensionality reduction method

**Project Task: Week 2**

Clustering Genes and Samples:

1. Our next goal is to identify groups of genes that behave similarly across samples and identify the distribution of samples corresponding to each cancer type. Therefore, this task focuses on applying various clustering techniques, e.g., k-means, hierarchical and mean shift clustering, on genes and samples.

* First, apply the given clustering technique on all genes to identify:
  + Genes whose expression values are similar across all samples
  + Genes whose expression values are similar across samples of each cancer type
* Next, apply the given clustering technique on all samples to identify:
  + Samples of the same class (cancer type) which also correspond to the same cluster
  + Samples identified to be belonging to another cluster but also to the same class (cancer type)

**Building Classification Model(s) with Feature Selection:**

1. Our final task is to build a robust classification model(s) for identifying each type of cancer. It also aims at the to do feature selection in order to identify the genes that help in classifying each cancer type.

**Sub-tasks:**

1. Build a classification model(s) using multiclass SVM, Random Forest, and Deep Neural Network to classify the input data into five cancer types
2. Apply the feature selection algorithms, forward selection and backward elimination to refine selected attributes (selected in Task-2) using the classification model from the previous step
3. Validate the genes selected from the last step using statistical significance testing (t-test for one vs. all and F-test)