NOVEMBER 9, 2019

COSTA RICAN HOUSEHOLD POVERTY LEVEL PREDICTION

Presented by An VO QUANG

COSTA RICA IS THE MOST-VISITED NATION IN THE CENTRAL AMERICAN REGION

- Total area: 51,100 km2
- Population (2018): 4,900,000 +
- 2.9 million foreign visitors in 2016, + 10% in 2015
- Tourism sector is responsible for 5.8% of Costa Rica's GDP, or $3.4 billion (2015)

© Sindicato de Trabajadores de MINAE

COSTA RICA HAS ONE OF THE HIGHEST STANDARDS OF LIVING IN CENTRAL AMERICA

- Human Development Index (HDI): 0.794.

- High quality health care is provided by the government at low cost to the users.

- Because of its educational system, Costa Rica has one of the highest literacy rates in Latin America (97%)

However...

1.1 MILLION

people currently live in poverty in Costa Rica

$155 / MONTH

20% of the population live below this national poverty line

QUESTION IS: CAN WE PREDICT POVERTY LEVELS BASED ON HOUSING DATA?

A supervized machine learning project

142 FEATURES

Among them:

- # of persons living in the household,
- # of children, males, females
- monthly rent payment
- urban area / rural area
- no level of education / # of years of education
- married / divorced / separated
- materials used for house building, floor, wall
- water provision yes / no
- if disable person in household
- etc.

RESPONSE VARIABLE: POVERTY LEVEL

1

EXTREME POVERTY

2

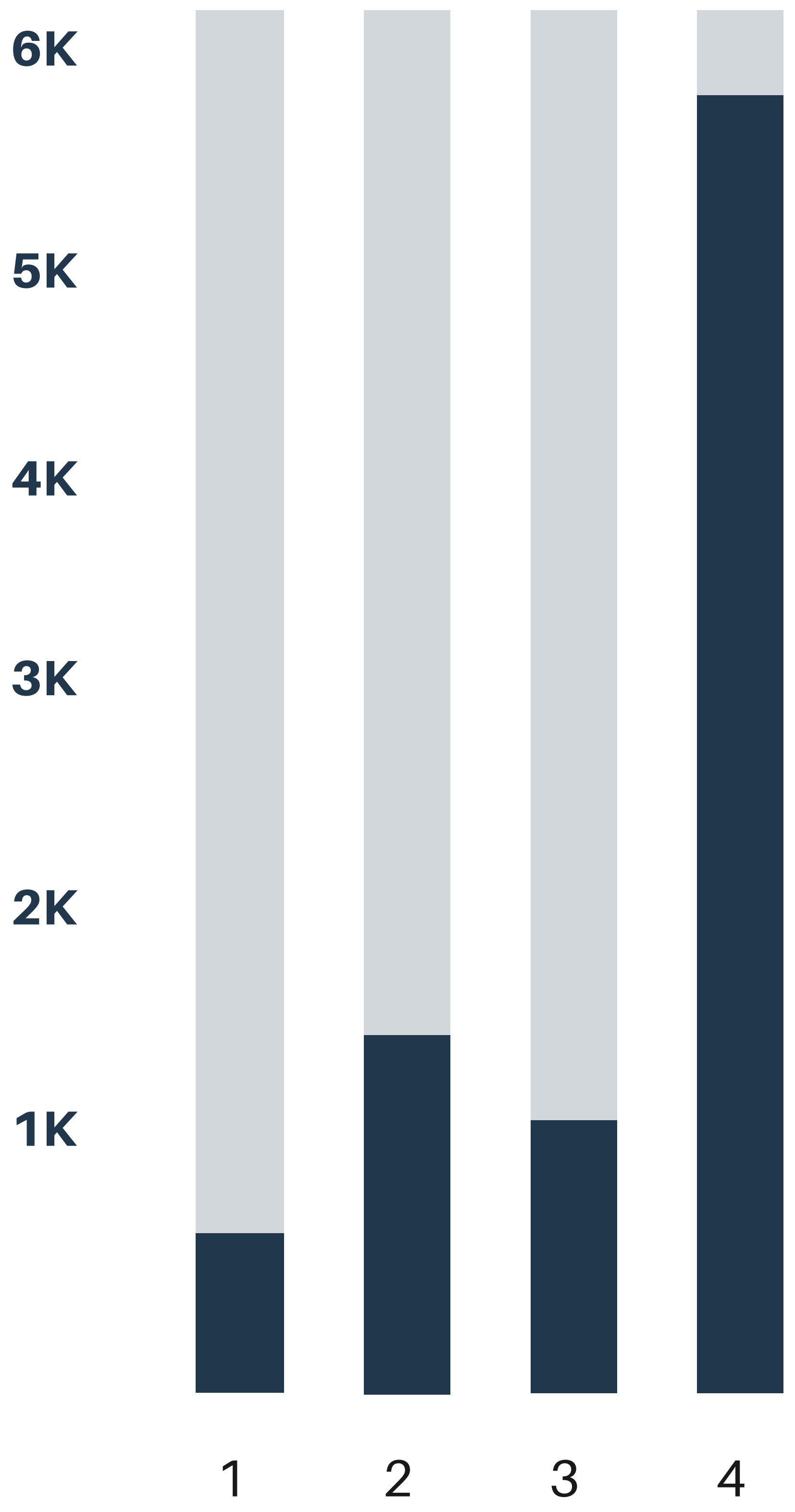
MODERATE POVERTY

3

VULNERABLE HOUSEHOLDS

4

NON VULNERABLE HOUSEHOLDS



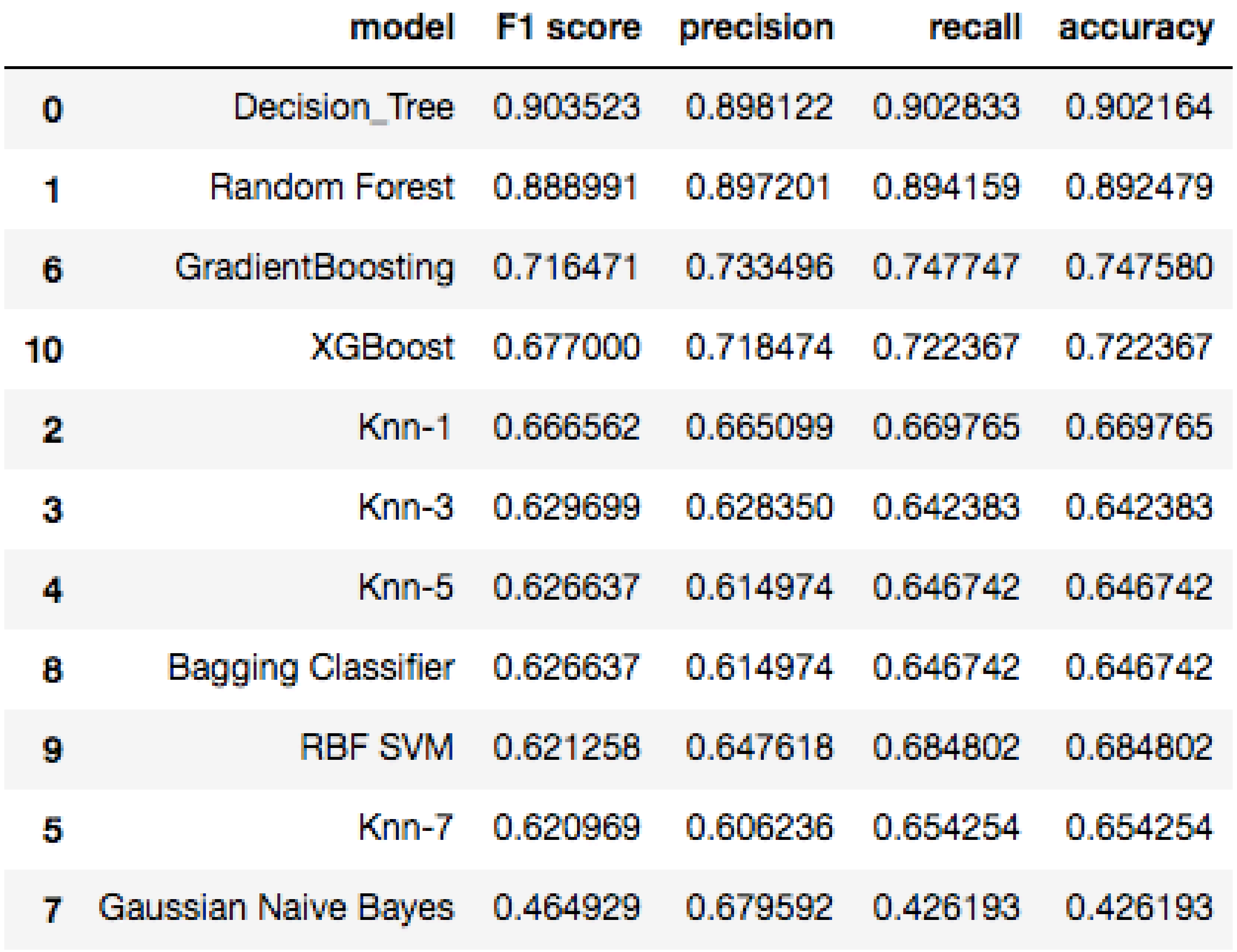
LEVEL DISTRIBUTION

FLOWCHART

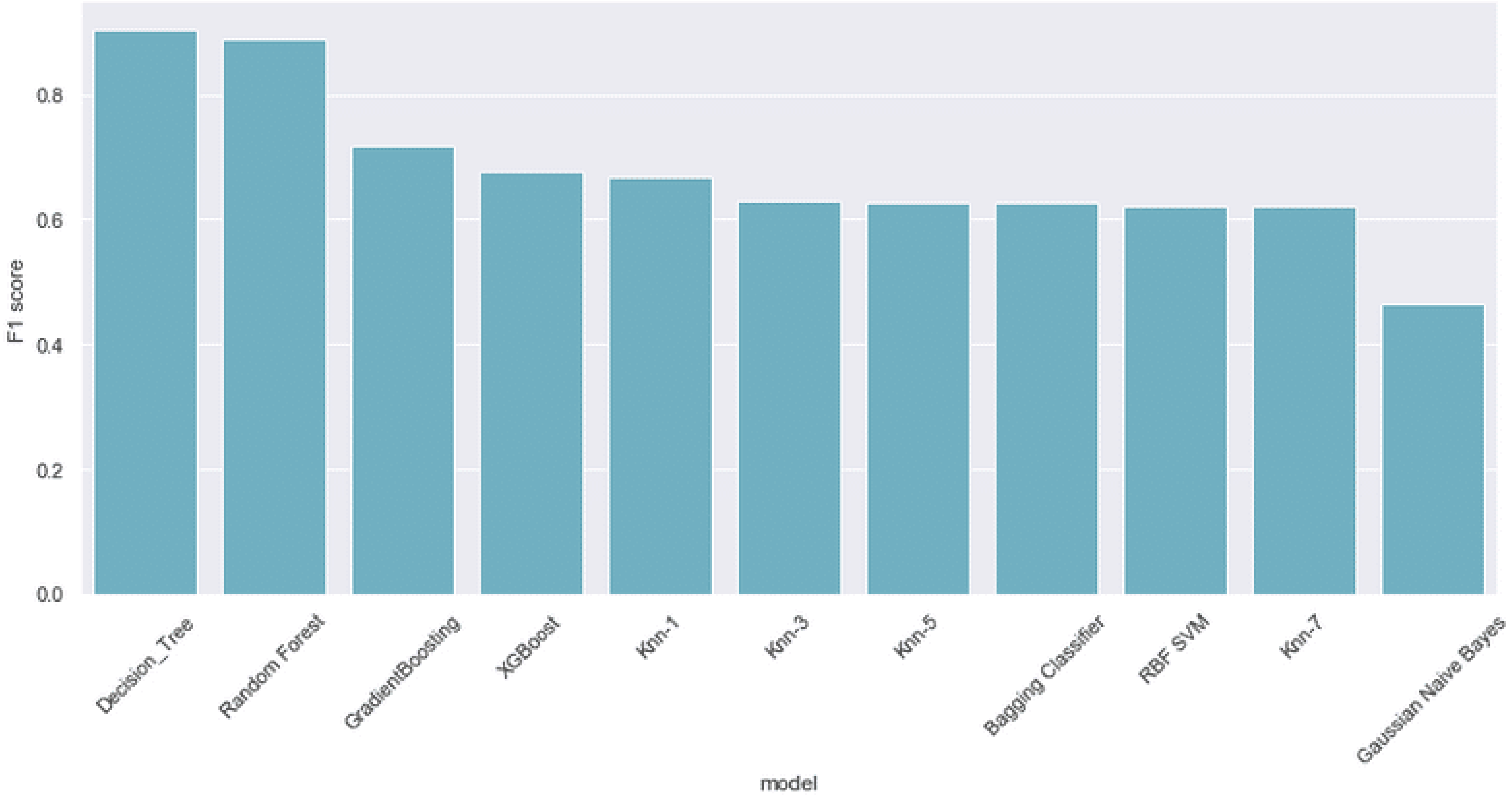
01	**Splitting TRAIN data**	Discard TEST data (there's no Target column in it) Re-splitting TRAIN data into New_train and New_test
02	***Data cleaning***	Removing columns with > 70% NAs Cleaning Object columns Removing columns with high multicollinearity > 0.9
03	***Model testing***	Comparing 8 ML models (+ cross-validation) 4 metrics: F1 scores, Precision, Recall, Accuracy
04	***GridSearch***	Hyperparameters tuning on selected model(s)
05	***Prediction***	Prediction on New_test data

MODEL COMPARISON AND MODEL SELECTION

CROSS VALIDATION = 10



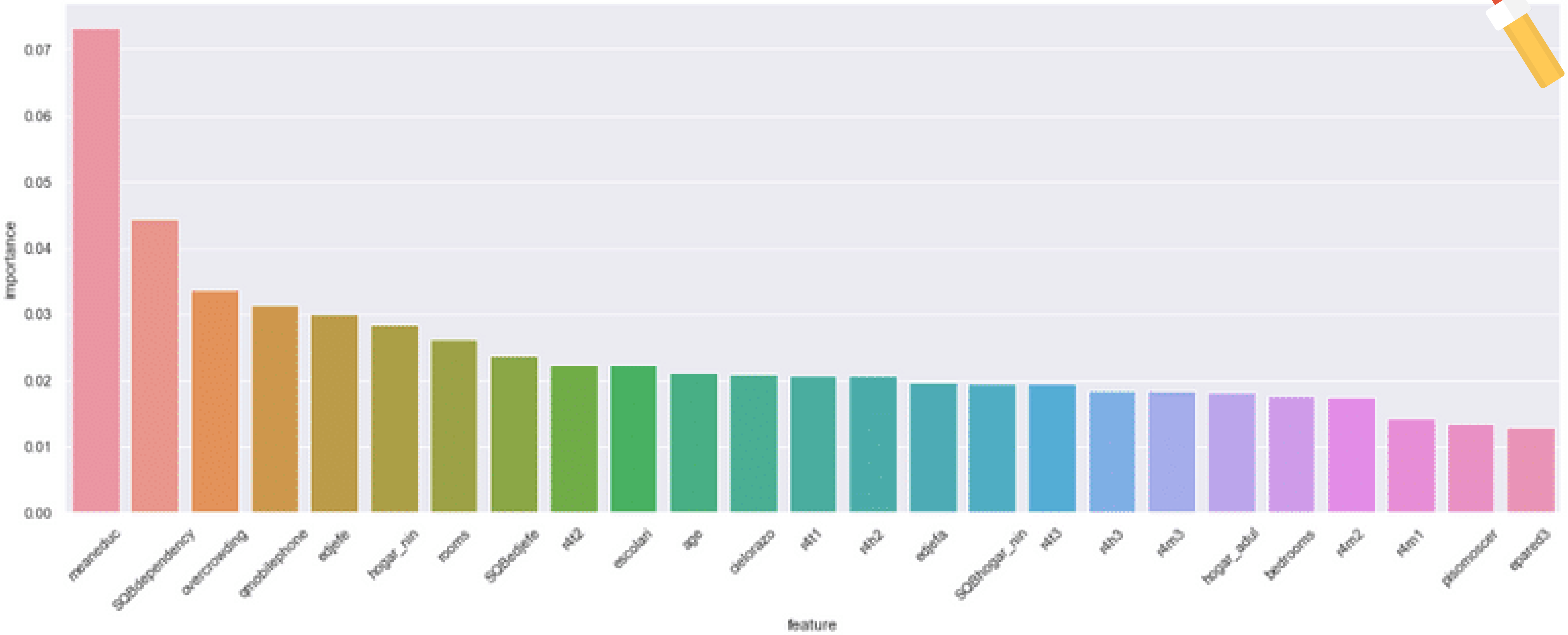
	model	F1 score	precision	recall	accuracy
0	Decision_Tree	0.903523	0.898122	0.902833	0.902164
1	Random Forest	0.888991	0.897201	0.894159	0.892479
6	GradientBoosting	0.716471	0.733496	0.747747	0.747580
10	XGBoost	0.677000	0.718474	0.722367	0.722367
2	Knn-1	0.666562	0.665099	0.669765	0.669765
3	Knn-3	0.629699	0.628350	0.642383	0.642383
4	Knn-5	0.626637	0.614974	0.646742	0.646742
8	Bagging Classifier	0.626637	0.614974	0.646742	0.646742
9	RBF SVM	0.621258	0.647618	0.684802	0.684802
5	Knn-7	0.620969	0.606236	0.654254	0.654254
7	Gaussian Naive Bayes	0.464929	0.679592	0.426193	0.426193



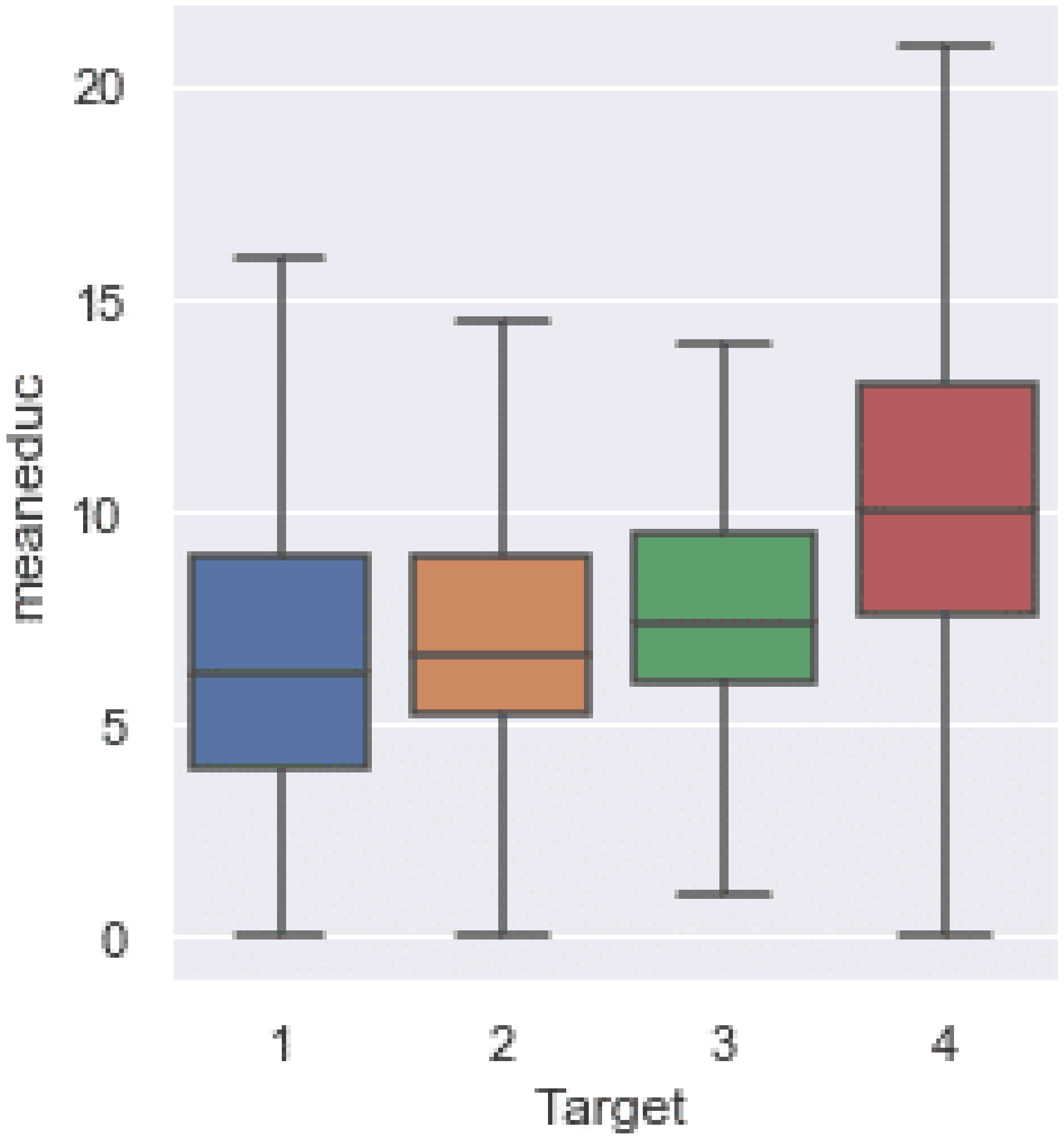
MODEL COMPARISON AND MODEL SELECTION

CROSS VALIDATION = 10

FEATURE IMPORTANCES



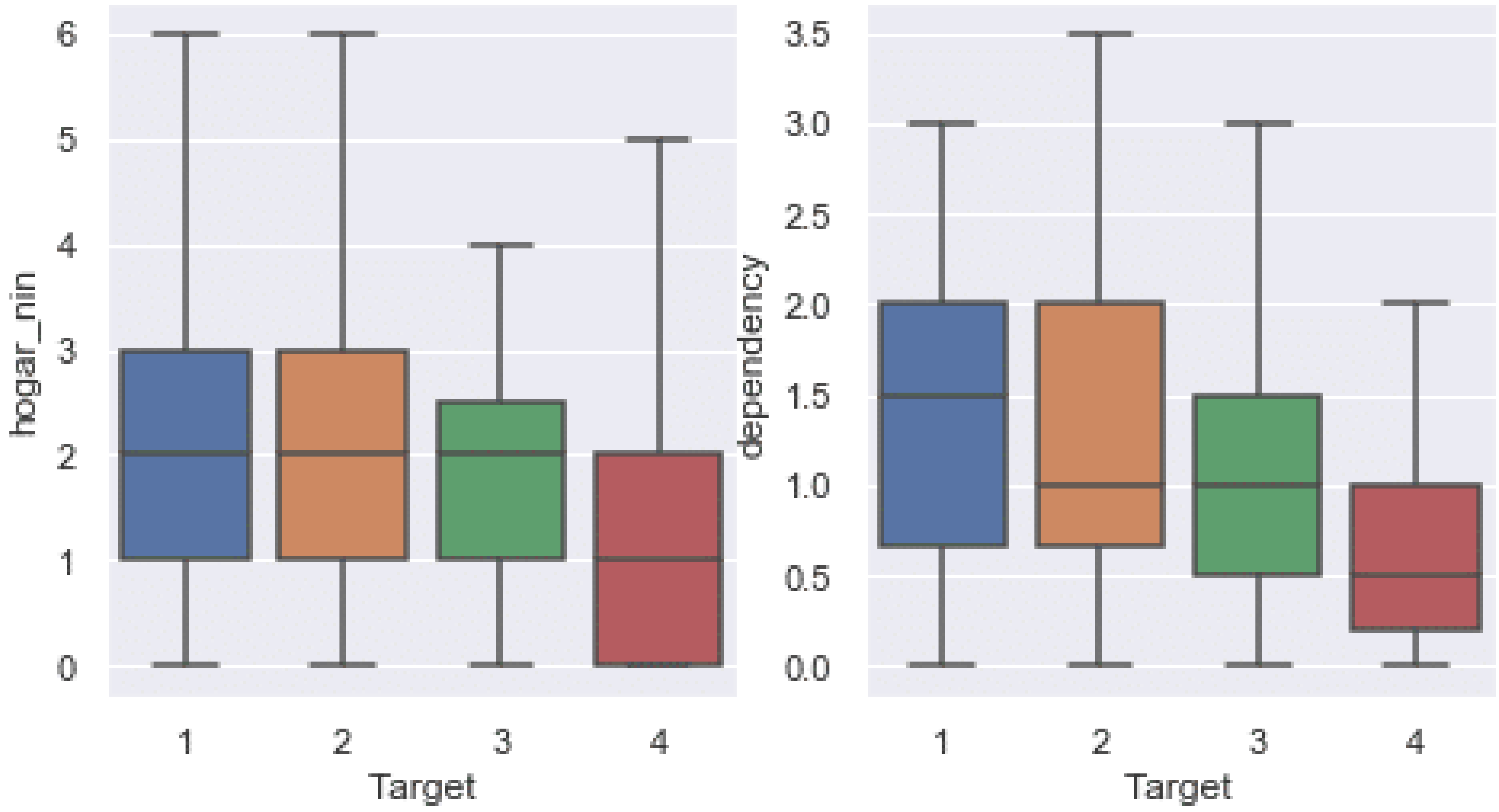
FEATURE IMPORTANCES



- Average years of education in a household is the strongest variable when predicting level of poverty

- Looking at this variable alone, average years of education is higher in non-vulnerable households (lever 4)

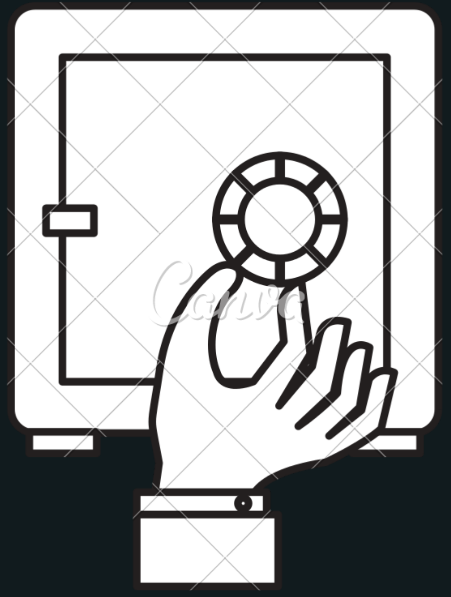
FEATURE IMPORTANCES



- Number of children in household is lower in non-vulnerable households

- Dependency level is calculated from number of seniors and children in household. It is lower in non-vulnerable households

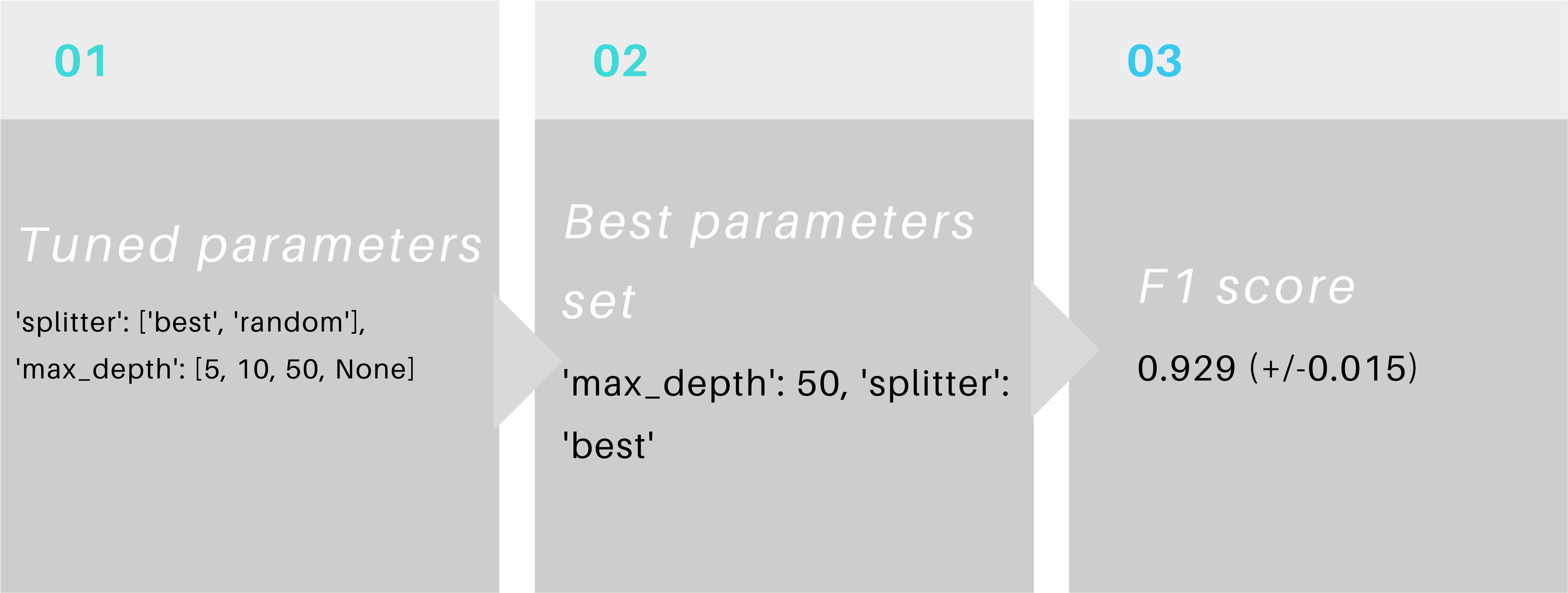
AN IN-DEPTH APPROACH



HYPERPARAMETERS TUNING ON SELECTED MODELS

Decision Tree, Random Forest, Gradient Boosting & XGBoost

GRIDSEARCH CV + DECISION TREE



01

Tuned parameters

'splitter': ['best', 'random'],

'max_depth': [5, 10, 50, None]

02

Best parameters set

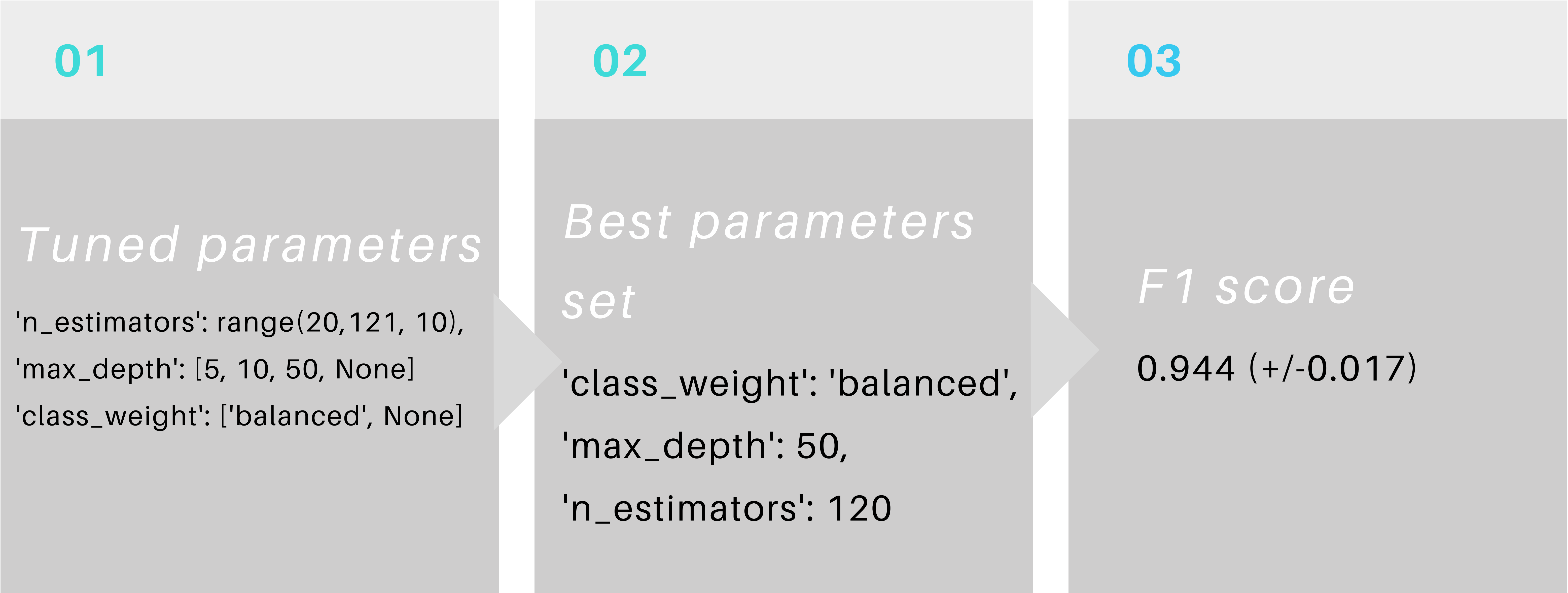
'max_depth': 50, 'splitter': 'best'

03

F1 score

0.929 (+/-0.015)

GRIDSEARCH CV + RANDOM FOREST



01

Tuned parameters

'n_estimators': range(20,121, 10),

'max_depth': [5, 10, 50, None]

'class_weight': ['balanced', None]

02

Best parameters set

'class_weight': 'balanced',

'max_depth': 50,

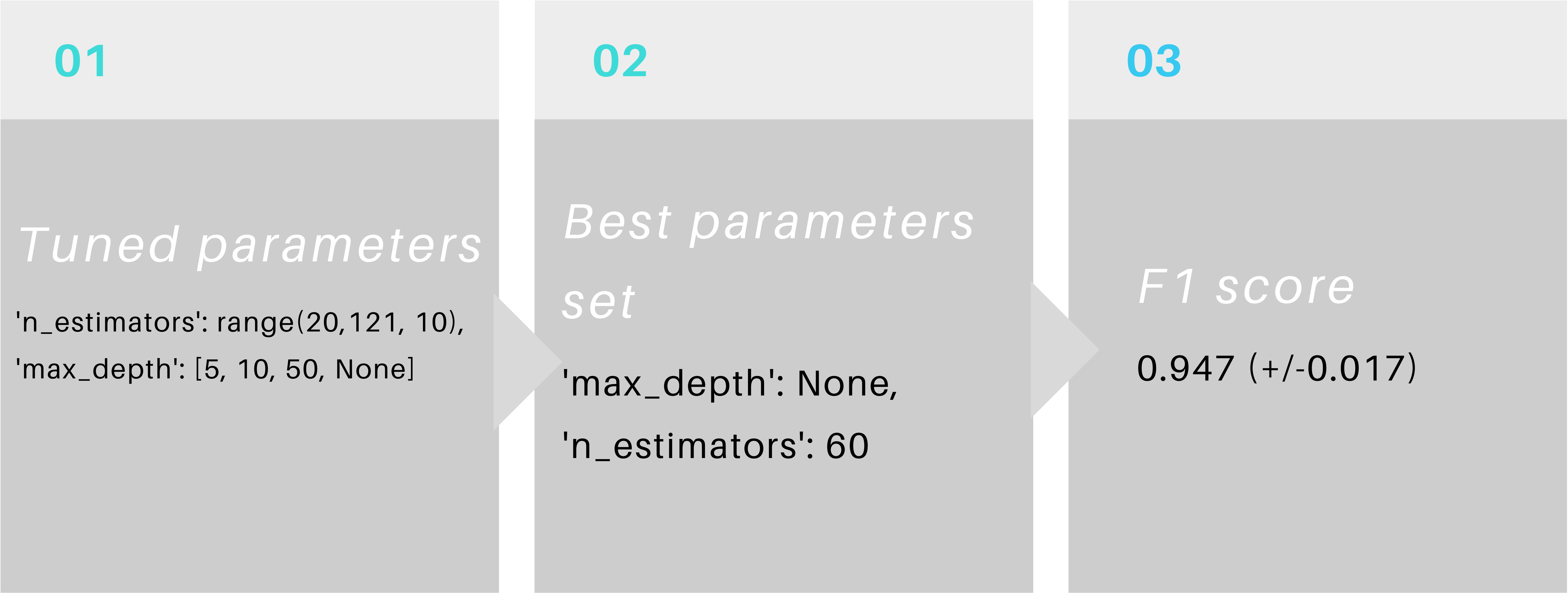
'n_estimators': 120

03

F1 score

0.944 (+/-0.017)

GRIDSEARCH CV + GRADIENT BOOSTING



01

Tuned parameters

'n_estimators': range(20,121, 10),

'max_depth': [5, 10, 50, None]

02

Best parameters set

'max_depth': None,

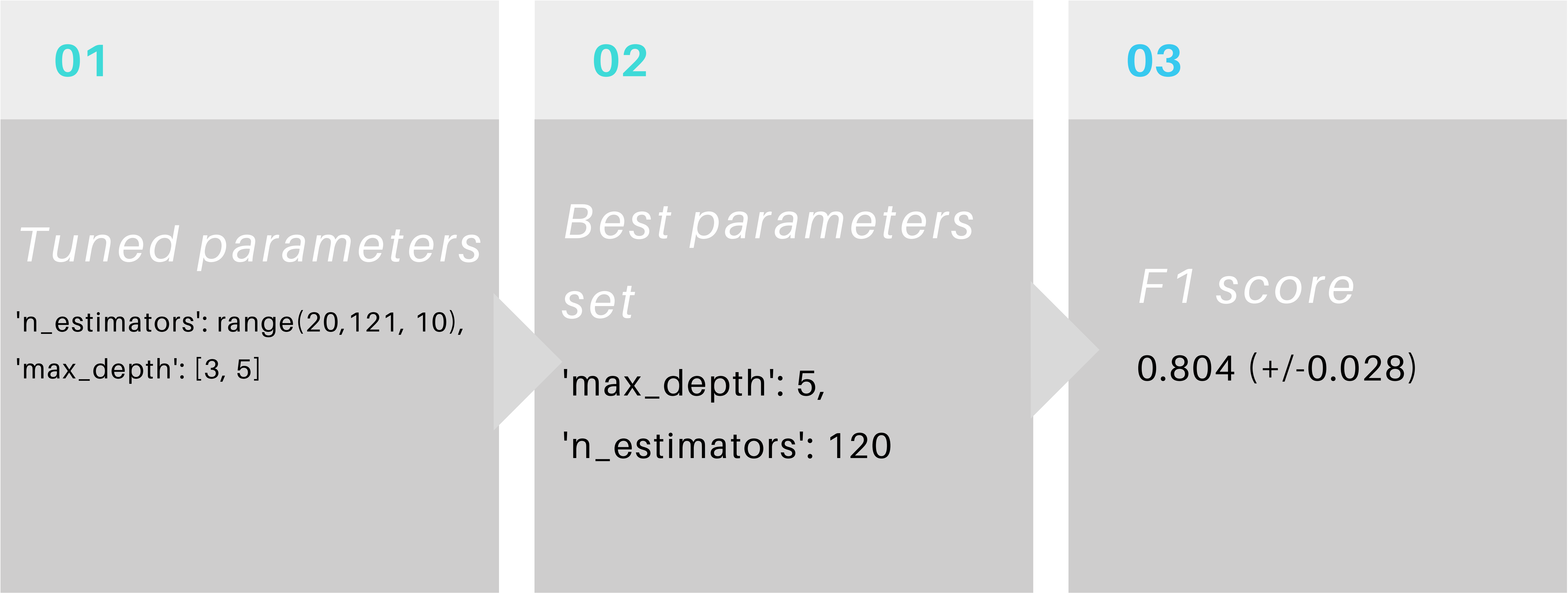
'n_estimators': 60

03

F1 score

0.947 (+/-0.017)

GRIDSEARCH CV + XGBOOST



01

Tuned parameters

'n_estimators': range(20,121, 10),
'max_depth': [3, 5]

02

Best parameters set

'max_depth': 5,
'n_estimators': 120

03

F1 score

0.804 (+/-0.028)

PREDICTION ON VALIDATION DATASET

CLASSIFIERS PERFORMANCE COMPARISON

XGBOOST ON NEW_TEST

with TUNED hyperparameters

Confusion matrix

```
[ 98   4   2   32]
 [  0 186   4   81]
 [  2  18 147   87]
 [  0  15   3 1193]
```

Accuracy score

0.867

F1 score

0.804

Balanced accuracy score

0.742

RANDOM FOREST ON NEW_TEST

with TUNED hyperparameters

Confusion matrix

```
[ 132    0    0    4]
 [   0  267    1    3]
 [   0    3  247    4]
 [   0    1    1 1209]
```

Accuracy score

0.990

F1 score

0.986

Balanced accuracy score

0.981

GRADIENT BOOSTING ON NEW_TEST

with TUNED hyperparameters

Confusion matrix

```
[ 130    0    1    5]
 [   0  268    1    2]
 [   1    1  251    1]
 [   0    0    5 1206]
```

Accuracy score

0.991

F1 score ✓

0.985

Balanced accuracy score

0.982

CONCLUSION

Machine Learning predicts accurately level of poverty in 99% of cases

Best suggested model

- Gradient Boosting
- 'max_depth': None, 'n_estimators': 70

Future improvements

- Linear regression + Thresholding
- PCA to reduce nb of features

CONTACT INFORMATION

Email Address

an.voquang@gmail.com

GitHub

github.com/Peau-Rouge