# Pattern Recognition Proposal Report

Group 3
2391007 Minjeong Kim, 2391016 Suahn Lee, 2391017 Yoonji Lee

## 1. Abstract

In this project, we used Euribor and basic years of education data, which allowed us to specify that this dataset was from Portugal. Moreover, we found correlations between variables and abnormalities through the data visualization process. Concerning new information, we conducted data preprocessing, which involves data reduction, duplicate value detection, dealing variables, etc. Finally, we made our model proposal considering significant features of the dataset.

## 2. Exploratory Data Analysis(EDA)

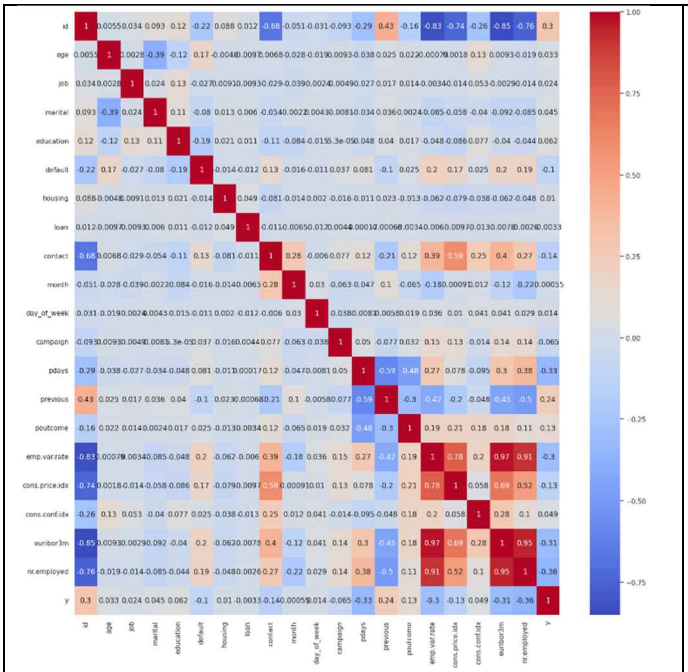
### A. Data Exploration

#### i. We just found that it is a “PORTUGUESE” Data Set!

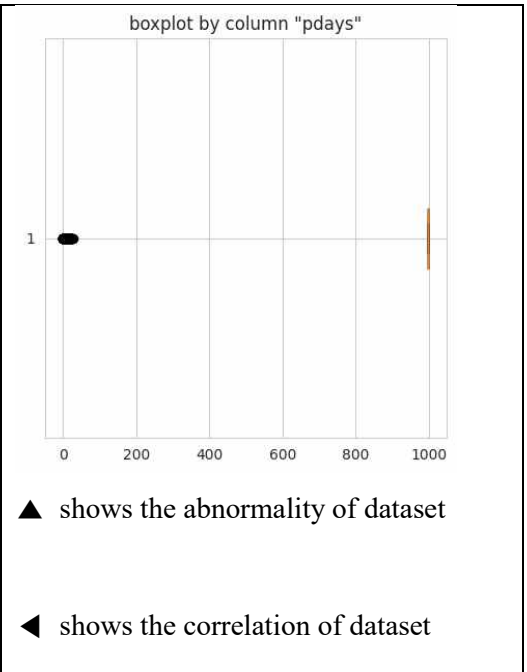
We found out this data source is from ‘Portugal’. Evidence follows; First, the acronym Euribor denotes the Euro Interbank Offered Rate. Since 'euribor3m' is included in the given data set, we assumed this data set is from one of the European Union member states. Additionally, according to the 'education' column, the education system of this country includes basic 4 years, basic 6 years, and basic 9 years of schooling before high school. Portugal is the only country that has the same education system among the 27 members of the EU. Finally, we compared the ‘cons.price.idx’ data within the dataset and Portugal’s actual Consumer Price Index (CPI). Every value from the data set matched the actual CPI from 2008 to 2010. Based on the information we gathered, we concluded that this data is from Portugal and added a new column called 'year'.

### B. Data Visualization

From the heatmap above, it is clear that three variables ‘emp.var.rate’, ‘euribor3m’, and ‘nr.employed’ have abnormally high Pearson correlation coefficients. The correlation coefficients for "euribor3m" and "nr.employed," "emp.var.rate" and "nr.employed," and "euribor3m" and "nr.employed" are 0.95, 0.91, and 0.97, respectively.



Heatmap between numerical variables




Boxplot of column “pdays”

### C. Data Preprocessing

As data visualization has shown, the values of ‘pdays’ were abnormal, so we deleted that column. Then we checked if there were any duplicate values, and nothing was found. The next step was changing the 'month’ variable into a discrete variable and labeling individuals over 65 years old as ‘retired’. The age of 65 came from the average retirement age in Portugal during 2008-2010. The ‘job’, ‘marital’, and ‘loan’ variables had some data labeled as ‘unknown’. We filled this with the mode value of each variable. Lastly, the given data set had a multicollinearity problem, which was caused by an abnormally high correlation between ‘nr.employed’ and ‘euribor3m’. To solve this problem, both variables were removed.

## 3. Model Proposal

Based on our exploratory data analysis (EDA), we intend to use tree-based models in conjunction with ensemble techniques. Tree-based models and ensemble techniques perform better than other methods when coping with categorical variables. Specifically, the models we are considering are as follows:

i. Decision Tree
ii. Random Forest Classifier

We will include Gradient Boosting for comparison with Decision Tree and Random Forest Classifier later. Also, due to the target imbalance problem (we found that the target ratio of “Yes” is around 11%), we plan to introduce both the Undersampling and Oversampling methods to solve this problem.