**Lab 2**

**조현식(team leader), 이연희, 장준, 함민혁**

**Project plan**

**<data exploration & data preprocessing (without scaling, encoding)>**

The dataset consists of a numerical value except for one feature. (median\_house\_value)

And the total\_bedrooms feature contains the NaN value. (207 records)

-> Drop the dirty record(NaN).

According to the professor's words, the median\_house\_value feature should be dropped.

-> Drop 'median\_house\_value' feature.

In addition, in order to process the Outlier, we need to plot the dataset and drop it for records that are more than(or less than) a specific value.

-> Drop outliers in each numerical features.

Make combinations of the features using PCA.

**<data preprocessing - scaling & encoding>**

Scaling :

We will use 5 different scaling methods. (MinMax, Robust, Standard, MaxAbs, Nomalizer)

Encoding :

For ocean\_proximity feature -> [ <1H OCEAN, INLAND, ISLAND, NEAR BAY, NEAR OCEAN ]

We will use 3 different encoding methods. (LabelEncoder, OneHotEncoder, OrdinalEncoder)

**<model building>**

We use 5 modeling algorithms and each parameters.

K-means :

n\_clusters

algorithm : {“lloyd”, “elkan”, “auto”, “full”}

EM(GMM) :

n\_components

init\_params{‘kmeans’, ‘k-means++’, ‘random’, ‘random\_from\_data’}

CLARANS :

number\_clusters

numlocal

maxneighbor

DBSCAN :

eps

metric : {'euclidean','manhattan'}

algorithm : {‘ball\_tree’, ‘kd\_tree’, ‘brute’}

Spectral clustering :

n\_components

gamma

n\_neighbors

eigen\_solver : {‘arpack’, ‘lobpcg’, ‘amg’}

**<testing>**

Through the silhouette score and elbow method, check how many clusters have high accuracy.

Test how evenly the data is distributed inside the cluster with purity.

**<Flowchart of simple code>**

def AutoML(scaler list, encoder list, model list, dataset, test\_size)

load\_and\_modification\_dataset()

for scaler in [scalers]:

Scaling\_Encoding()

for model in [models]:

Modeling()

Testing\_Plotting()

main(){

// original dataset

AutoML(...)

// feature combination dataset

AutoML(...)

}