MNIST number dataset is a set of 70,000 small images of digits handwritten by high school students and employees of the US Census Bureau. Each image is labelled with the digit it represents. This set has been studied so much that it is often called the “hello world” of Machine Learning: whenever people come up with a new classification algorithm they are curious to see how it will perform on MNIST, and anyone who learns Machine Learning tackles this dataset sooner or later.

Use the following steps to complete this task:

Step 1: Use Jupyter Notebook for the interactive practice of Python and related Machine Learning packages – Google Colab or Anaconda.
Anaconda is recommended for this task as it is the most widely used Python distribution for data science and comes pre-loaded with all the most popular libraries and tools.
You can use these instructions to set up Jupyter with Ananconda on your machine: (5%）
a. Before installing Jupyter notebook, install Anaconda here: https://docs.anaconda.com/anaconda/install/
b. Create a virtual environment for each Python project.
c. For installing libraries, follow this guide: Installing Conda Packages
d. For creating a Jupyter notebook, follow this Using Jupyter Notebook guide. Note: always use kernel 3.X.
e. Familiarize yourself with cells in Jupyter notebook and practice mixing texts and python coding.
Step 2: Complete the following tasks:
a. Download the dataset (5%)
b. Explore the dataset, and output the following information:（10%）
i. the total number of images
ii. how many features there are and the range of feature values (for example, a histogram of the data value). Describe the link between the data (e.g numbers) and their real-world context: for example, if you're dealing with data to do with images, you might describe the dimensions, resolution and colour palette of the outputted images.
iii. how many types of categories/labels (discrete or continuous) there are
iv. visualize at least three randomly selected samples within each category (feel the variance of the data)
v. visualize enough data samples to see whether there are any bad data samples that need to be removed. How would you define a "bad data sample" in this particular context?
c. Do more data manipulation（10%）
i. Use two-dimensional reduction algorithms, PCA and t-SNE, to reduce the MNIST dataset down to two dimensions and plot each result using Matplotlib. Use a scatterplot using 10 different colours to represent each image’s target class.
ii. Load the MNIST dataset and split it into a training set and a test set (take the first 60,000 instances for training, and the remaining 10,000 for testing). Train a Random Forest classifier on the dataset and time how long it takes, then evaluate the resulting model on the test set.
iii. Next, use PCA to reduce the dataset’s dimensionality to 174. Train a new Random Forest classifier on the reduced dataset and see how long it takes. Was training much faster? Next, evaluate the classifier on the test set. How does it compare to the previous classifier?
Explain the functions you’ve chosen in detail. Try to understand the results you got and interpret your results in detail.
iv. Write a 200-300 word summary/conclusion of your discoveries and insights.
Structure:

Prepare a Jupyter notebook for this assignment. The structure of the Jupyter notebook should alternate texts and python codes and cover topics listed in the steps above.