Datasets & Problem statement

1. Fashion-MNIST Dataset

Fashion-MNIST is a dataset consisting of 60,000 training examples and 10,000 test examples. Each example is a 28x28 pixels gray-scale image. Each image is labeled with 10 class categories.

Figure 1 shows example images in this dataset.

Here, each image is considered to be 784 dimensional data sample. So, there is a need to reduce the dimension of the data. Principal component analysis can be used for selecting the important features and create a lower dimensional feature vector for classification task. So, implement a module or a function to get the principal components for each of the data sample. Again, implement your own code for implementing PCA. Also, evaluate the variation of classification performance with variation in number of principal components that are included as part of feature vectors.

You can retrieve this dataset from https://github.com/zalandoresearch/fashion-mnist

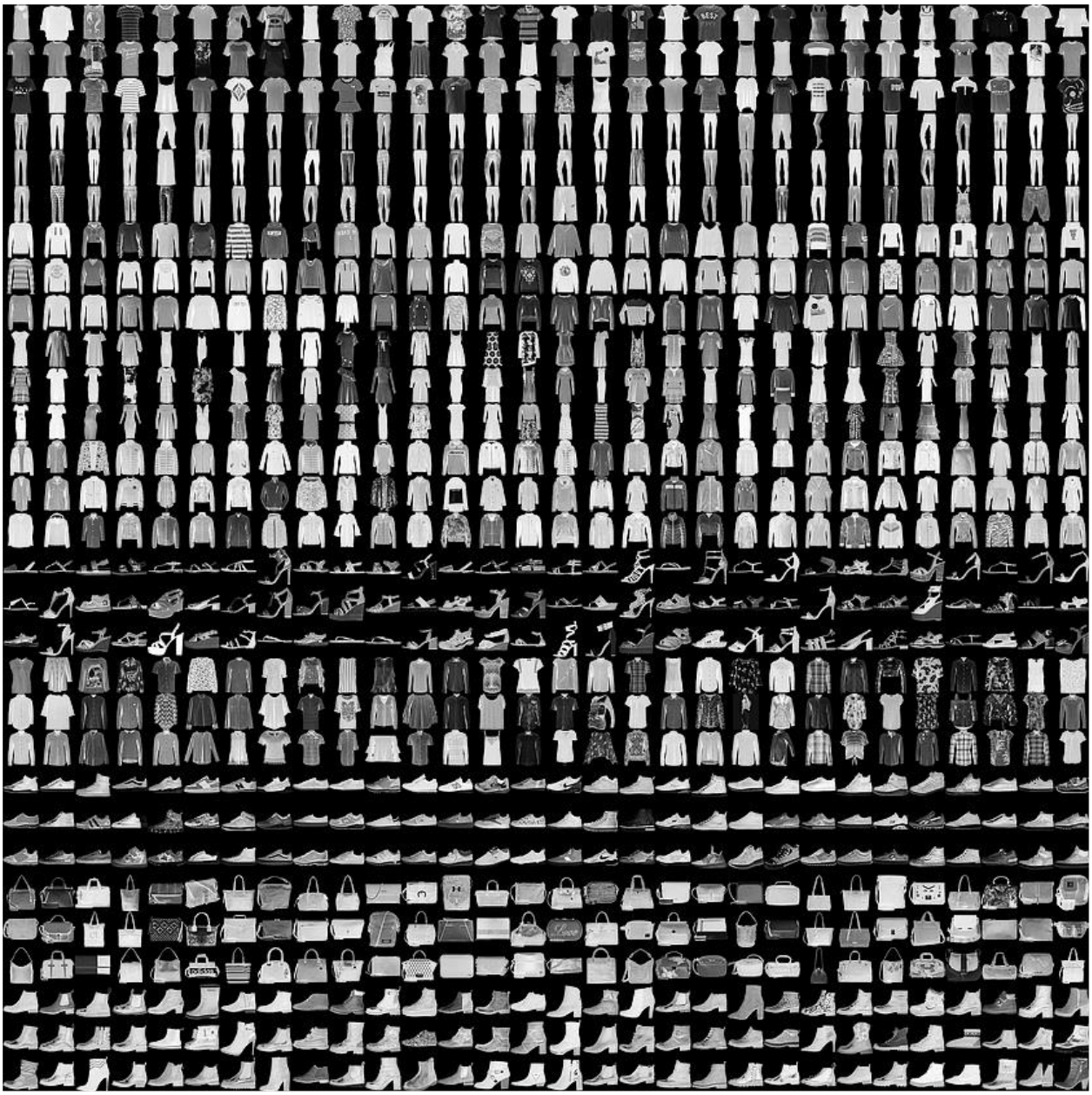

Figure 1: Fashion-MNIST Dataset

2. Blood Test

This dataset consist of outcomes of three Blood Tests (Test1, Test2 and Test3) for analyzing the condition of Heart of a patient. Doctors in a hospital is analyzing these outcomes and are providing the report for patient indicating whether the Heart is Healthy or the patient needs medication or there is a need of any kind of Surgery. This dataset is also containing the doctor's advice for whether the Heart is HEALTHY, MEDICATION and SURGERY based on the outcomes of the three tests.

You are required to create various kinds classifiers classifying patients in categories of Healthy, need medication or undergo surgery.
The snapshot of the dataset is shown in Figure 2.

Health	TEST1	TEST2	TEST3
MEDICATION	4.309531	-0.83201	0.051151
HEALTHY	2.452432	0.067027	1.825669
SURGERY	-0.16333	1.827832	-0.38598
HEALTHY	2.069746	-0.0386	2.77623
SURGERY	-0.32149	1.982645	1.475755
MEDICATION	1.289905	2.202437	0.129948
HEALTHY	1.875493	-0.22852	2.033736

Figure 2: Heart Health Test Dataset

3. Train Selection

Indian Railways has introduced a new luxury train from Mumbai to New Delhi. This train has all facilities like WiFi, Club, Lounge, Playing, SPA etc. Each of the facilities are chargeable along
with the travel fare. To analyze the interest shown by public, they floated a form with information such as Age, Sex, fare paid, number of members traveling with, Travel class etc. This form was filled by the person while booking the ticket for the train. After the first day launch of the train, the department analyzed whether the person has boarded the train or not.
The dataset that is provided contains all the information about the person along with whether the person as boarded the train or not. You need to create classifiers to classifying whether a person will board the train or not if provided with information such as age, fare paid, number of members traveling with etc.
The snapshot of the dataset is shown in Figure 3.

caseID	Whether boarded the Train?	Train Fare to be Paid	Number of Family Members travelling with	preferred Class	sex	Age Category
111131089	0	2201	0	FIRST_AC	female	2
2489059216	0	1775	3	FIRST_AC	male	0
1565109576	1	1775	3	FIRST_AC	female	0
1373075087	1	1775	3	FIRST_AC	male	3
1598041082	1	1775	3	FIRST_AC	female	2
3825576434	0	852	0	FIRST_AC	male	4
644658844	0	1207	1	FIRST_AC	female	6
688816386	1	710	0	FIRST_AC	male	3

Figure 3: Train Selection Dataset