Họ tên: **Nguyễn Hữu Hiệu**
MSSV: **20520506**

BANK MARKETING DESCRIPTION

1. **Code book**

- **Bank**

Thông tin	Nội dung
Tên bộ dữ liệu	Bank Marketing
Nguồn thu thập	[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.
Số thuộc tính	17
Số mẫu dữ liệu	49732
Thông tin thuộc tính	Các biến đầu vào: 1 - age (int) 2 - job : loại công việc (phân loại: "admin.", "unknown", "unemployed", "management", "housemaid", "entrepreneur", "student", "blue-collar", "self-employed", "retired", "technician", "services") 3 - marital: tình trạng hôn nhân (phân loại: "đã kết hôn","ly hôn","độc thân"; lưu ý: "ly hôn" có nghĩa là đã ly hôn hoặc góa bụa) 4 - education (phân loại: "không rõ","trung học","tiểu học","đại học") 5 -default: có tín dụng trong vỡ nợ? (nhị phân: "có","không") 6 - balance: số dư trung bình hàng năm, tính bằng euro (số) 7 - housing: có vay mua nhà không? (nhị phân: "có","không") 8 -loan: có khoản vay cá nhân? (nhị phân: "có","không") # có liên quan với liên hệ cuối cùng của chiến dịch hiện tại: 9 -contact: loại liên lạc của liên hệ (phân loại: "không xác định","điện thoại","di động") 10 - day: ngày liên hệ cuối cùng trong tháng (số) 11 - month: tháng liên hệ cuối cùng trong năm (phân loại: "jan", "feb", "mar", ..., "nov", "dec") 12 -duration: thời lượng liên lạc cuối cùng, tính bằng giây (số) # thuộc tính khác: 13 -campaign: số lượng liên hệ được thực hiện trong chiến dịch này và cho khách hàng này (số, bao gồm liên hệ cuối cùng) 14 - pdays: số ngày trôi qua sau khi khách hàng được liên hệ lần cuối từ chiến dịch trước đó (số, -1 có nghĩa là khách hàng chưa được liên hệ trước đó) 15 -previous: số lượng liên hệ được thực hiện trước chiến dịch này và cho khách hàng này (số) 16 - poutcome: kết quả của chiến dịch tiếp thị trước đó (phân loại: "không xác định","khác","thất bại","thành công") Biến đầu ra (mục tiêu mong muốn): 17 - y - khách hàng đã đăng ký tiền gửi có kỳ hạn chưa? (nhị phân: "có","không")
Thông tin tác giả	Paulo Cortez (Univ. Minho) and Sérgio Moro (ISCTE-IUL) @ 2012

- **Bank-additional**

Thông tin	Nội dung
Tên bộ dữ liệu	Bank Marketing (with social/economic context)
Nguồn thu thập	[Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing.
Số thuộc tính	21
Số mẫu dữ liệu	45307
Thông tin thuộc tính	Các biến đầu vào: 1 - age (số) 2 -job : loại công việc (phân loại: "quản trị viên.","cổ cồn xanh","doanh nhân","người giúp việc nhà","quản lý","đã nghỉ hưu","tự kinh doanh","dịch vụ","sinh viên" "kỹ thuật viên","thất nghiệp","không xác định") 3 - marital : tình trạng hôn nhân (phân loại: "ly hôn","đã kết hôn","độc thân", "không xác định"; lưu ý: "ly hôn" có nghĩa là đã ly hôn hoặc góa bụa) 4 -education (phân loại: "cơ bản.4y", "cơ bản.6y", "cơ bản.9y", "trung học phổ thông", "mù chữ", "chuyên nghiệp.khóa học", "đại học.bằng cấp", "không xác định") 5 - default: có tín dụng trong vỡ nợ? (phân loại: "không","có","không xác định") 6 -housing: có vay mua nhà không? (phân loại: "không","có","không xác định") 7 -loan: có khoản vay cá nhân? (phân loại: "không","có","không xác định") 8 -contact: loại liên lạc của liên hệ (phân loại: "di động","điện thoại") 9 - month: tháng liên hệ cuối cùng trong năm (phân loại: "jan", "feb", "mar", ..., "nov", "dec") 10 -day_of_week: ngày liên hệ cuối cùng trong tuần (phân loại: "thứ hai","thứ ba","thứ tư","thu","thứ sáu") 11 -duration: thời lượng tiếp xúc cuối cùng, tính bằng giây (số). Lưu ý quan trọng: thuộc tính này ảnh hưởng lớn đến mục tiêu đầu ra (ví dụ: nếu thời lượng = 0 thì y = "không"). 12 -campaign: số lượng liên hệ được thực hiện trong chiến dịch này và cho khách hàng này (số, bao gồm liên hệ cuối cùng) 13 - pdays: số ngày trôi qua sau khi khách hàng được liên hệ lần cuối từ chiến dịch trước đó (số; 999 có nghĩa là khách hàng chưa được liên hệ trước đó) 14 -previous: số lượng liên hệ được thực hiện trước chiến dịch này và cho khách hàng này (số) 15 - poutcome: kết quả của chiến dịch tiếp thị trước đó (phân loại: "thất bại","không tồn tại","thành công") 16 - emp.var.rate: tỷ lệ thay đổi việc làm - chỉ số hàng quý (số) 17 - cons.price.idx: chỉ số giá tiêu dùng - chỉ số hàng tháng (số) 18 - cons.conf.idx: chỉ số niềm tin của người tiêu dùng - chỉ số hàng tháng (số) 19 - euribor3m: tỷ lệ euribor 3 tháng - chỉ báo hàng ngày (số) 20 - nr.employed: số lượng nhân viên - chỉ số hàng quý (số) Biến đầu ra (mục tiêu mong muốn):

	21 - y - khách hàng đã đăng ký tiền gửi có kỳ hạn chưa? (nhị phân: "có","không")
Thông tin tác giả	Sérgio Moro (ISCTE-IUL), Paulo Cortez (Univ. Minho) and Paulo Rita (ISCTE-IUL) @ 2014

2. **Raw data**

Raw data gồm các file **bank-additional-full.csv, bank-additional.csv, bank-full.csv, bank.csv**.

3. **Tidy data**

Tidy data được lưu lại thành 2 file **bank_additional.csv, bank.csv**.

4. **Instruction list**

```
# Import libraries
import os
import time
import pandas as pd
import glob

# Insert the directory path in here
path = "./Raw_Data"

# Create a list of dataframes to concat later
lst_df_additional = []
lst_df = []

# Extracting all the contents in the directory corresponding to path
l_files = os.listdir(path)
print(l_files)

# create full file path for each file in l_files:
full_file_paths = glob.glob(os.path.join(path, "*.csv"))
print(full_file_paths)

# for additional bank
file_paths_additional = os.path.join(path, "*additional*.csv")
list_file_paths_additional = glob.glob(file_paths_additional)
print(list_file_paths_additional)

for file_additional in list_file_paths_additional:
  df = pd.read_csv(file_additional, sep=";")
  lst_df_additional.append(df)
  # print(df.shape)

final_df_additional = pd.concat(lst_df_additional)
# print(final_df_additional.shape)

bank_file_paths = list(set(full_file_paths) -
set(list_file_paths_additional))
# print(file_path)

for file in bank_file_paths:
  df = pd.read_csv(file, sep=";")
  lst_df.append(df)

final_df_bank = pd.concat(lst_df)
```

```
df_additional_drop_dup = final_df_additional.drop_duplicates()
df_bank_drop_dup = final_df_bank.drop_duplicates()

print(df_additional_drop_dup.shape)
print(df_bank_drop_dup.shape)
df_additional_drop_dup.to_csv('bank_additional.csv', index=True)

df_bank_drop_dup.to_csv('bank.csv', index=True)
```