Streamplate ML Challenge

The following task will assess your ability to train unsupervised ML models and your ability to write python code.

The problem

The following dataset contains images of handwritten digits from 0 to 9. Each image has dimensions of 28 x 28, meaning that each image has 784 pixels.

train.csv has 784 columns. Each row represents an image. Each of the 784 columns represent the pixels of the image. The number contained within each pixel indicates the brightness/darkness of that pixel, with higher numbers representing darker pixels.

test.csv has the same format, however, there is an additional first column (which is followed by the same 784 columns explained before). The first column of each row contains a number which indicates which number is drawn in the image.

Your task is to create and train an **unsupervised** machine learning model using the data in train.csv. Note that since this is an unsupervised model, you are **not expected** to know the number drawn in each image at the time of training. This unsupervised model should determine which images represent the same number and be able to cluster them into groups (your model **does not** need to label each of the groups).

After the model is created, use it to group the images from test.csv. In order to assess the accuracy of the groupings made, calculate and output the NMI score between the groupings created by your model and the manual labels provided in test.csv.

Format required

Write your solution using **python**. Your unsupervised model should be created using **Keras**, however, there is **no limitation** on what other modules you may use in your solution. Present your solution in a Google Collab notebook and share the notebook with bryan@streamplate.com and kim@streamplate.com.