**961733 Energy Data Analytics**

**Programming Assignment 0**

**“A Toy dataset and Simple Hold-out Sampling”**

**Program Description**

The goal is to get you more familiar with Python programming as a data science. You will be used numpy to generate a toy dataset and implement the r hold out sampling method that split the dataset into training and testing set.

Your function would act pretty much like the predefined method from *sklearn.model\_seleaction.train\_test\_split*. It should take three parameters including numpy array, test size, and shuffle state.

**Part 1 A toy dataset**

1. Create a numpy array as independent variables. The array must have n rows and n columns, the value inside that array must by an integer ranging between [0,99]
2. Create another column as a dependent variable and concatenate with the first array
3. Create another column as a indexer stack in front of the first array

**Part 2 A simple holdout method**

1. The test\_size can’t be greater than 50%
2. Return the training set and testing (train, train\_y, test, test\_y) based on the given test\_size and shuffle\_state.

**Example of input for the hold out method**

train, train\_y, test, test\_y = holdout(arr, 0.2, 1)

**Example of output**

**Creating a toy dataset...**

**array shape is:**

**(1000, 7)**

**Top 5 row is:**

**[[ 0 68 48 81 39 52 1]**

**[ 1 59 50 81 22 20 0]**

**[ 2 99 52 75 55 30 0]**

**[ 3 86 22 58 3 95 1]**

**[ 4 89 11 2 2 6 1]]**

**Buttom 5 row is:**

**[[995 22 14 87 36 92 1]**

**[996 26 45 56 73 89 1]**

**[997 27 1 80 90 64 1]**

**[998 98 18 85 0 48 1]**

**[999 10 52 89 3 76 0]]**

**Performing a simple hold-out method..**

**shuffle state is 1**

**After the simple holdout method:**

**Training set:**

**top 5 row(train) is:**

**[[180 68 54 71 40 13 0]**

**[141 92 35 30 18 14 0]**

**[586 95 2 55 55 92 1]**

**[827 34 52 95 40 98 1]**

**[279 64 93 46 98 13 1]]**

**buttom 5 row(train) is:**

**[[757 1 65 36 61 96 0]**

**[506 86 62 27 60 20 1]**

**[186 62 62 9 28 88 1]**

**[479 5 19 95 29 74 0]**

**[136 28 42 40 28 28 0]]**

**top 5 row(train\_y) is:**

**[0 0 1 1 1]**

**buttom 5 row(train\_y) is:**

**[0 1 1 0 0]**

**Testing set:**

**top 5 row(test) is:**

**[[999 10 52 89 3 76 0]**

**[484 96 71 58 1 25 0]**

**[649 97 92 4 27 98 0]**

**[704 34 4 52 53 24 1]**

**[402 20 89 6 48 16 1]]**

**buttom 5 row(test) is:**

**[[286 13 79 97 56 11 0]**

**[918 49 52 1 73 78 0]**

**[725 17 20 28 66 68 1]**

**[985 2 37 73 79 90 0]**

**[ 72 81 76 52 36 68 1]]**

**top 5 row(test\_y) is:**

**[0 0 0 1 1]**

**buttom 5 row(test\_y) is:**

**[0 0 1 0 1]**

**train shape: (800, 7)**

**train\_y shape: (800,)**

**test shape (200, 7)**

**test\_y shape (200,)**

**program terminated properly.**

**Submission:**

* Submit your source code .py file the Google classroom under the programming assignment section.
* Submit a pdf file showing your output for a dataset with 5555 rows and 5 columns
* Submit a second pdf showing your output for a dataset with 12000 rows and 10 columns