# ML for Diabetes Prevention

Vittorio Haardt & Luca Porcelli

# 01

# Project Goals

Effective diagnosis of diabetes can save lives

# Make it Real

The goal was to produce a solution that could actually be useful for the diabetic diagnosis

| | |
|---|---|
| **A performing model** | Achieved by an intensive and a careful ML procedure |
| **An usable solution** | Achieved by a deployment phased aimed to usability |
| **A really useful tool** | Achieved by an App build for generic users |

**The best technology is the one that everyone can use, without even thinking about it.**
**- Clayton Christensen**

# Sub-Goals

## Training

Deployment of the right ML strategy in order to reach the optimal solution

## Deployment

Deployment of the best model in order to make it useful and usable

## DataApp

Deployment of the application in order to make the solution real


# 02

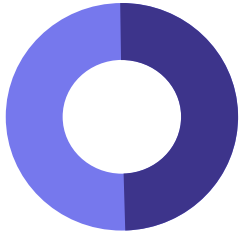
# Data

Data as the foundation of ML solutions

# The Right Data for the Right Model

**Data:** 40,108 observation from a bigger dataset

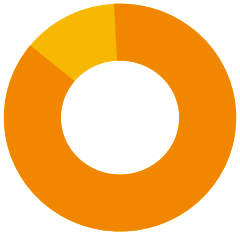
$y$ — **Diabetes: yes/no**



**51/49**

Balanced class ideal for machine learining

$x_i$ — **Input variables about life habits**



**5/17**

5 integer variables

**12/17**

12 string variables

# Pre-Processing

## Missing

There was no missing values

## Near-Zero Var

There wasn't any near zero variance

## Correlation

There wasn't redundant features

## Outlier Removal

Outlier removal based on *BMI*

## Variable Transformation

*BMI* from continuous in to categorical

# BMI Pre-Processing

## Outliers Removal

We removed them, as suggested by domain experts

- $BMI < 15$
- $BMI > 50$

Only from the training set, in order not to distort performance

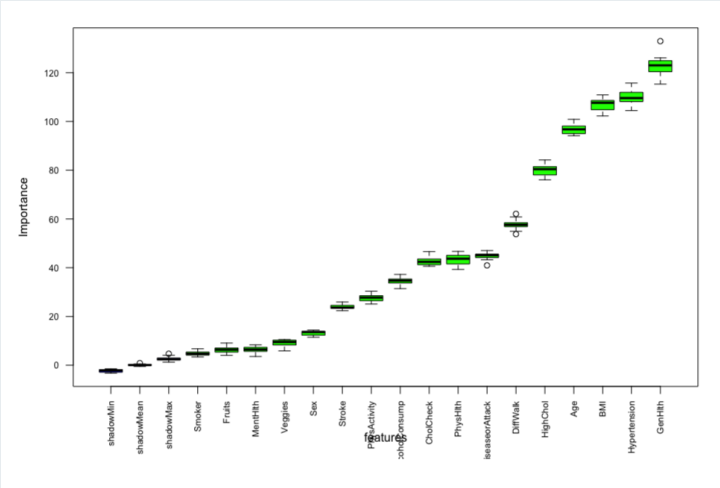
## Variable Transformation

On suggestions from domain experts.

- $BMI \leqslant 16$ → INANITION
- $16 < BMI \leqslant 17.50$ → UNDERWEIGHT
- $17.5 < BMI \leqslant 18.5$ → SLIGHTLY UNDERWEIGHT
- $18.5 < BMI \leqslant 25$ → NORMAL
- $25 < BMI \leqslant 30$ → OVERWEIGHT
- $30 < BMI \leqslant 35$ → CLASS I OBESE
- $35 < BMI \leqslant 40$ → CLASS II OBESE
- $BMI > 40$ → CLASS III OBESE

# Feature Selection

Assessed with Boruta




## All Attribute are Important

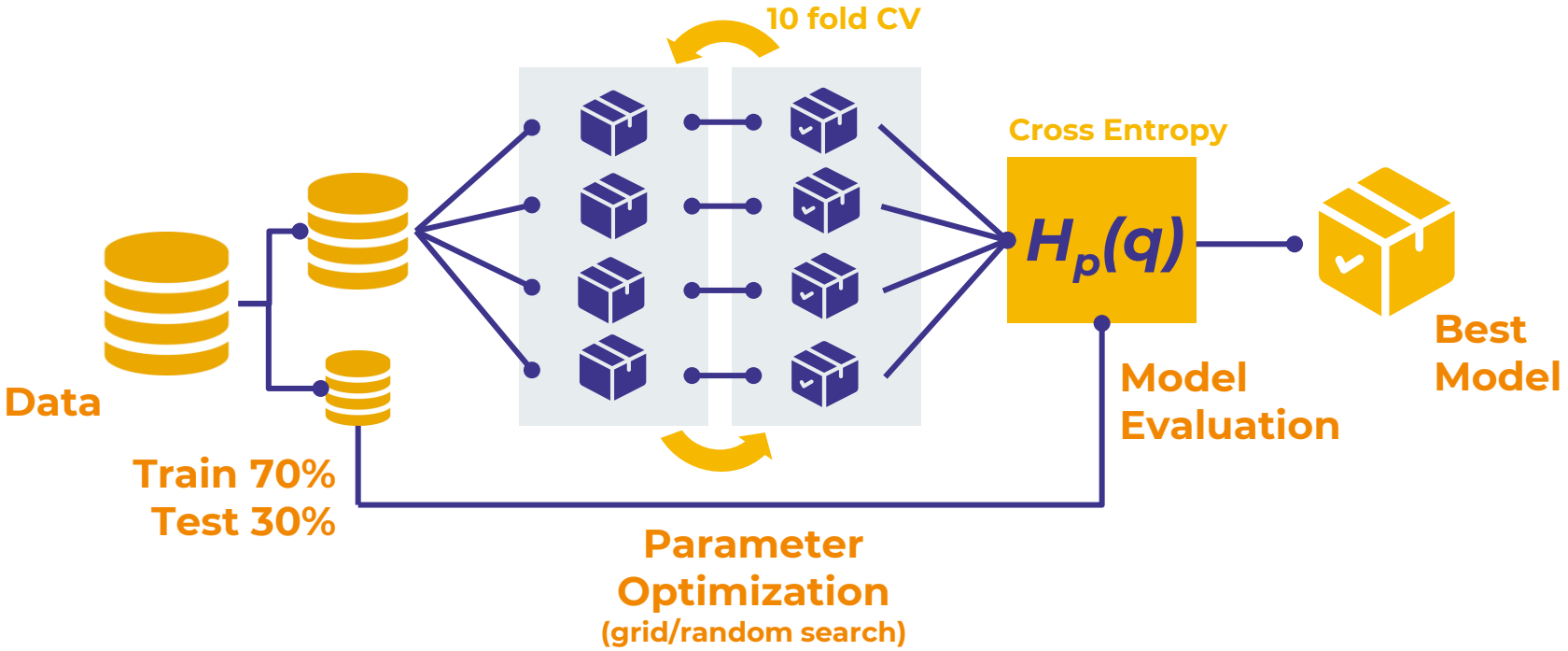
Since no attribute was placed under shadow attributes, we used it all for the models

# 03

# ML Approach

Best performance to perform the best

# ML Pipeline

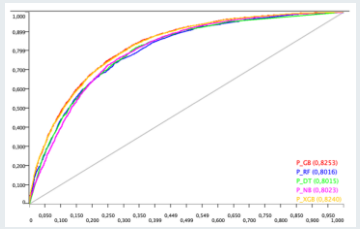

10 fold CV
Cross Entropy
$H_p(q)$
Data
Train 70%
Test 30%
Model
Evaluation
Best
Model
Parameter
Optimization
(grid/random search)

# Models and Performances

## Models

- Random Forest
- Naïve Bayes
- Decision Tree
- Gradient Boosting
- XGBoost
- Staked Model

## Evaluation



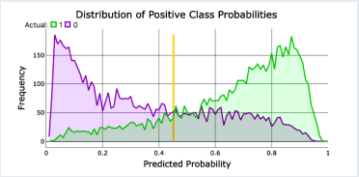
LogLoss

Accuracy

## Best Model

Gradient Boosting

## Threshold




0.451

# Future Development

## Better Model Selection

Specific model selection for models can increase performance

## Different Optimization

Random is not the best option to find the best performance

## Explainability

AI where humans can understand or predictions made by the AI.

# Hands On

Actual use of the developed solution

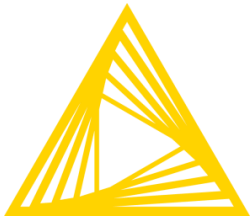
# A Tour on Knime

## Training

Workflow used to show the training phase, and explain the techniques involved in the process

## Deployment

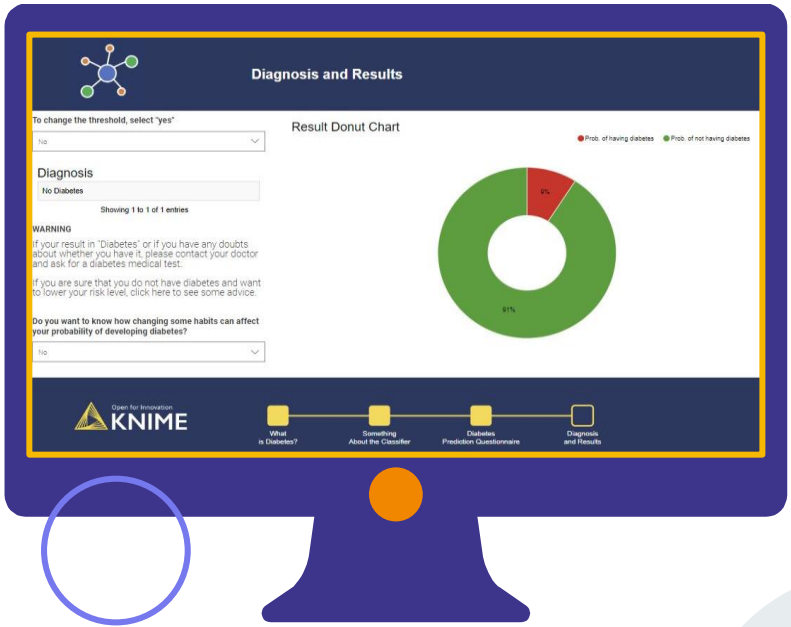
Workflow built to an actual use of the solution developed in the training phase



# DataApp

An effective and simple application for diagnosing diabetes, which could potentially save lives

**To the DataApp**



# The End

Thanks for the Attention