Salma Chaaban – 301216551

COMP257 - Assignment 1 – Analysis Report

**Exercise 1**

1. The MNIST dataset, consisting of 70,000 instances of 28x28 pixel images of handwritten digits, was retrieved from the ARFF file and loaded to a Pandas data frame for easier manipulation.
2. The “class” column of the data frame, which represents the digit labels from 0-9, was converted from bytes into a string format for clarity, and the data was split into X (features) which are pixel values of the image, and y (labels) which represent the digits 0-9. To ensure that the data was loaded correctly and to see a sample of it, the first 10 digits of the data were displayed as grayscale images with their corresponding labels.
3. PCA was applied to the data, scaled and unscaled, for the unscaled data, the explained variance ratio for PCA for the first components was 0.097, and 0.071 for the second. Meaning that these components captured 9.7% and 7.1% of the total variance of the dataset. With scaled data, using StandardScaler, the results were 0.056 for the first component, and 0.04 for the second. The higher explained variance ratio values have a higher influence on the variance, which may not be representative of the overall data structure. So, scaling helps standardize pixel intensity values to have the same range, giving equal importance to all features. However, the results from the unscaled data are not that high; therefore, it was kept unscaled.
4. The first two principal components were projected on a 1D hyperplane. From the plots, both components seem to preserve a large amount of variance, with the first component preserving more. The results were similar with scaled data as well.
5. Incremental PCA was applied to reduce the dimensionality of the dataset down to 154 components.
6. To assess the results of dimensionality reduction, the original and compressed images were visualized. The first 10 digits were plotted side by side with their compressed, which demonstrated how much of the image quality gets affected by compression. Although the compressed images were slightly blurred, they are still easily recognizable. This means that reducing the dimensionality of the dataset to 154 components still preserves the essential structure of the digits. The results with scaled data were of much lower quality, they looked more blurred; however, they were still easily distinguishable.

**Exercise 2**

1. The Swiss Roll dataset was generated using the make\_swiss\_roll function from Scikit-learn. This dataset consists of 3D points arranged in a spiral (Swiss roll) shape.
2. The dataset was plotted into a 3D space to visualize its structure.
3. Kernel PCA (KPCA) was applied to the un-scaled dataset using three different kernels: Linear kernel, RBF kernel, and Sigmoid kernel.
4. The results of these kernels were plotted. With the Linear kernel, the data points were in a spiral shape, as the linear kernel does not capture the non-linear nature of the Swiss roll, with the RBF kernel capturing the non-linear structure and “unrolled” the Swiss roll, the Sigmoid kernel was also in a spiral shape like the Linear kernel.
5. A pipeline was created combining kernel PCA and logistic regression. A grid search was performed using GridSearchCV to find the best parameters. The parameter grid included three kernels (linear, RBF, and sigmoid) and a range of gamma values (0.03 to 0.05). The best parameters identified were: 'kpca\_\_gamma': 0.05, 'kpca\_\_kernel': 'rbf'.
6. The best parameters found by GridSearchCV were used to transform the data and plot it., which showed a similar result as the RBF plot from earlier.