Amirkabir University of Technology
(Tehran Polytechnic)

1. Clustering Football Players: Is Mbappé More Similar to Ronaldo Than Messi? (90 Pts.)

There are thousands of professional football players across the world, each with specific skills and unique playing style. Given their quantitative playing attributes, however, it might be possible to determine players with similar playing styles, and find groups of players with (almost) identical qualities.

To accomplish this task, we make use of *FIFA 23* player ratings. The dataset contains the information corresponding to over 18000 football players, each with 87 various features. Our goal is to investigate how these players can be assigned to distinctive clusters.

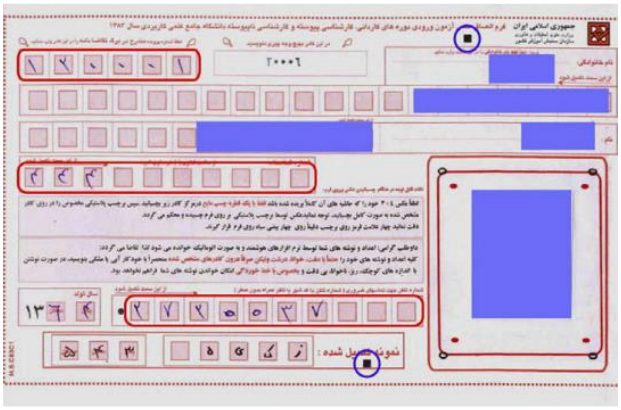
First, we assume the players with a rating above 85 (91 players in total), and discard all non-numeric features.

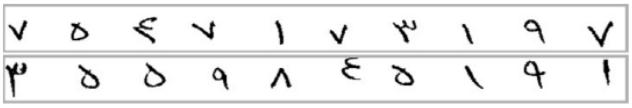
a. Normalize the data, and apply PCA to them so that the dimensions are reduced to 2.
b. Assuming $k = 5$, perform k-means clustering. Visualize clusters with players' names attached to each point.
c. Is Kylian Mbappé playing style more comparable to that of Lionel Messi or Cristiano Ronaldo? How about Mehdi Taremi?

Now, we wish to see how accurate this approach is in categorizing football players based on their positions. As can be seen in the dataset, each player is given a 'Best Position' attribute, denoting his most preferred position on the pitch.

d. Perform clustering on all the players with $k = 16$ (since there are 16 distinct positions listed), and calculate the clustering accuracy. Which position is clustered more accurately?

2. MDC for OCR: An Attempt to Classify Farsi Digits (70 Pts.)





A **Minimum Distance Classifier** attempts to classify an unlabelled sample to a class which minimise the distance between the sample and the class in multi-feature space. As minimising distance is a measure for maximising similarity, **MDC** actually assigns data to its most similar category.

While **MDC** might look too basic, it works pretty well in some problems. One of them could be **Optical Character Recognition** (**OCR**), where the goal is to distinguish handwritten or printed text characters inside digital images of documents. Here, we aim to apply this technique to the problem of Farsi digits classification. We use a dataset named Hoda, which contains 102353 samples of digits written by candidates of Karshenasi Arshad entrance exam in their registration forms, Figure 2. You are given a shorter version of this dataset in which the images are binary.

a. Use the training set to calculate the prototype of each class. Display the results.

b. Now use the test samples to evaluate your MDC classifier. Report the error, and display five erroneous predictions.

3. Please answer the following questions as clear as possible: (40 Pts.)

a. You have a DataFrame df with columns 'A', 'B', and 'C'. How would you efficiently find the top N rows in column 'C' for each distinct value in column 'A'?

b. If you have a DataFrame sales_data containing sales figures for different products across various regions and months, how can you create a new column that shows the percentage contribution of each product's sales to the total sales in each month?

c. Given a NumPy array arr of shape (m, n) representing a grayscale image, how would you efficiently apply a median filter of size 3x3 to the image without using explicit loops?

d. Assume you have two 1-dimensional NumPy arrays, a and b, both of size n. How can you efficiently find the index of the element in array a that is closest to each element in array b?

Good Luck!