Assignment 3

1. Recognizing hand-written Digits dataset

The digits dataset consists of 8x8 pixel images of digits. The images attribute of the dataset stores 8x8 arrays of grayscale values for each image. We will use these arrays to visualize the first 4 images. The target attribute of the dataset stores the digit each image represents.

*# Standard scientific Python imports*

**import** matplotlib.pyplot **as** plt

*# Import datasets, classifiers and performance metrics*

**from** sklearn **import** datasets**,** metrics**,** svm

**from** sklearn.model\_selection **import** [train\_test\_split](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html#sklearn.model_selection.train_test_split)

Plot 4 digits in a single row with corresponding possible training result. To apply a classifier on this data, we need to flatten the images, turning each 2-D array of grayscale values from shape (8, 8) into shape (64,).

We can then split the data into train and test subsets and fit a support vector classifier on the train samples. The fitted classifier can subsequently be used to predict the value of the digit for the samples in the test subset.

Hence do the following:

* flatten the images
* Create a classifier: a support vector classifier with gamma=0.001
* Split data into 70% train and 30% test subsets
* Learn by fitting the digits on the train subset
* Predict the value of the digit on the test subset
* Visualize the first 4 test samples and show their predicted digit value in the title.
* Now build a [classification\_report — scikit-learn 1.5.1 documentation](https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html#sklearn.metrics.classification_report) showing the main classification metrics.
* Plot a confusion matrix of the true digit values and the predicted digit values. Is this confusion matrix a sparse matrix?

1. Apply SelectKBest on load\_digits dataset.Use at least 3 different **score\_func like chi2,** f\_classif etc
2. Apply SelectPercentile on load\_digits dataset. Use at least 3 different **score\_func like chi2,** f\_classif etc and at least 3 different percentile values.

2.

· Use the iris dataset (load\_iris). Split the data into a train and test dataset. What happens when we do not put random\_state value?

· We want to use a k-nearest neighbors classifier considering a neighborhood of 11 data points. Since our k-nearest neighbors model uses euclidean distance to find the nearest neighbors, it is therefore important to scale the data beforehand. Which scaler do you want to use?

· Now, fit two classifiers with different values of the parameter weights and plot the decision boundary of each classifier as well as the original dataset to observe the difference.