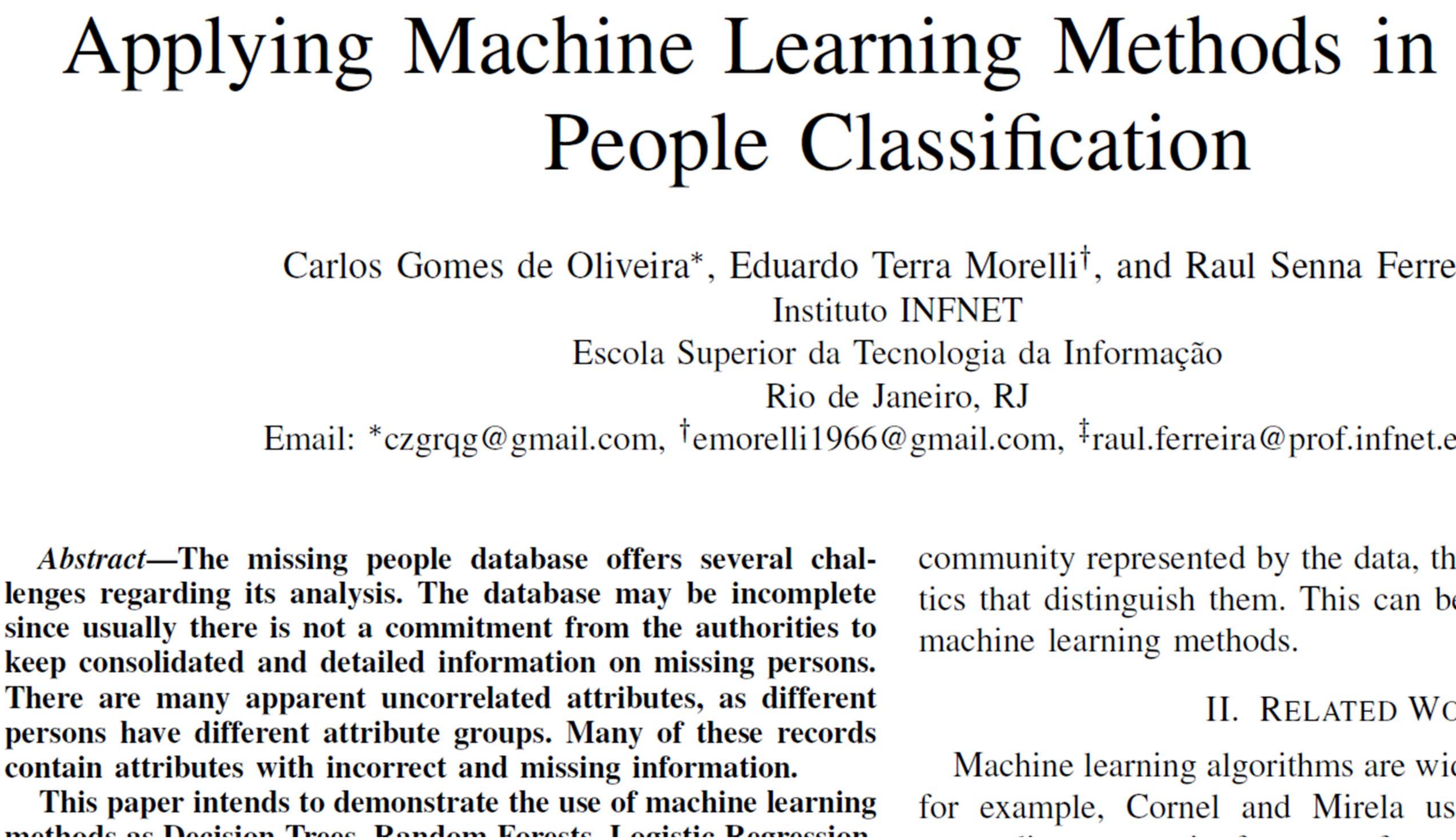
Machine Learning Methods in Vulnerable Communities Classification

- Carlos Gomes de Oliveira
  DBA and Data Engineer

  Linkedin.com/in/CarlosOliveira



Applying Machine Learning Methods in People Classification

Carlos Gomes de Oliveira*, Eduardo Terra Morelli†, and Raul Senna Ferre
Instituto INFNET
Escola Superior da Tecnologia da Informação
Rio de Janeiro, RJ
Email: *czgrqg@gmail.com, †emorelli1966@gmail.com, ‡raul.ferreira@prof.infnet.e

***Abstract*—The missing people database offers several challenges regarding its analysis. The database may be incomplete since usually there is not a commitment from the authorities to keep consolidated and detailed information on missing persons. There are many apparent uncorrelated attributes, as different persons have different attribute groups. Many of these records contain attributes with incorrect and missing information.**

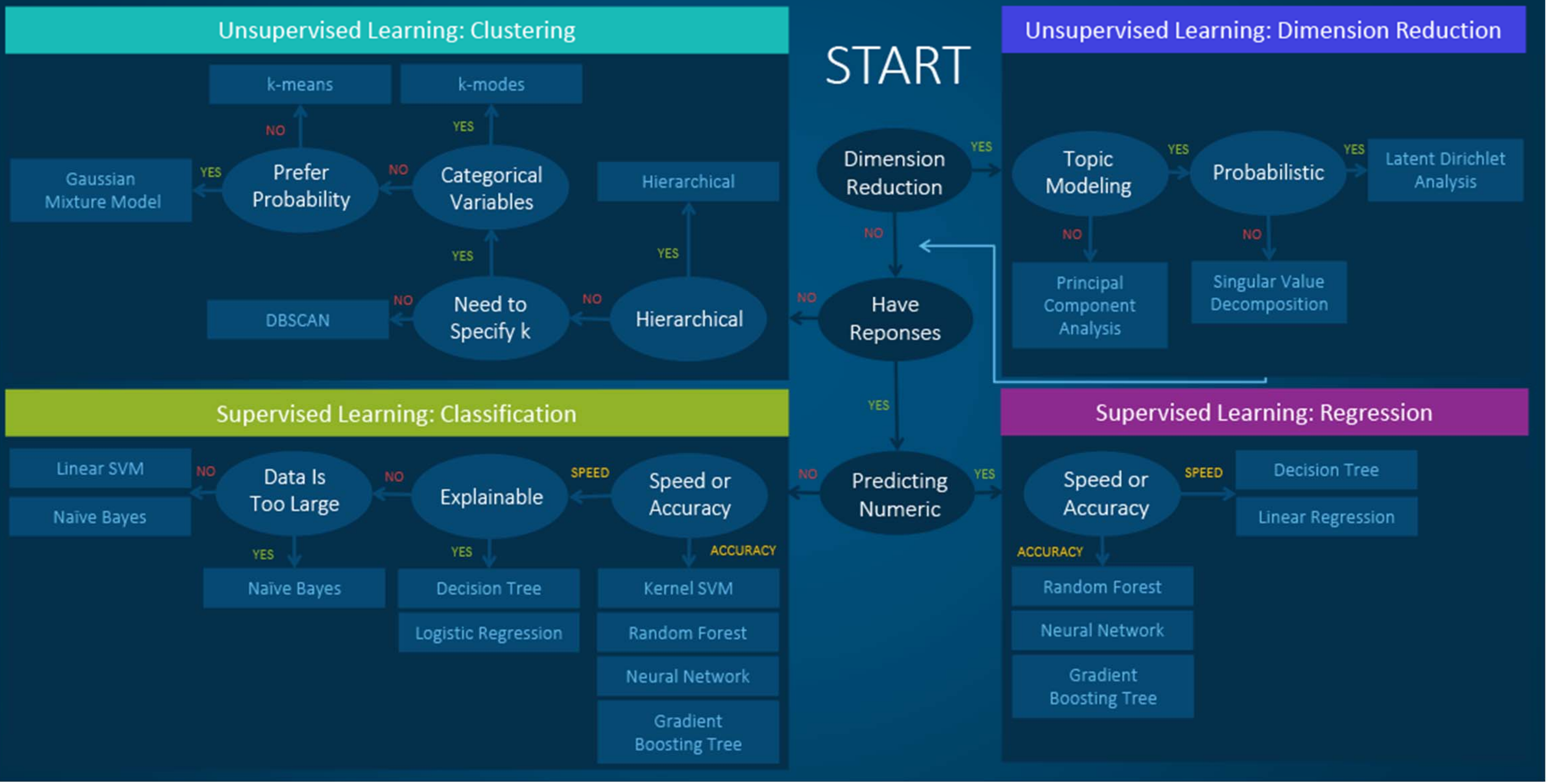
This paper intends to demonstrate the use of machine learning methods as Decision Trees, Random Forests, Logistic Regression,

community represented by the data, th tics that distinguish them. This can be machine learning methods.

II. Related Wo

Machine learning algorithms are wi for example, Cornel and Mirela us

Machine Learning Algorithms Cheat Sheet



START

Dimension Reduction

YES

NO

Have Reponses

NO

YES

Predicting Numeric

NO

YES

Unsupervised Learning: Clustering

k-means

k-modes

NO

YES

Gaussian Mixture Model

YES

Prefer Probability

NO

Categorical Variables

Hierarchical

YES

YES

DBSCAN

NO

Need to Specify k

NO

Hierarchical

Unsupervised Learning: Dimension Reduction

Topic Modeling

YES

Probabilistic

YES

Latent Dirichlet Analysis

NO

NO

Principal Component Analysis

Singular Value Decomposition

Supervised Learning: Classification

Linear SVM

Naïve Bayes

NO

Data Is Too Large

NO

Explainable

SPEED

Speed or Accuracy

YES

YES

ACCURACY

Naïve Bayes

Decision Tree

Logistic Regression

Kernel SVM

Random Forest

Neural Network

Gradient Boosting Tree

Supervised Learning: Regression

Speed or Accuracy

SPEED

Decision Tree

Linear Regression

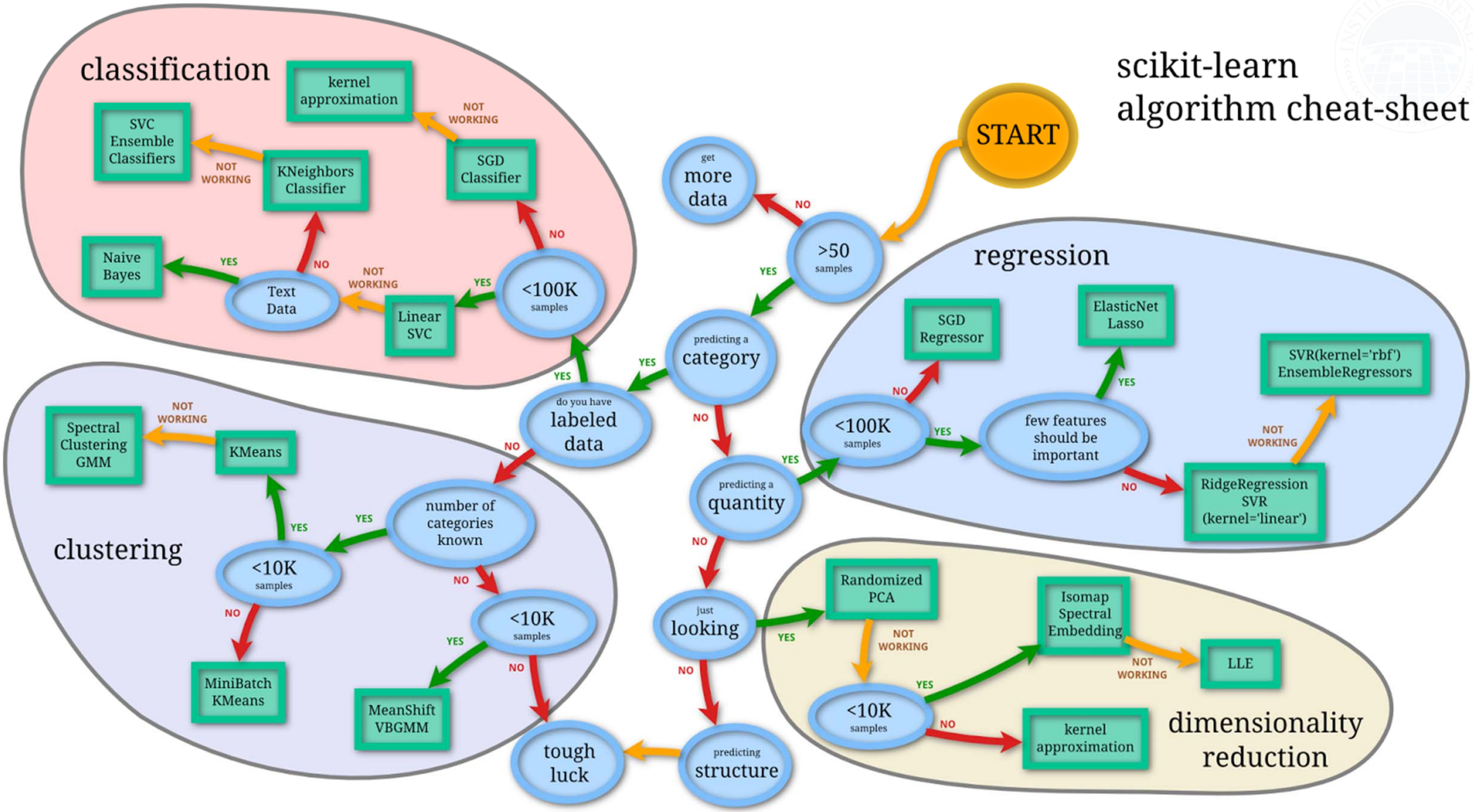
ACCURACY

Random Forest

Neural Network

Gradient Boosting Tree

scikit-learn algorithm cheat-sheet



Objective

- Identify risk communities in a community using machine learning methods
  - Decision Trees
  - Random Forests
  - XGBoost
  - Logistic Regression

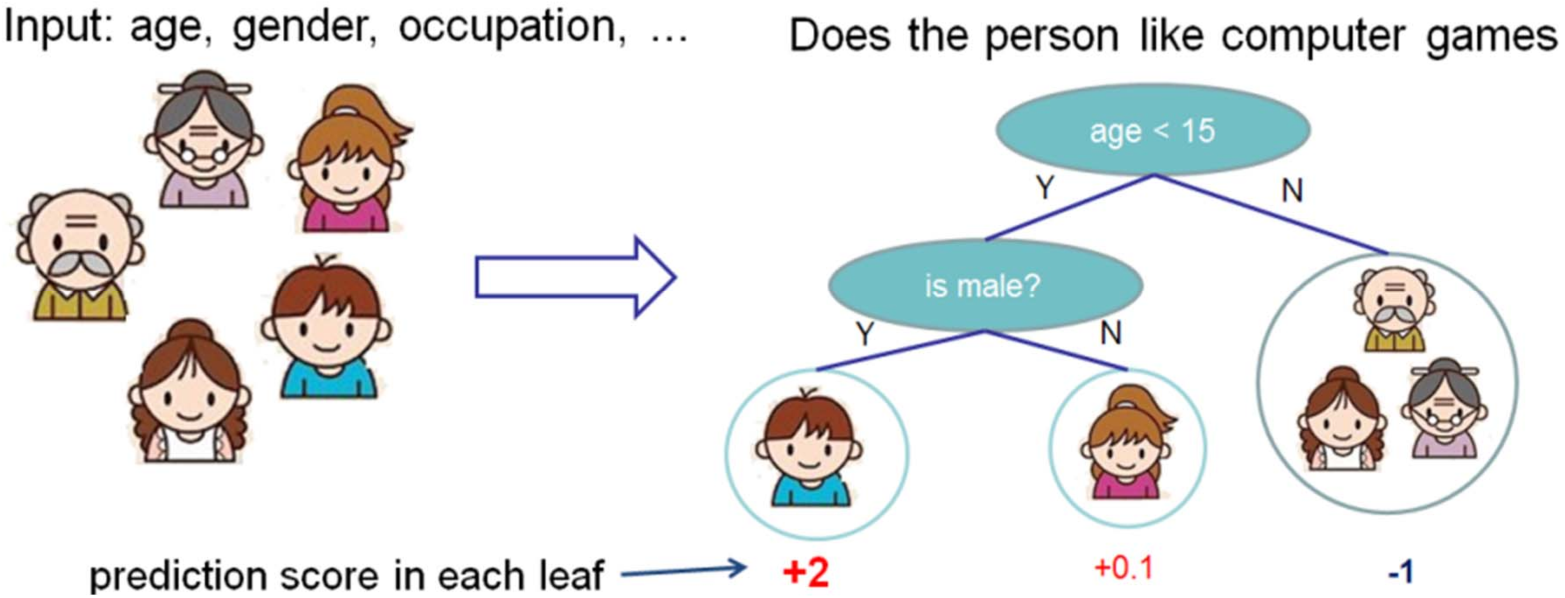
Machine Learning

Basic equation: $Obj(\theta) = L(\theta) + \Omega(\theta)$

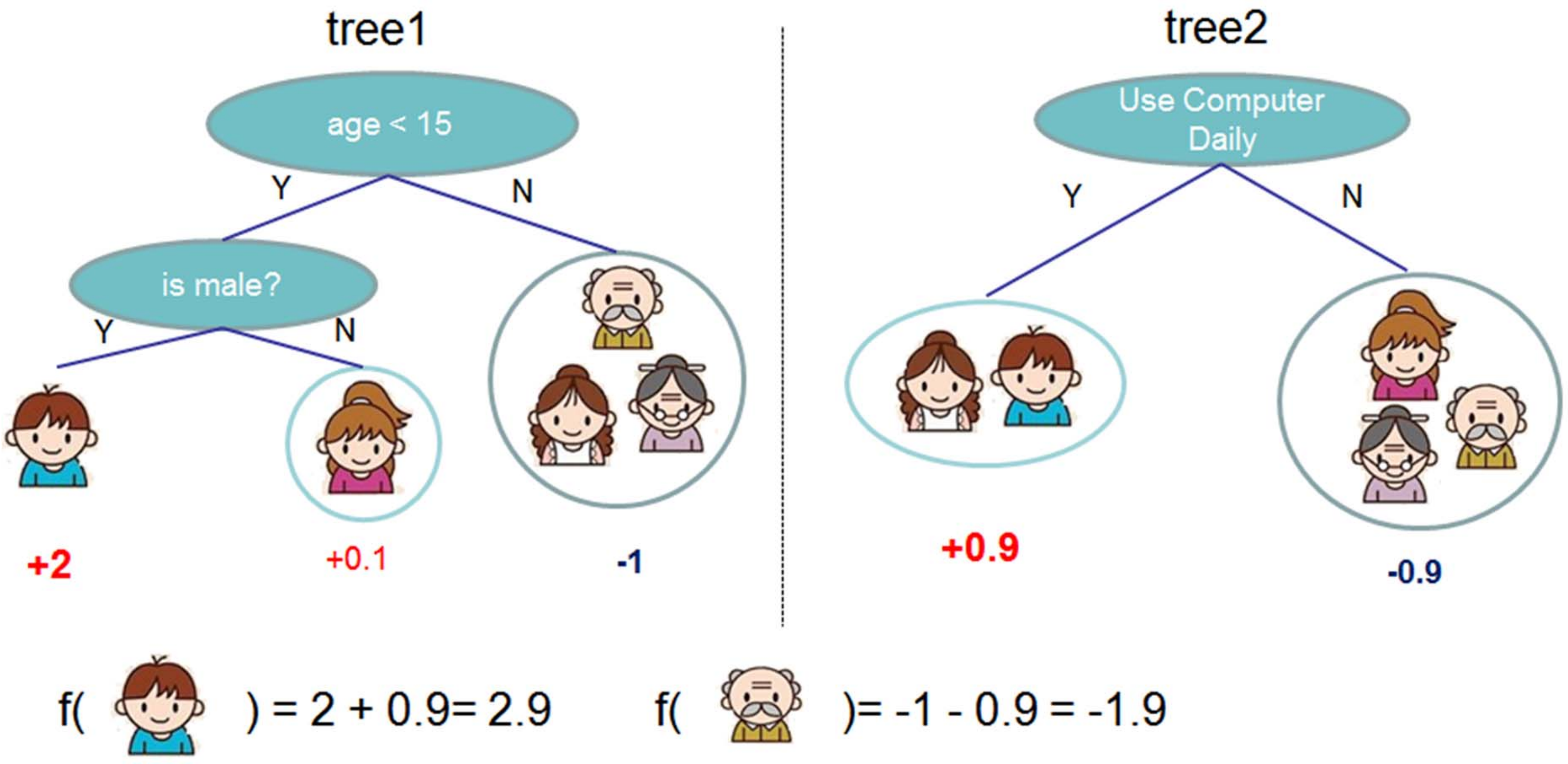
- L – Training Loss
- Ω – Regularization

Ex: Linear regression, Logistic Regression, Decision Trees

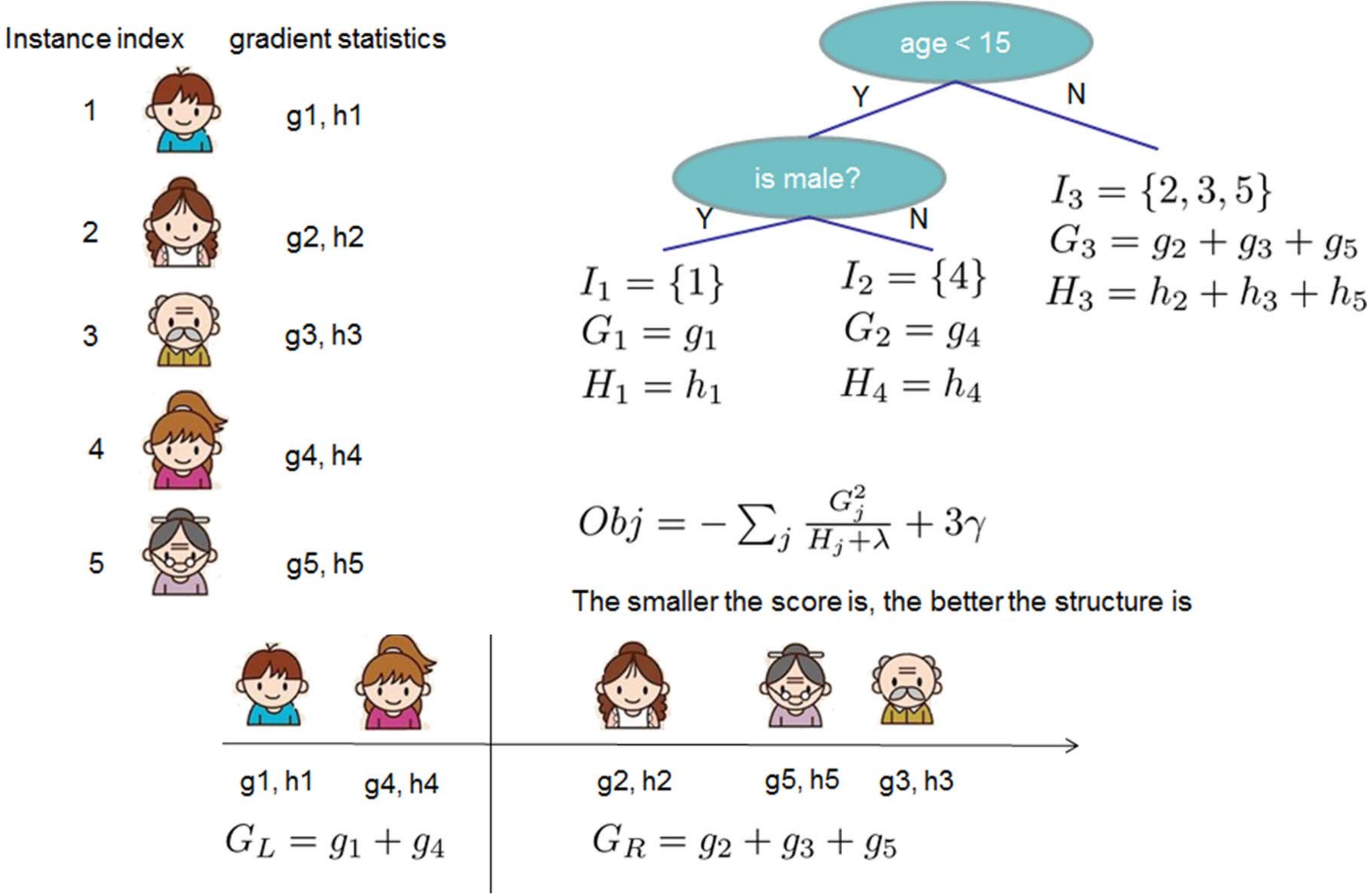
Decision Trees


Input: age, gender, occupation, ...
Does the person like computer games
age < 15
Y
N
is male?
Y
N
prediction score in each leaf
+2
+0.1
-1

Random Forests



Gradient Boosting (XGBoost)



Instance index	gradient statistics
1	g1, h1
2	g2, h2
3	g3, h3
4	g4, h4
5	g5, h5

$I_1 = \{1\}$
$G_1 = g_1$
$H_1 = h_1$

$I_2 = \{4\}$
$G_2 = g_4$
$H_4 = h_4$

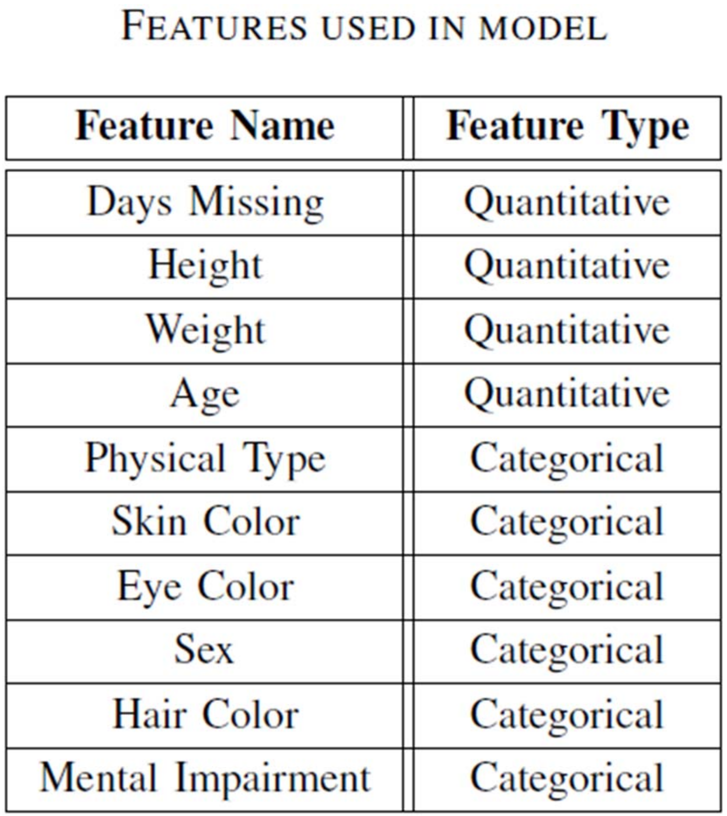
$I_3 = \{2, 3, 5\}$
$G_3 = g_2 + g_3 + g_5$
$H_3 = h_2 + h_3 + h_5$

$$Obj = -\sum_j \frac{G_j^2}{H_j + \lambda} + 3\gamma$$

The smaller the score is, the better the structure is

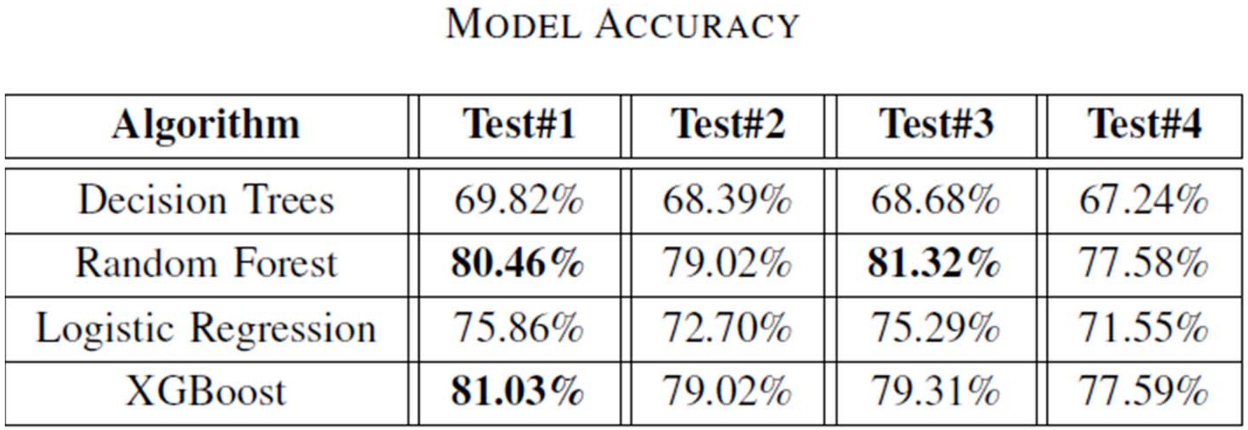
g1, h1 g4, h4 | g2, h2 g5, h5 g3, h3

$G_L = g_1 + g_4$ | $G_R = g_2 + g_3 + g_5$



Features used in model

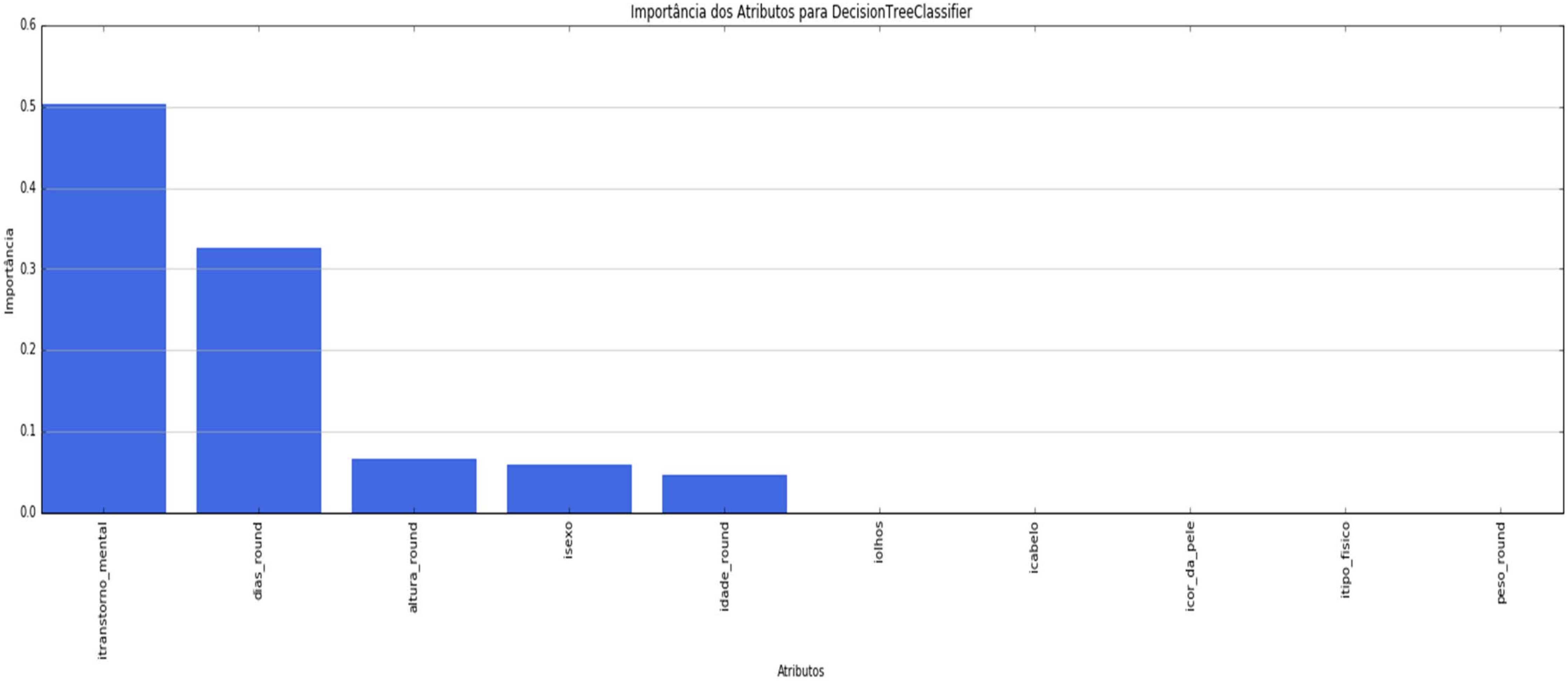
Feature Name	**Feature Type**
Days Missing	Quantitative
Height	Quantitative
Weight	Quantitative
Age	Quantitative
Physical Type	Categorical
Skin Color	Categorical
Eye Color	Categorical
Sex	Categorical
Hair Color	Categorical
Mental Impairment	Categorical



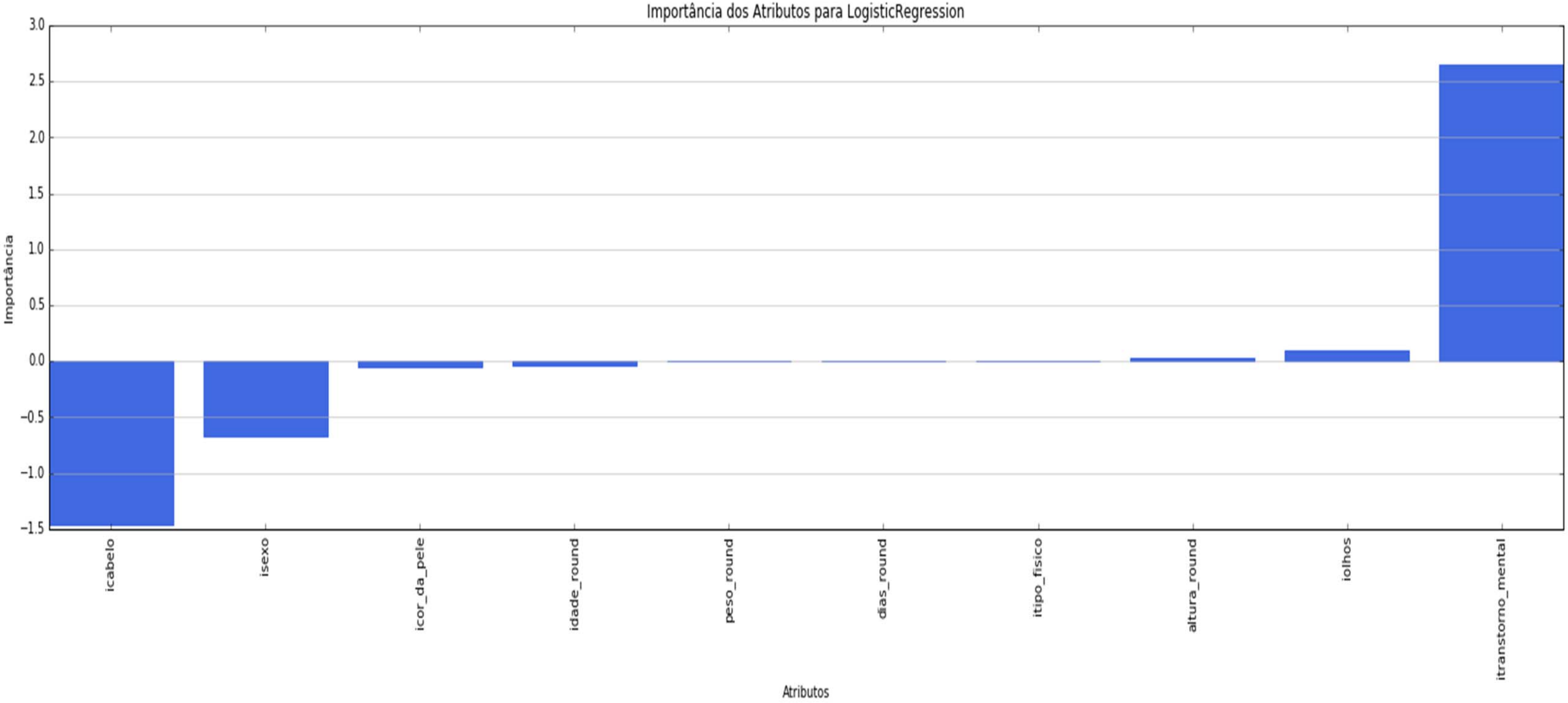
Model Accuracy

Algorithm	**Test#1**	**Test#2**	**Test#3**	**Test#4**
Decision Trees	69.82%	68.39%	68.68%	67.24%
Random Forest	**80.46%**	79.02%	**81.32%**	77.58%
Logistic Regression	75.86%	72.70%	75.29%	71.55%
XGBoost	**81.03%**	79.02%	79.31%	77.59%

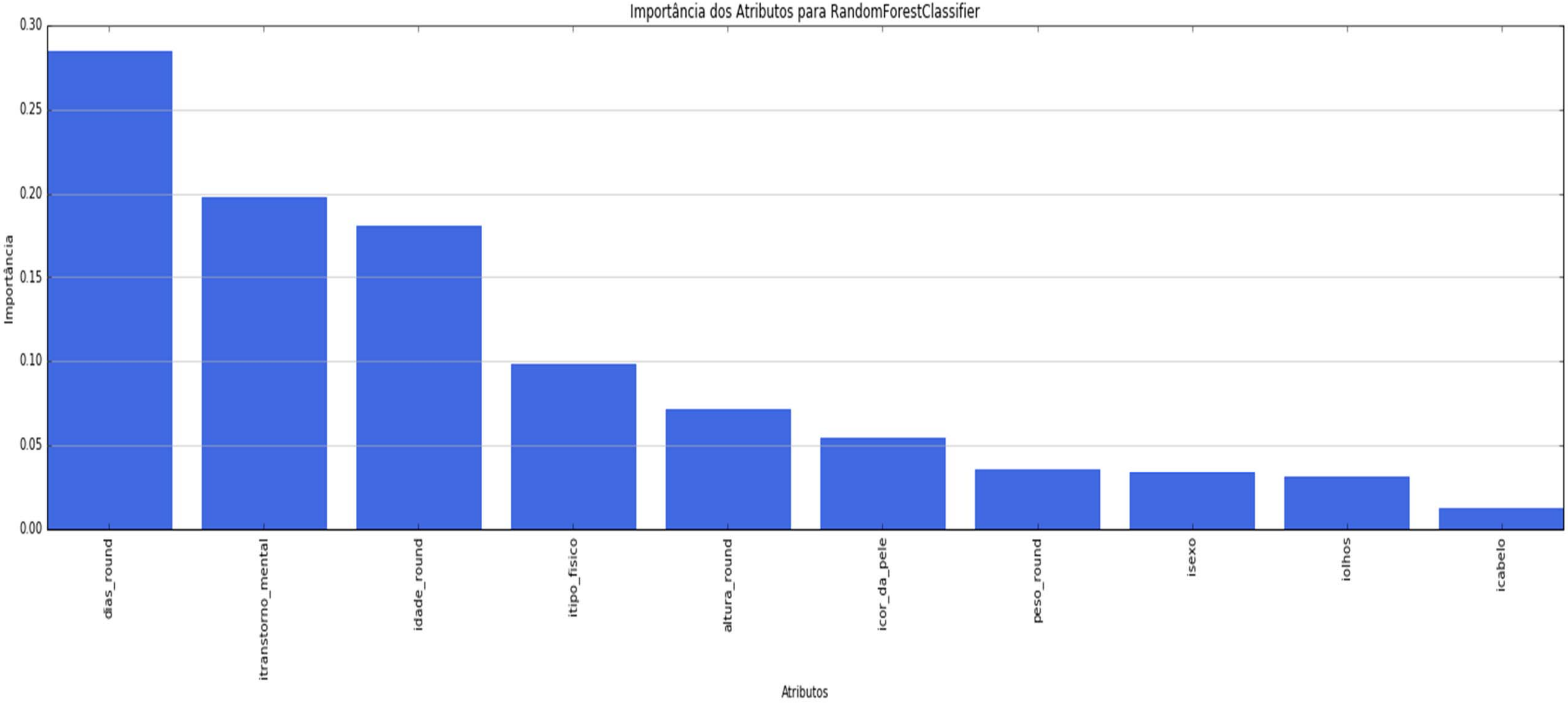
Decision Tree Classifier 69%



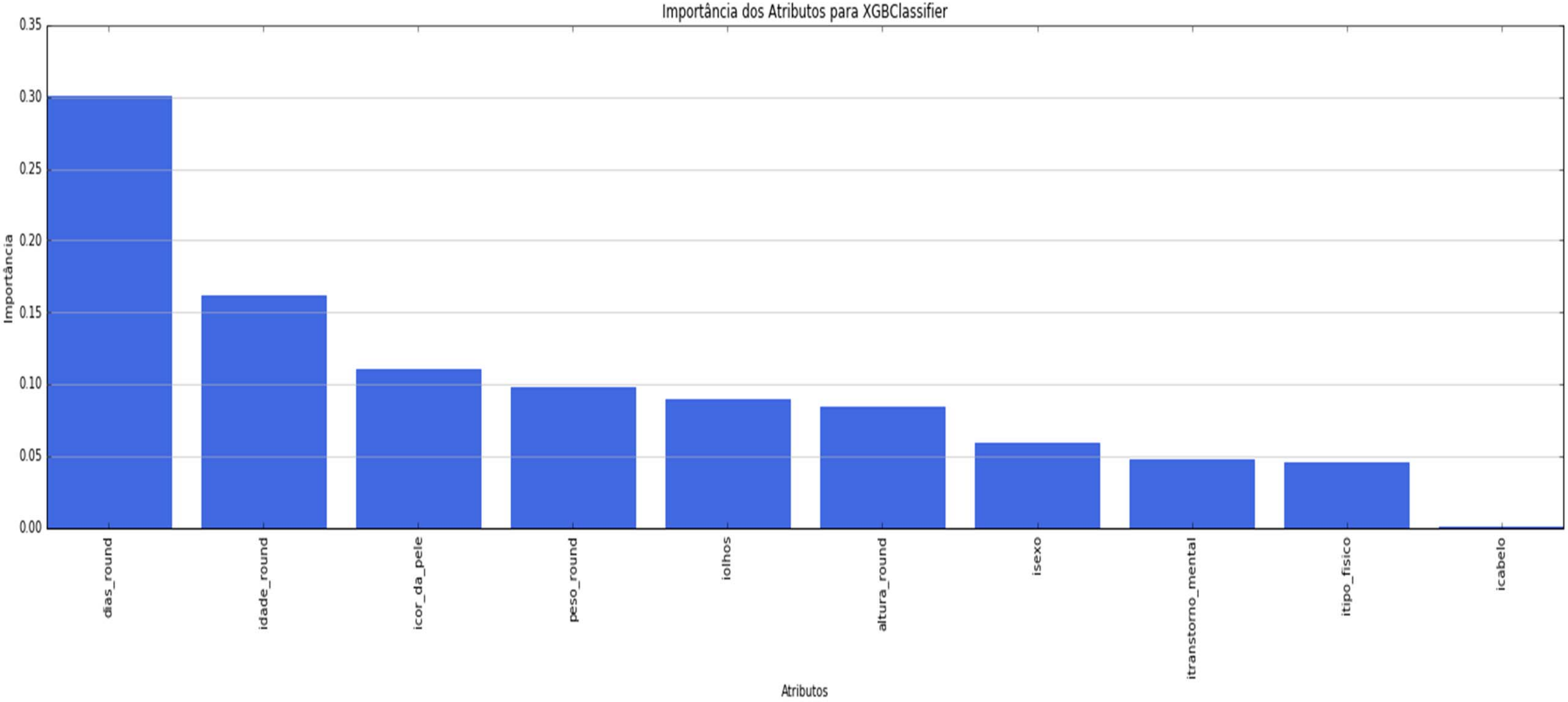
LogisticRegression 73%

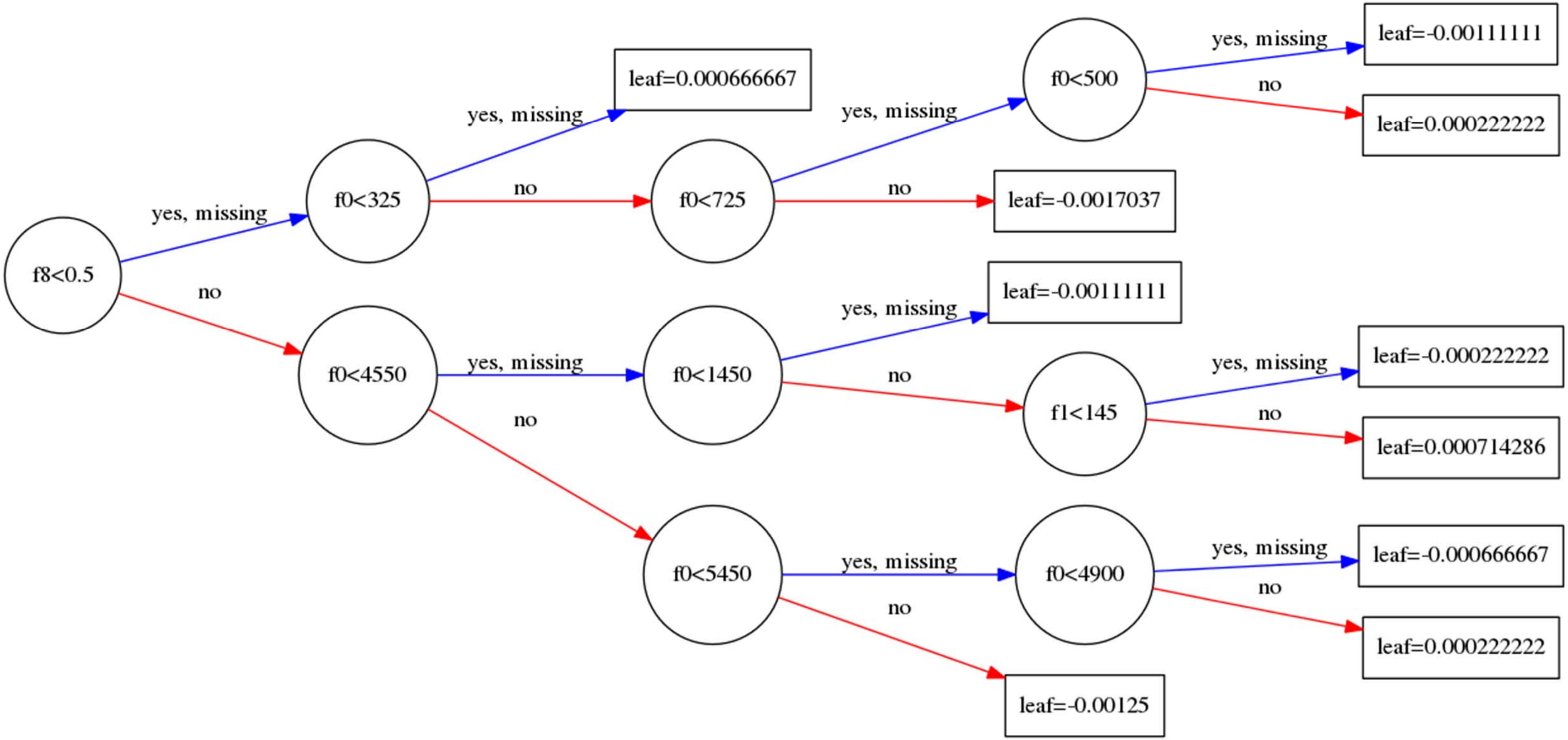


RandomForestClassifier 77%

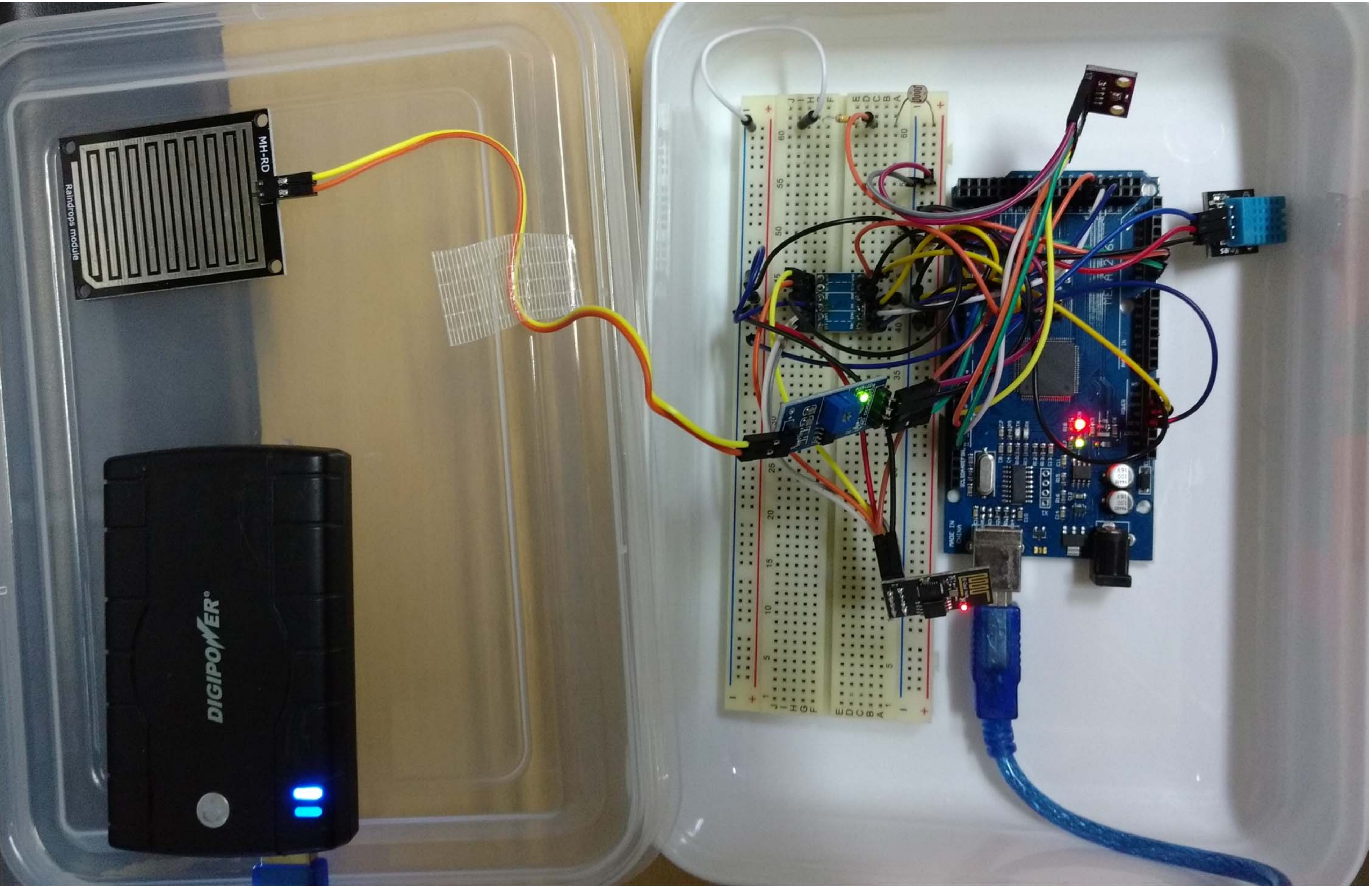


XGBClassifier 79%




f8<0.5
yes, missing
no
f0<325
yes, missing
no
leaf=0.000666667
f0<725
yes, missing
no
f0<500
yes, missing
no
leaf=-0.00111111
leaf=0.000222222
leaf=-0.0017037
f0<4550
yes, missing
no
f0<1450
yes, missing
no
leaf=-0.00111111
f1<145
yes, missing
no
leaf=-0.000222222
leaf=0.000714286
f0<5450
yes, missing
no
f0<4900
yes, missing
no
leaf=-0.000666667
leaf=0.000222222
leaf=-0.00125

OTHER PROJECTS


MH-RD
Raindrops module
DIGIPOWER®
MADE IN CHINA

QUESTIONS?