Artificial Intelligence and Machine Learning Fundamentals

Activity 12: k-means Clustering of Sales Data

In this section, we will detect product sales that perform similarly to recognize trends in product sales.
We will be using the Sales Transactions Weekly Dataset, found at the following URL: https://archive.ics.uci.edu/ml/datasets/Sales_Transactions_Dataset_Weekly
Perform clustering on the dataset using the k-means algorithm. Make sure that you prepare your data for clustering based on what you have learned in the previous lessons. Use the default settings for the k-means algorithm:

1. Load the dataset using pandas. If you examine the data in the CSV file, you will realize that the first column contains product ID strings. These values just add noise to the clustering process. Also, notice that for weeks 0 to 51, there is a W-prefixed label and a normalized label. Using the normalized label makes more sense so that we can drop the regular weekly labels from the dataset. Create a k-means clustering model and fit the data points into 8 clusters. Retrieve the center points and the labels from the clustering algorithm.
2. The labels belonging to each data point can be retrieved using the labels_property. These labels determine the clustering of the rows of the original data frame. How are these labels beneficial?

Suppose that, in the original data frame, the product names are given. You can easily recognize the fact that similar types of products sell similarly. There are also products that fluctuate a lot, and products that are seasonal in nature. For instance, if some products promote fat loss and getting into shape, they tend to sell during the first half of the year, before the beach season.