과목 명: 시스템프로그래밍

담당 교수 명: 김 지 환

<<Project 6>>

**서강대학교 컴퓨터공학과**

**[20141556]**

**[이성문]**

목 차

1. 프로젝트 목표 3

2. 프로젝트 설명 3

2.1 합성

2.2 처리와 결과

2.3 시험

3. ML Algorithm

3.1 알고리즘 이름

3.2 알고리즘 상세 내용

# 프로그램 목표

본 프로젝트에서는 사용자가 직접 데이터를 가지고 학습 및 서비스를 진행할 수 있는 Azure Machine Learning 서비스를 이용하여 ML 모델을 구축한다.

주어진 데이터와 Azure Machine Learning 사이트에서 제공하는 ML algorithm으로 Experiment를 구성한다.

# 프로그램 설명

## 합성

- 93개의 feature를 갖는 데이터 (train/test set)을 이용한다.

- train 데이터 : 45,678개

- test 데이터 : 16,201개

- 데이터는 .csv 형식으로 주어지며 각 item이 각 feature 값으로 총 9개의 class로 나뉜다.

## 처리와 결과

주어진 train data로 모델을 학습하도록 하고 test data로 scoring을 거친 후 evaluation한다

Test data에 대한 결과(Accuracy를 확인할 수 있는 table)

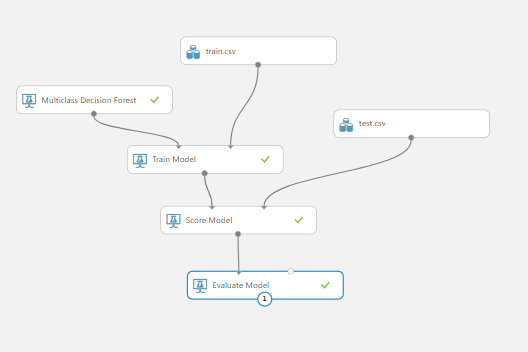
Evaluation Model – Evaluation results – visualize Overall accuracy

## 시험

입력 파일은 사이버 캠퍼스에 주어진 train data와 test data를 이용한다.

더 나은 성능을 위해 ML algorithm을 구성한다.

# ML algorithm



위의 사진에서 보는 바와 같이 본인은 이번 ML 알고리즘의 흐름도를 작성하였다.

명세서에서 제시한대로, train 데이터와 Multiclass Decision Forest 알고리즘을 이용하여 train 시켜서 Model을 하나 만들었고 그 모델에 대해 test 데이터를 이용해 Scoring하여 그 결과를 Evaluate할 수 있도록 하였다.

## 알고리즘 이름 : Multiclass Decision Forest

우리가 여기서 값을 예측하려는 feature은 “target” 이란 feature인데 이는 나올 수 있는 class의 개수가 총 9개이다. 즉, 수업시간에 이용했던 two-class 알고리즘은 2개의 결과밖에 보여주지 못하므로 이용할 수가 없다. 대신에 우린 9개의 calss를 예측할 수 있는 multiclass 알고리즘을 선택한다. Multiclass 알고리즘 5개 중에서도 제일 정확도가 높게 나온 알고리즘인 Multiclass Decision Forest 알고리즘으로 선택하였다.

## 알고리즘 상세 내용

\*Multiclass Decision Forest의 구성

-Resampling method :

**Bagging**: Bagging is also called bootstrap aggregating. In this method, each tree is grown on a new sample, created by randomly sampling the original dataset with replacement until you have a dataset the size of the original. The outputs of the models are combined by voting, which is a form of aggregation. For more information, see the Wikipedia entry for Bootstrap aggregating.

**Replicate**: In replication, each tree is trained on exactly the same input data. The determination of which split predicate is used for each tree node remains random, creating diverse trees.

-Create trainer mode:

**Single Parameter**: Select this option if you know how you want to configure the model, and provide a set of values as arguments.

**Parameter Range**: Use this option if you are not sure of the best parameters, and want to use a parameter sweep.

-Number of decision trees

Type the maximum number of decision trees that can be created in the ensemble. By creating more decision trees, you can potentially get better coverage, but training time might increase.

This value also controls the number of trees displayed in the results, when visualizing the trained model. To see or print a single tree, you can set the value to 1; however, this means that only one tree can be produced (the tree with the initial set of parameters), and no further iterations are performed.

-Maximum depth of the decision trees

Type a number to limit the maximum depth of any decision tree. Increasing the depth of the tree might increase precision, at the risk of some overfitting and increased training time.

-Number of random splits per node

Type the number of splits to use when building each node of the tree. A split means that features in each level of the tree (node) are randomly divided.

-Minimum number of samples per leaf node

Indicate the minimum number of cases that are required to create any terminal node (leaf) in a tree. By increasing this value, you increase the threshold for creating new rules.

<출처 – <https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/multiclass-decision-forest>>

이 알고리즘에서 각 변수가 담당하는 역할에 따라 변수값을 여러 케이스로 조정하면서 나타나는 overall accuracy를 비교해보아 제일 큰 accuracy가 나오는 변수들의 값을 정하였다. 그 결과 Resampling method은 “Bagging” , Create trainer mode은 “Single Parameter”, Number of decision trees은 100, Maximum depth of the decision trees은 40, Number of random splits per node은 3000, Minimum number of samples per leaf node은 1로 결정하였다.

