**Lab 1**

**조현식(team leader), 이연희, 장준, 함민혁**

**Project plan**

**<data preprocessing - without scaling & encoding>**

**1. data exploration**

The number of records is 699 ,but the unique value is 645. There is a record with overlapping IDs. Because the contents of the data in the overlapping ID are different, we do not remove the overlapping ID data. (We will drop ID feature)

Bare Muclei data contains missing data.('?')

-> So, Drop the whole feature OR Delete missing data only.

Q. Is there anything to drop for feature?

A. No. Even if it is concentrated on one side, it is because it is a disease search, so a small number of data can be important.

Therefore, We think to save all data because a small number of data is likely to be important.

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

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

Encoding : dataset is all numerical type. So, Encoding is unnecessary.

**<model building>**

Data split : test\_size={0.1, 0.2, 0.3}

We use 4 modeling algorithms and each parameters.  
logistic regression : solver = {‘newton-cg’, ‘lbfgs’, ‘liblinear’, ‘sag’, ‘saga’}

DT (entropy) : criterion = entropy, max\_features= {“auto”, “sqrt”, “log2”},

DT (gini) : criterion = gini, max\_features= {“auto”, “sqrt”, “log2”}

SVC : kernel = {‘linear’, ‘poly’, ‘rbf’, ‘sigmoid’, ‘precomputed’}, gamma = {‘scale’, ‘auto’}, C = {0.01, 0.1, 1}

In addition, in the case of model instances, it is collected separately in the array for score calculation in testing.

**<testing>**

Calculate the score by adjusting the k(Cross validataion paramater 3,5,7) value with the model instances stored in the array. Also, print the top 5 accuracy model.

**<Flowchart of simple code>**

def do\_classification()

load\_and\_modification\_dataset()

for scaler in [scalers]:

Scaling()

DataSplit() ← test\_size = {0.1, 0.2, 0.3}

for model in [models]:

Modeling()

newArr ← saving model instance

CrossValidation() ← k={3,5,7}

top\_5 = findTop5()

return top\_5

main(){

top\_5 = do\_classification()

print(top\_5)

}