▾ MLOps Assessment

▾ Create Dataset `(OBJECTIVE 1)`
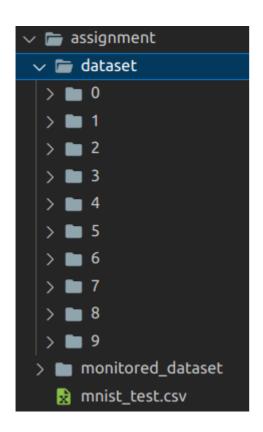
- A sample of MNIST data (`mnist_test.csv`) will be created once provided code in notebook is executed
- Data Description

  - First column is `label`
  - Next 784 columns represents the pixel value of 28x28 image marked from `1x1, 1x2 ... 28x28`

- Create Image dataset from the csv and store them in directories under `dataset`. Directory Structure will look like as shown below and each directory will contain `~1000 images`

Develop a Classification MLOps Pipeline `(OBECTIVE 2)`

- The pipeline should automatically retrain a new classification model as soon as there are `significant changes` in dataset.
- Some of the major steps in pipeline can include:

  - Monitor the dataset - `monitored_dataset` directory

- - if a new class is added with less than 200 examples - `Do Nothing`
    - if new class contains more than 200 examples - `trigger next step`
  - Split dataset into train/test/validation
  - Train classification model on train dataset with validation
  - Generate and store classification report with evaluation metrics (`Precision, Recall, F1-score`) on test set.
- Any other improvement over this would be a plus.
- Consider the scenario given below for better understanding.

Scenario

- Initially, only classes `0` and `1` are available in `monitored_dataset` with ~500 samples each, then `start pipeline -> train model -> generate report`
- A new class `2` is added with 150 samples, `Do Nothing`
- Another 100 samples of class `3` are added, `Do Nothing`
- Another 100 samples of class `2` are added, `start pipeline` and train only for classes having greater than 200 samples i.e. `0, 1, 2`.
- And, so on.