

WQD 7004

PROGRAMMING FOR DATA SCIENCE

SEMESTER II

SESSION 2018 / 2019

GROUP PROJECT

TOPIC

BANKING TERM DEPOSIT CLASSIFICATION

GROUP NAME

???????

GROUP MEMBERS

CHEAH JUN YITT WQD180107

TAN YIN YEN WQD180108

CHOO JIAN WEI WQD180124

CHOONG ENJUN WQD180113

INSTRUCTORS

PROFESOR DR. KASTURI DEWI A/P VARATHAN

SUBMITTED BY

21st MAY 2019

**Introduction:**

The Inter-American Development Bank (IDB) works to improve lives in Latin America and the Caribbean by providing financial and technical support for countries working to reduce poverty and inequality. Household poverty is an important key to IDB’s strategies. In Latin America, Proxy Means Test is one popular method used in Latin America to verify income qualification. However, IDB wishes to improve on PMT method and thus, has turned to the Kaggle Community for help with poverty prediction on some of the world’s poorest households.

**1.0 Details of Dataset**

Title of Dataset : Costa Rican Household characteristics Dataset

Year : 2018

Number of Observations : 9557 observations on Train Dataset;

: 23856 observations on Test Dataset

Number of Variables : 142 columns

Data Characteristics : Multivariate

Source : Uploaded for Kaggle Community Competition by Inter-

American Development Bank (IDB)

<https://www.kaggle.com/c/costa-rican-household-poverty-prediction/>

Purpose of Dataset : To challenge existing Proxy Means Test (PMT) algorithm on

household poverty prediction

**1.0 Preparation and cleaning of Dataset**

There are multiple discrepancies in the dataset. Also, we refer to the Kaggle discussion for additional insights regarding the data.

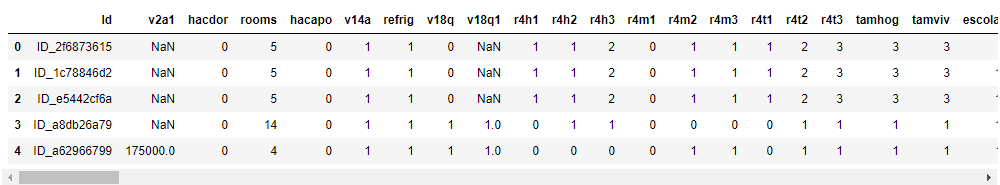


Figure1: Partial snapshot of data

Following can be said about the raw dataset:

1. Each line represents to an individual in the household
2. The target (poverty level) is segregated to 4 levels (1: extreme poverty, 2: moderate poverty, 3: vulnerable households, 4: non-vulnerable households)
3. There are individuals without household
4. There are individuals within the same household that does not share the same target poverty level
5. Prediction is scored based on head of household only