

# Assignment #3

## Competition using decision tree

Introduction to Big Data Analytics

TA: Dongmin Hyun and Junyoung Hwang

([dm.hyun@postech.ac.kr](mailto:dm.hyun@postech.ac.kr), [jyhwang@postech.ac.kr](mailto:jyhwang@postech.ac.kr))

# Machine Learning Practice with Kaggle

- In this Homework, you will practice to implement, train, and tune a decision tree on two datasets using Kaggle.
- Kaggle is a framework for evaluating the performance of ML models (you can find a lot of tutorials for Kaggle).
- General procedures
  - Train your ML model using given training data from given links
  - Produce predictions using the trained model on test data (You can find the output format as '[data name]\_sample\_submission.csv' from the given links).
  - Submit the output for test data to Kaggle to evaluate the performance of your model.

# Notes

- Your score is based on your ranking at Kaggle (private leader board).
  - The ranking of public and private leader board could be different.
- You have to submit your code and report (1~2 pages) to LMS.
  - The report includes summary of your process and your submission score.
- You can use ***any*** python library for this competition.
- You must use your student ID number (ex. 20181234) as your team name.
- *You must read all competition rules carefully and comply with them.*

# Kaggle links for two Datasets


- San Francisco Crime Dataset (Multi-class classification)

<https://www.kaggle.com/t/c73f61dbb1804cb7965fd7898e20efab>

- Bank Dataset (Binary Classification)

<https://www.kaggle.com/t/22b72934f5c342aaad360509efb995f4>

# Kaggle Competition (Join using the given links)

 InClass Prediction Competition

## Binary Classification using Decision Tree

The classification goal is to predict if a client will subscribe a term deposit

14 days to go

[Overview](#) [Data](#) [Kernels](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Join Competition](#)

### Overview

<b>Description</b>	The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be (or not) subscribed.
<b>Evaluation</b>	The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y). <i>Please note: Only provided training data must be used to train your decision tree.</i>
	<b>Sources</b>  [Moro et al., 2011] S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology.  In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimarães, Portugal, October, 2011. EUROSIS.

# Kaggle Competition (Join Competition)

The screenshot shows the Kaggle competition interface for the 'InClass Prediction Competition'. The header includes the competition name and a 'Join Competition' button. The main content area is titled 'Binary Classification using Decision Tree' and describes the goal: 'The classification goal is to predict if a client will subscribe a term deposit'. A navigation bar contains links for Overview, Data, Kernels, Discussion, Leaderboard, and Rules. The 'Overview' section is active, showing a 'Description' and 'Evaluation' tab. A modal dialog is overlaid on the page, prompting the user to accept the competition rules. The dialog contains the text: 'By clicking on the "I understand and accept" button below, you are indicating that you agree to be bound to the competition rules.' Below this text are two buttons: 'I Do Not Accept' and 'I Understand and Accept'. The 'I Understand and Accept' button is highlighted with a red border. The background content is partially obscured by the modal dialog.

InClass Prediction Competition

## Binary Classification using Decision Tree

The classification goal is to predict if a client will subscribe a term deposit

14 days to go

Overview Data Kernels Discussion Leaderboard Rules [Join Competition](#)

### Overview

#### Description

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client (or not) subscribed.

#### Evaluation


By clicking on the "I understand and accept" button below, you are indicating that you agree to be bound to the [competition rules](#).

[I Do Not Accept](#) [I Understand and Accept](#)

[Moro et al., 2011] S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology.

In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimarães, Portugal, October, 2011. EUROSIS.

# Kaggle Competition (Submit Predictions)

 InClass Prediction Competition

## Binary Classification using Decision Tree

The classification goal is to predict if a client will subscribe a term deposit

14 days to go

[Overview](#) [Data](#) [Kernels](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Team](#) [My Submissions](#) [Submit Predictions](#)

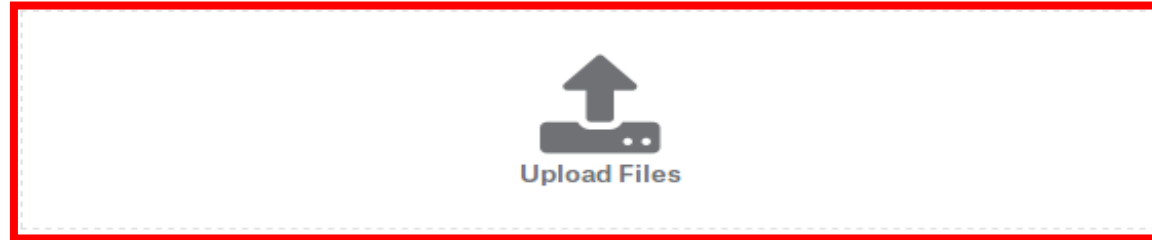
### Overview

<b>Description</b>	The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be (or not) subscribed.
<b>Evaluation</b>	The classification goal is to predict if the client will subscribe (yes/no) a term deposit (variable y). <i>Please note: Only provided training data must be used to train your decision tree.</i>
	<b>Sources</b>  [Moro et al., 2011] S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology.  In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimarães, Portugal, October, 2011. EUROSIS.

# Kaggle Competition (Submit Predictions)

You have 2 submissions remaining today. This resets 11 hours from now (00: 00 UTC).

## Step 1 Upload submission file



### File Format

Your submission should be in CSV format. You can upload this in a zip/gz/rar/7z archive, if you prefer.

### Number of Predictions

We expect the solution file to have 7234 prediction rows. This file should have a header row. Please see sample submission file on the [data page](#).

## Step 2 Describe submission

**B** *I* | 🔗 “ </> 🖼️ | ☰ ☷ H 📏 | ↺ ↻ [M+] Styling with Markdown supported

Briefly describe your submission.

Make Submission



# Kaggle Competition (Submit Predictions)

## Step 1

Upload submission file



✓ **bank\_sample\_submission.csv** (56.77 kB)

### File Format

Your submission should be in CSV format.  
You can upload this in a zip/gz/rar/7z  
archive, if you prefer.

### Number of Predictions

We expect the solution file to have 7234 prediction rows. This  
file should have a header row. Please see sample submission file  
on the [data page](#).

## Step 2

Describe submission



 Styling with Markdown supported

Briefly describe your submission.

**Make Submission**

# Kaggle Competition (Check Score)

InClass Prediction Competition

Binary Classification using Decision Tree

The classification goal is to predict if a client will subscribe a term deposit

14 days to go

OverviewDataKernelsDiscussionLeaderboardRules

Submit Predictions

Your most recent submission

Name	Submitted	Wait time	Execution time	Score
bank_sample_submission.csv	a few seconds ago	22 seconds	0 seconds	0.19928

Complete

[Jump to your position on the leaderboard](#)

Public LeaderboardPrivate Leaderboard

This leaderboard is calculated with approximately 30% of the test data.  
The final results will be based on the other 70%, so the final standings may be different.

[Raw Data](#)[Refresh](#)

#	Δ1w	Team Name	Kernel	Team Members	Score	Entries	Last
📍		Random prediction			0.19928		
1	new				0.19928	1	<10s

Your Best Entry

Your submission scored 0.19928

[Tweet this!](#)

# Python Library for Competitions

- Anaconda(<https://www.anaconda.com/download/>)
  - Almost all packages are installed in anaconda including below.
- Data preprocessing
  - Pandas (<http://pandas.pydata.org/>)
  - matplotlib (<https://matplotlib.org/>)
- Machine Learning algorithm
  - Scikit-learn (<http://scikit-learn.org/stable/>)
  - Numpy (<http://www.numpy.org/>)
  - ...