Trinity College Dublin
Coláiste na Trínóide, Baile Átha Cliath
The University of Dublin

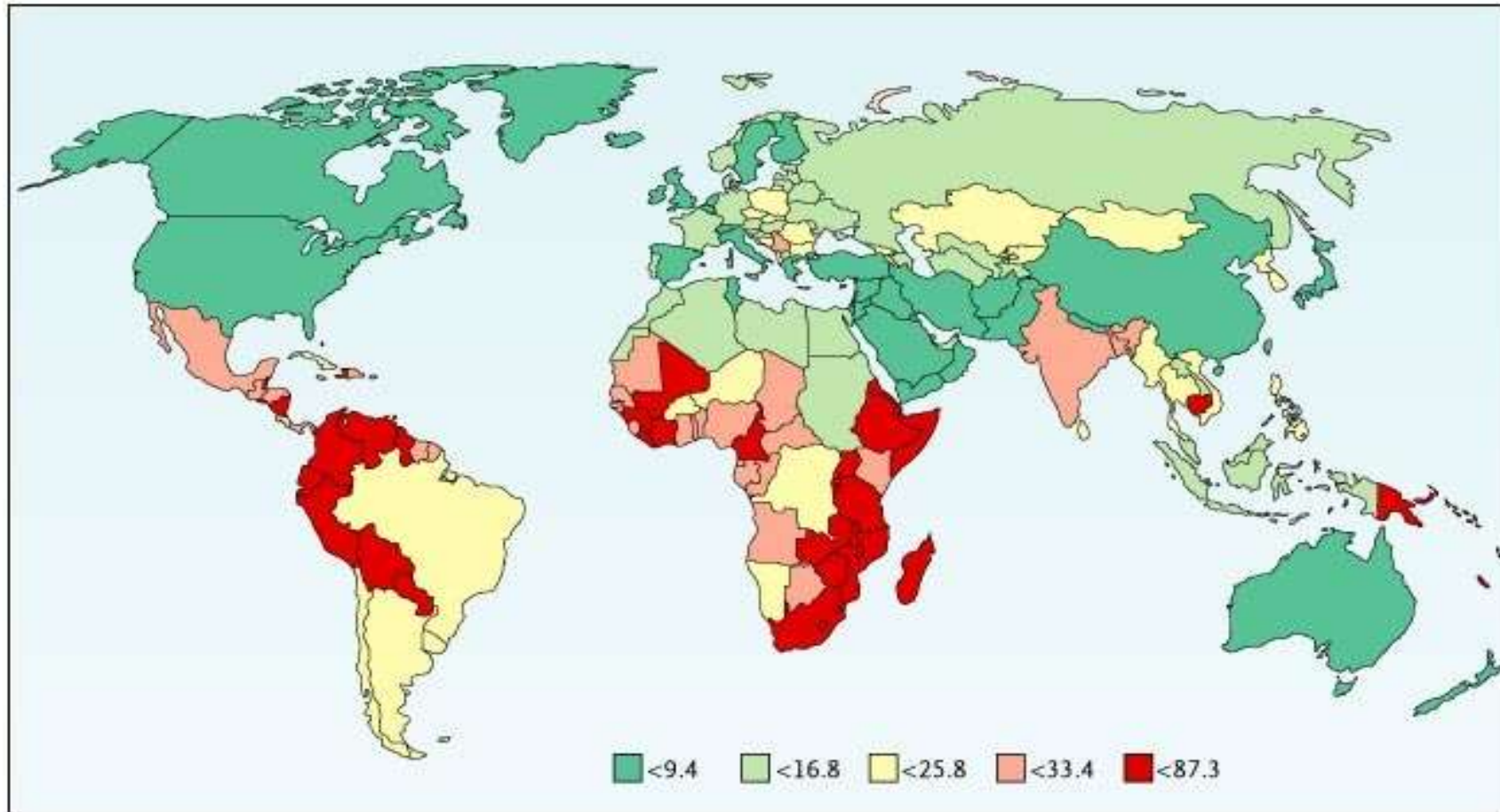# Explainable Machine Learning for Cervical Cancer Risk Factors Assessment

dr. ir. Sultan K. Imangaliyev

Dublin, 25/11/2020
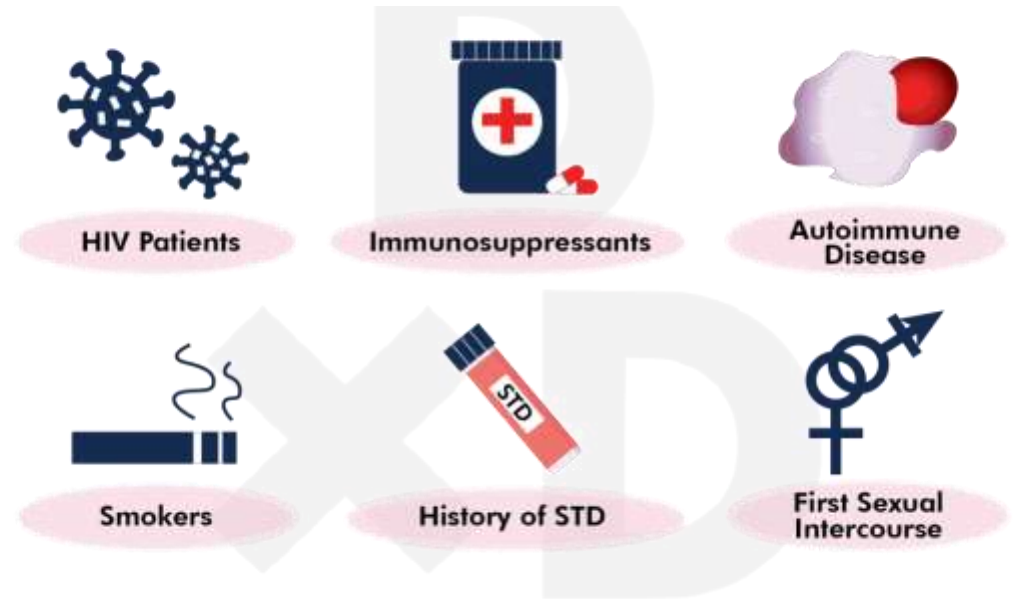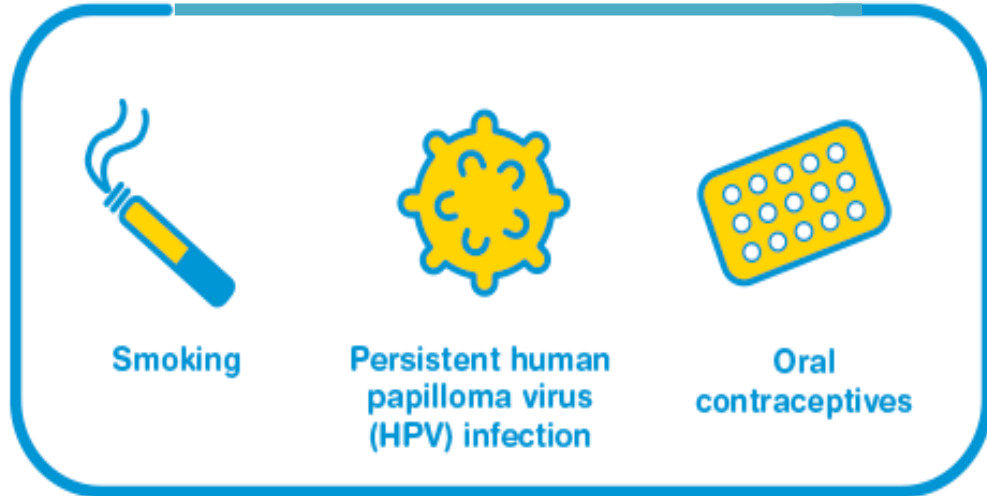
Research question

Which epidemiological and/or demographical variables are predictive for cervical cancer screening?

Cervical cancer worldwide

Incidence of Cervical Cancer, cases per 100K population

Cervical cancer risk factors

Smoking

Persistent human papilloma virus (HPV) infection

Oral contraceptives

HIV Patients

Immunosuppressants

Autoimmune Disease

Smokers

History of STD

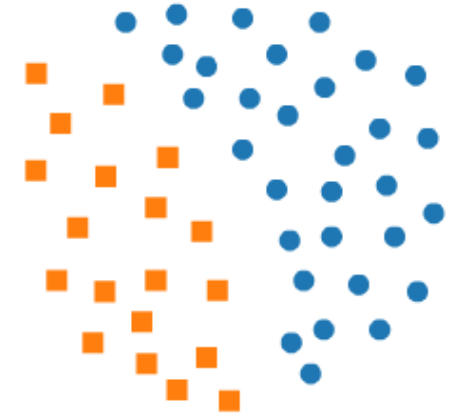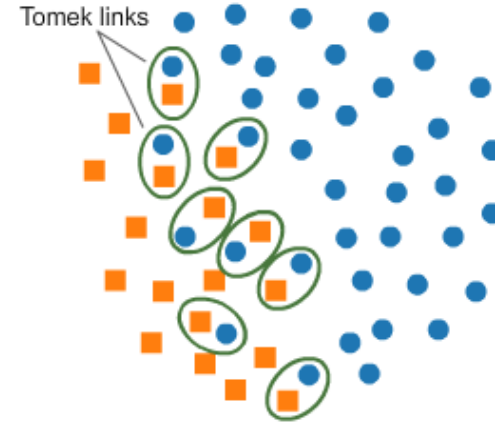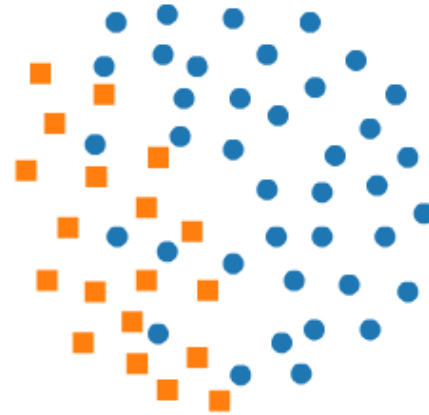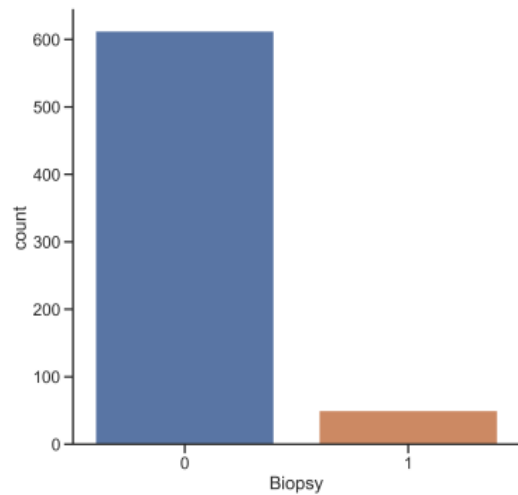First Sexual Intercourse

Case study

- Publicly available dataset[*]
  - 858 patients of *Hospital Universitario de Caracas*, Venezuela

- 36 features
  - demographics, lifestyle, medical history *etc*.
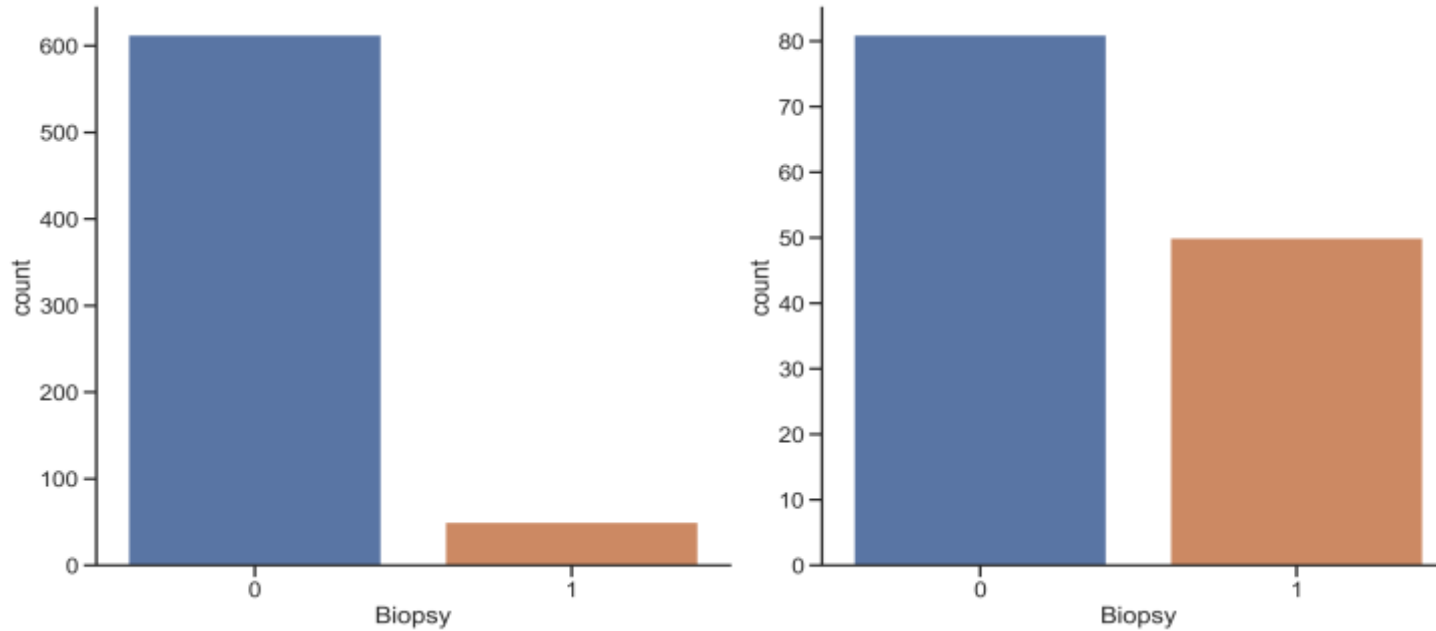
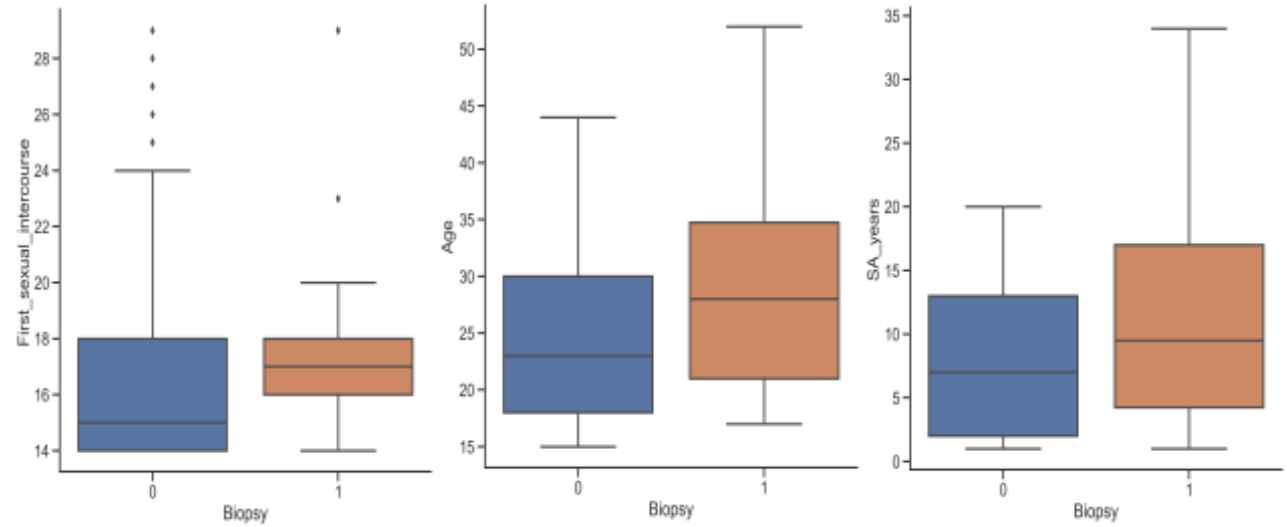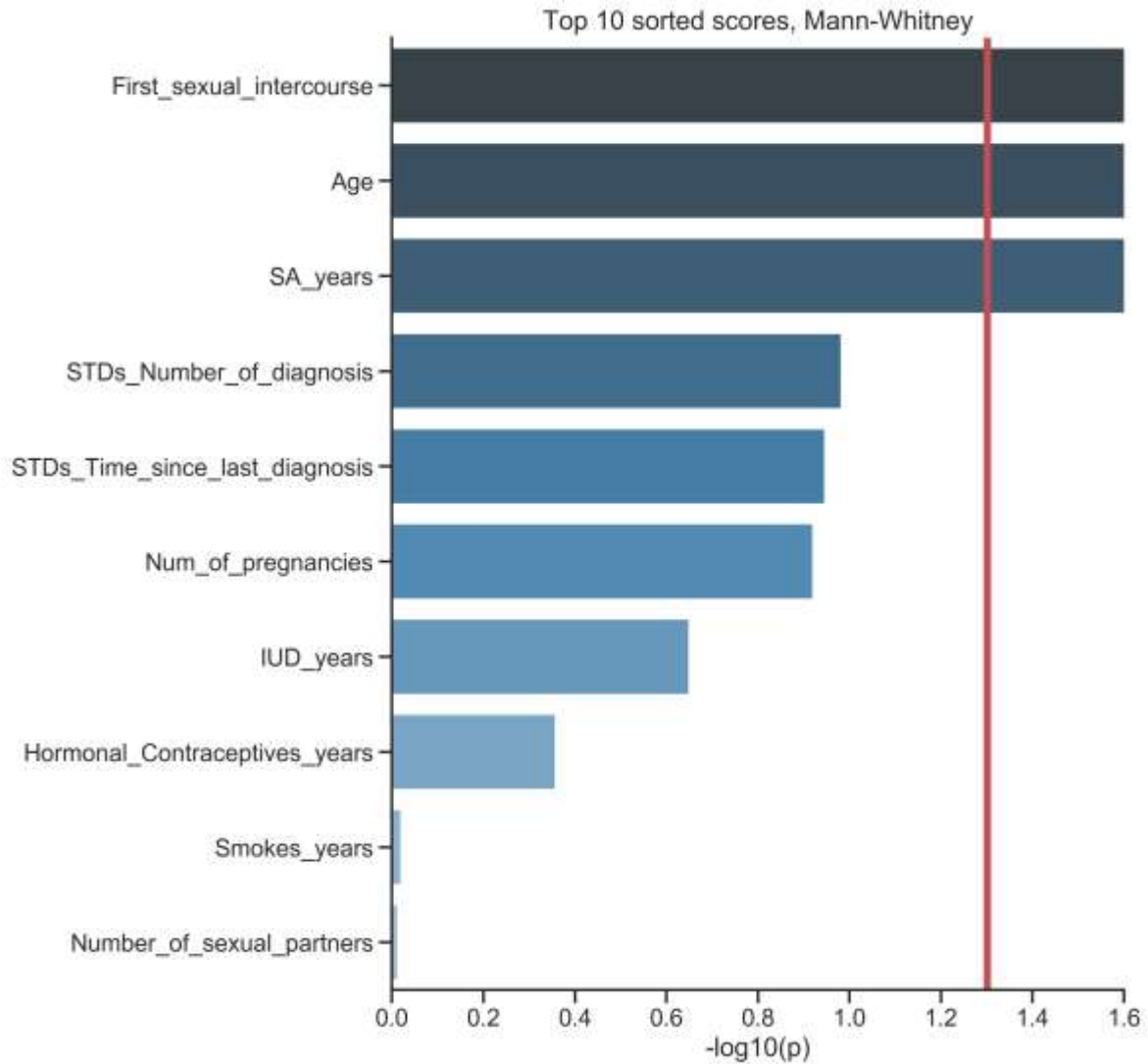- Biopsy as a binary classification label

Missing values map
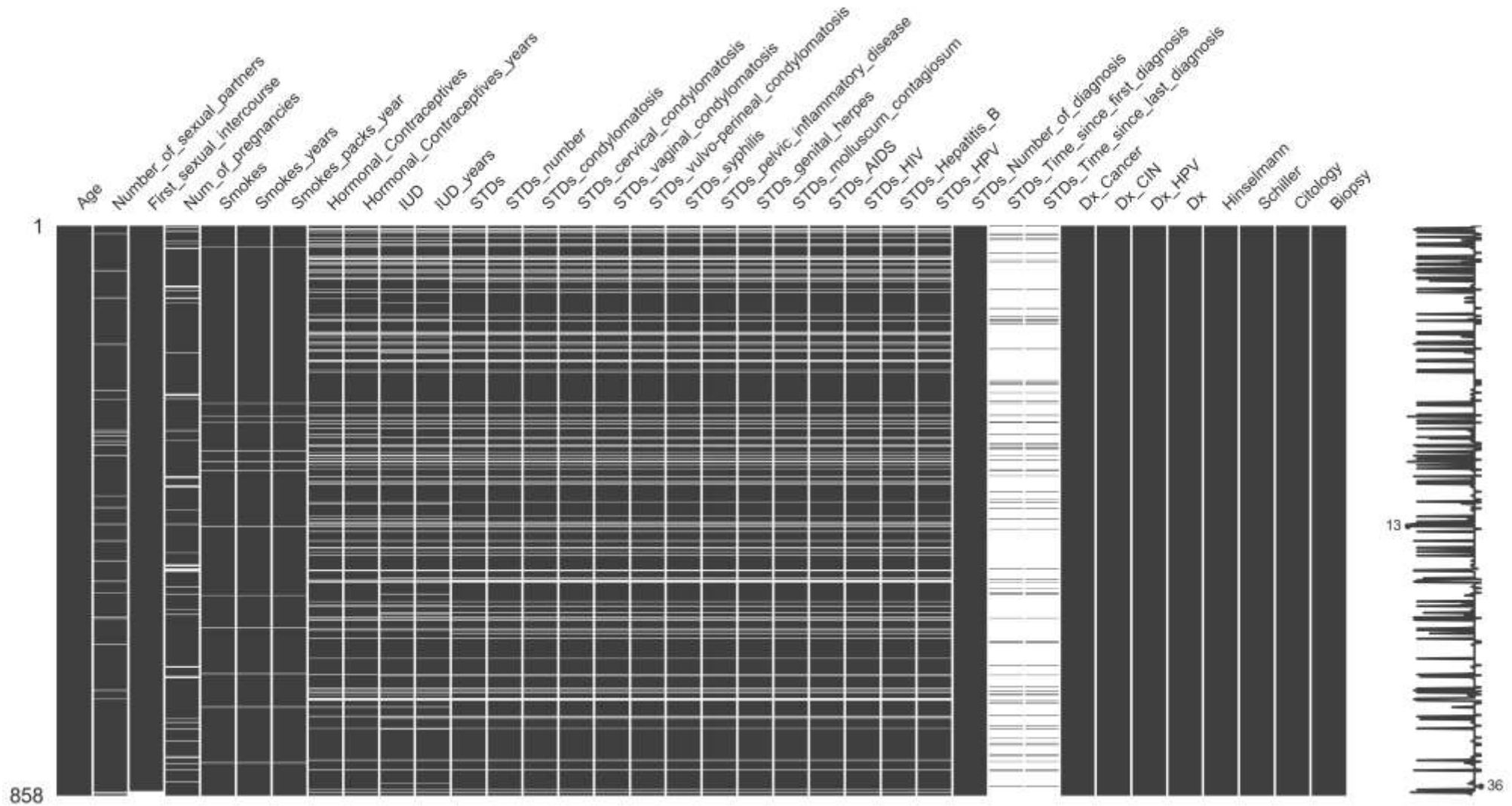
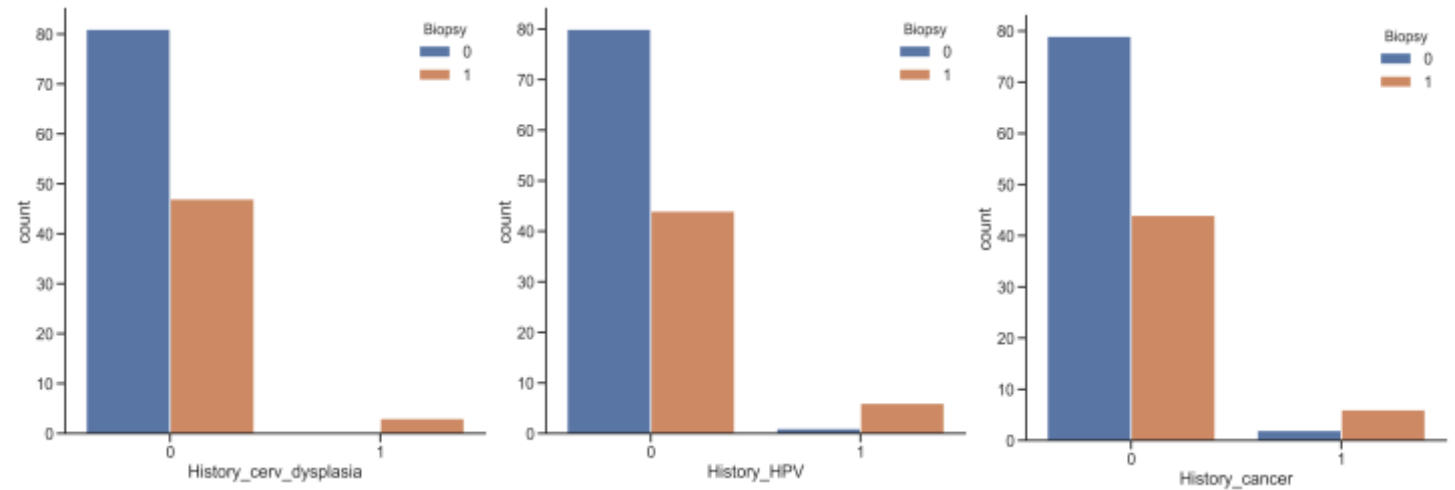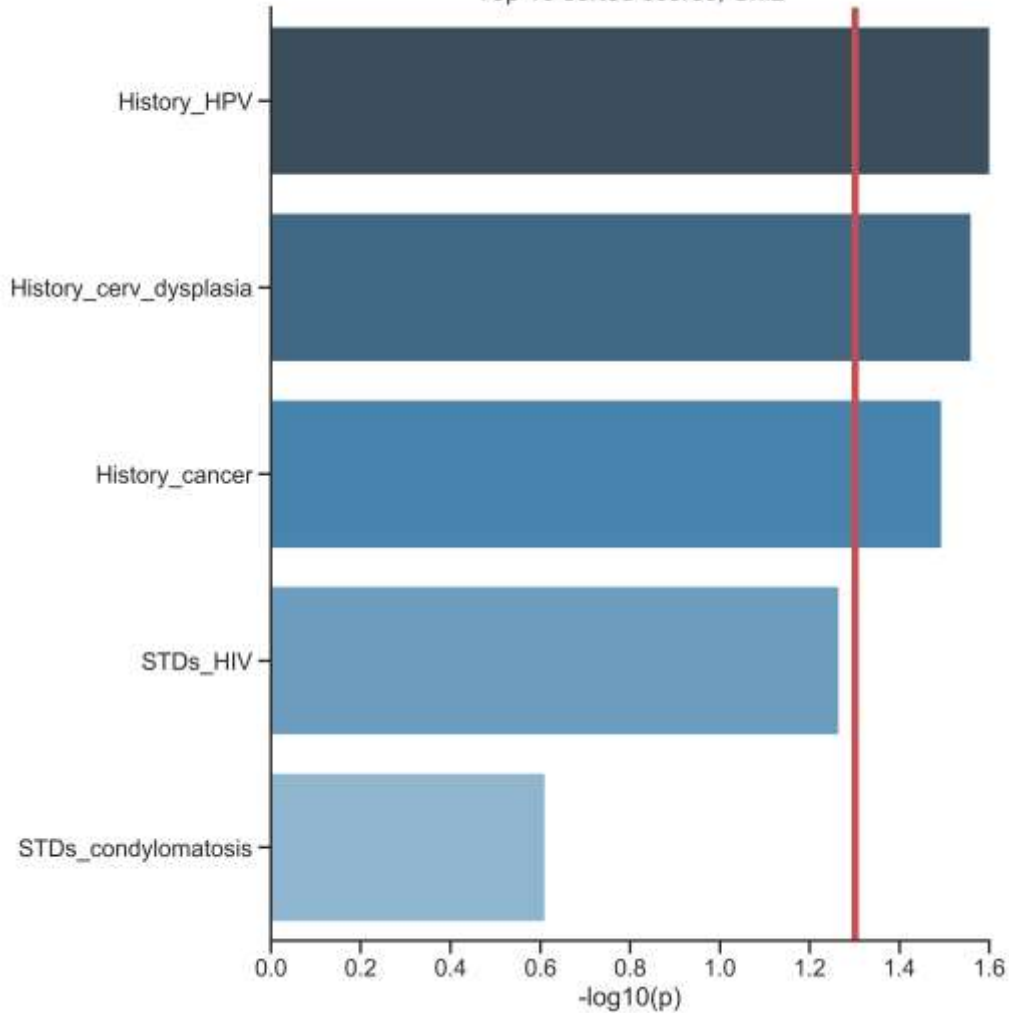# Class imbalance and data subsampling

Majority class subsampling

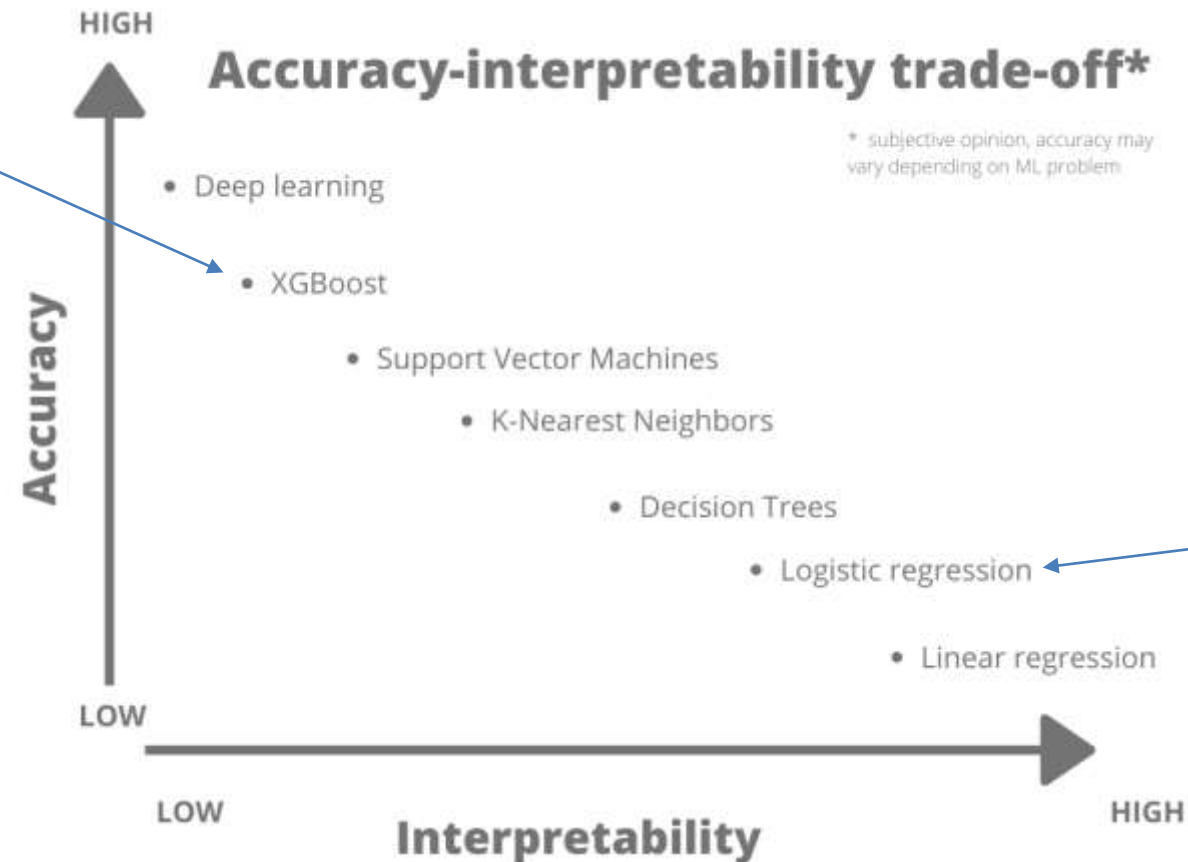# Statistical tests, Mann-Whitney U-test

Statistical tests, Chi-squared test

Motivation for Explainable ML

Published in 2016

Published in 1845

Accuracy-interpretability trade-off*

* subjective opinion, accuracy may vary depending on ML problem

HIGH

- Deep learning
- XGBoost
- Support Vector Machines
- K-Nearest Neighbors
- Decision Trees
- Logistic regression
- Linear regression

Accuracy

LOW

LOW — Interpretability — HIGH

Linear models

The Logistic Function

$$\text{Log}\left[\frac{Y}{(1-Y)}\right] = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_nX_n$$
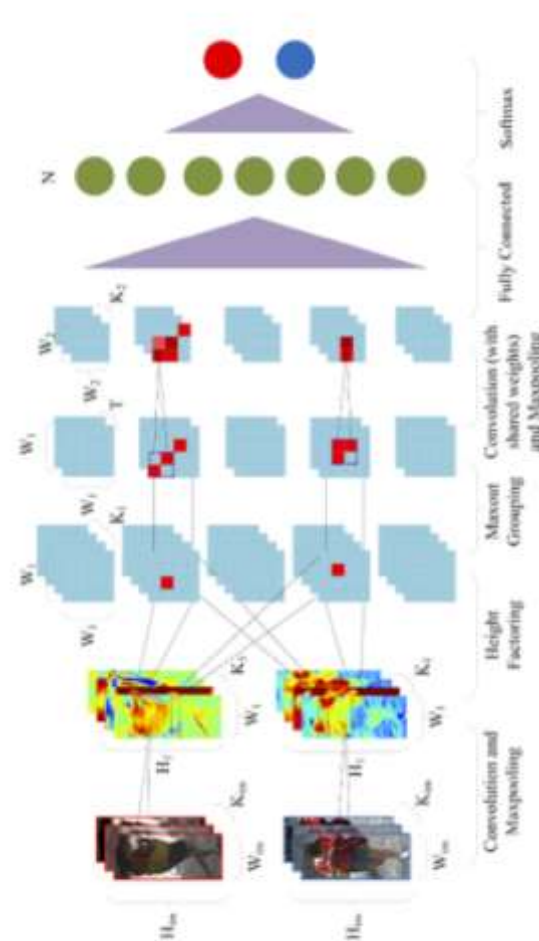
Log(Likelihood)

diet
score
(0-15)

age
group
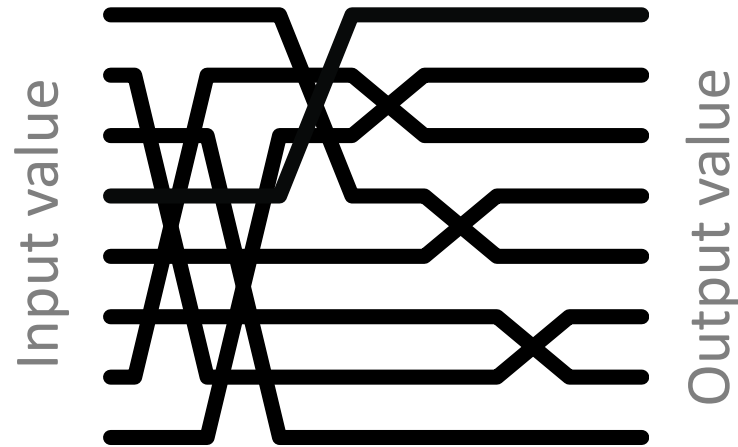(0/1)

sex
(0/1)

Non-linear models

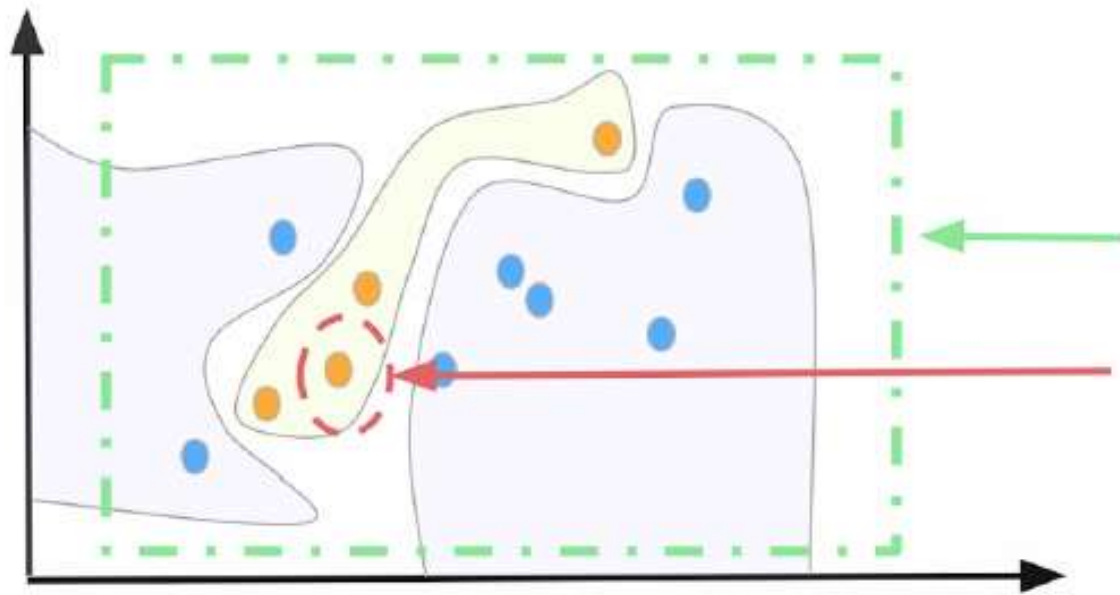# Local explanations

Input value

Output value

Complex models are inherently complex!

But a single prediction involves only a small piece of that complexity

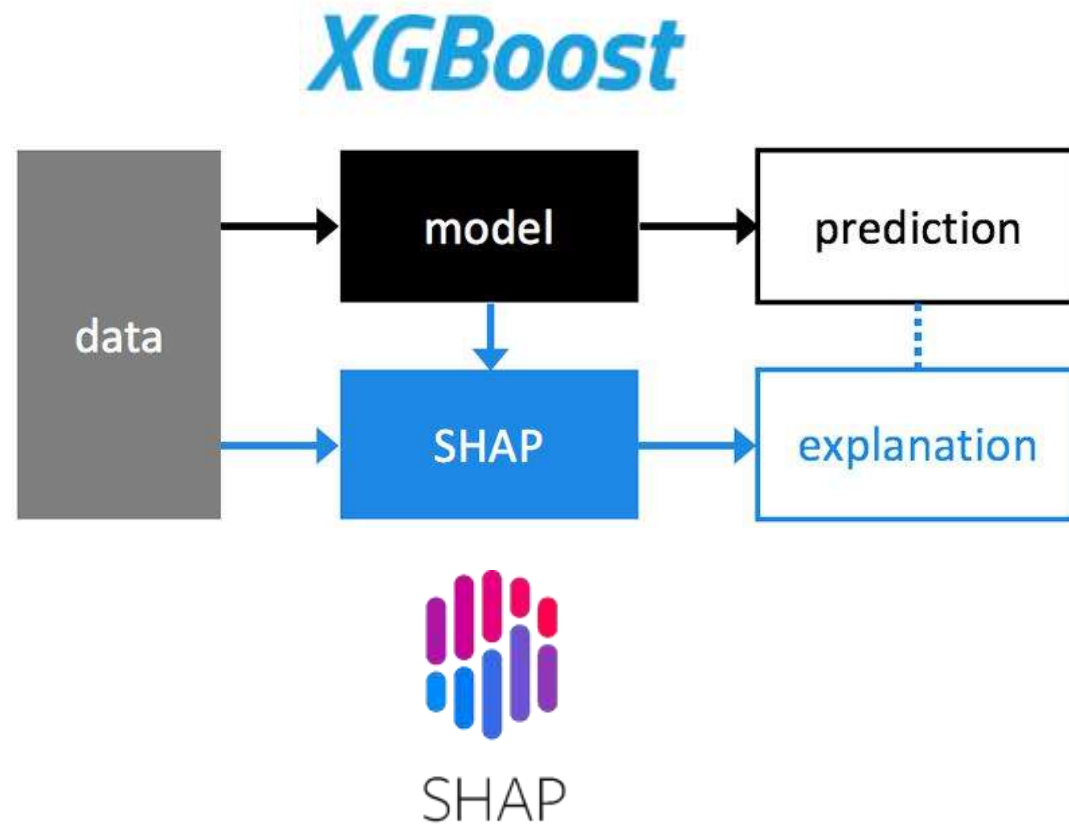# Surrogate linear model

Global Interpretation

Local Interpretation

Averages effects over data dimensions

Assesses individual effects

Framework

SHAP Explainer

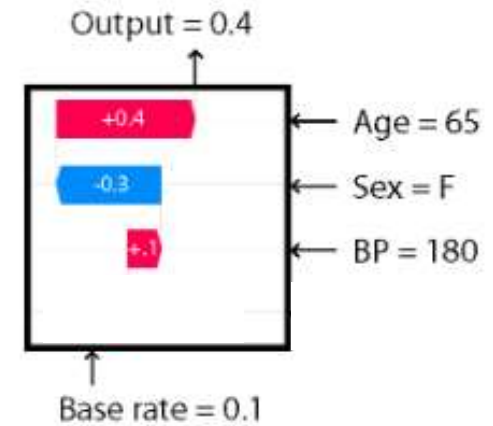**SH**apley **A**dditive ex**P**lanations

Lloyd Shapley (*1923 – 2016*),
2012 Nobel Prize in Economics

- A method from coalitional game theory
  - Players cooperate in a coalition and receive a certain profit from this cooperation

- Features of the model
  - *"Players"*
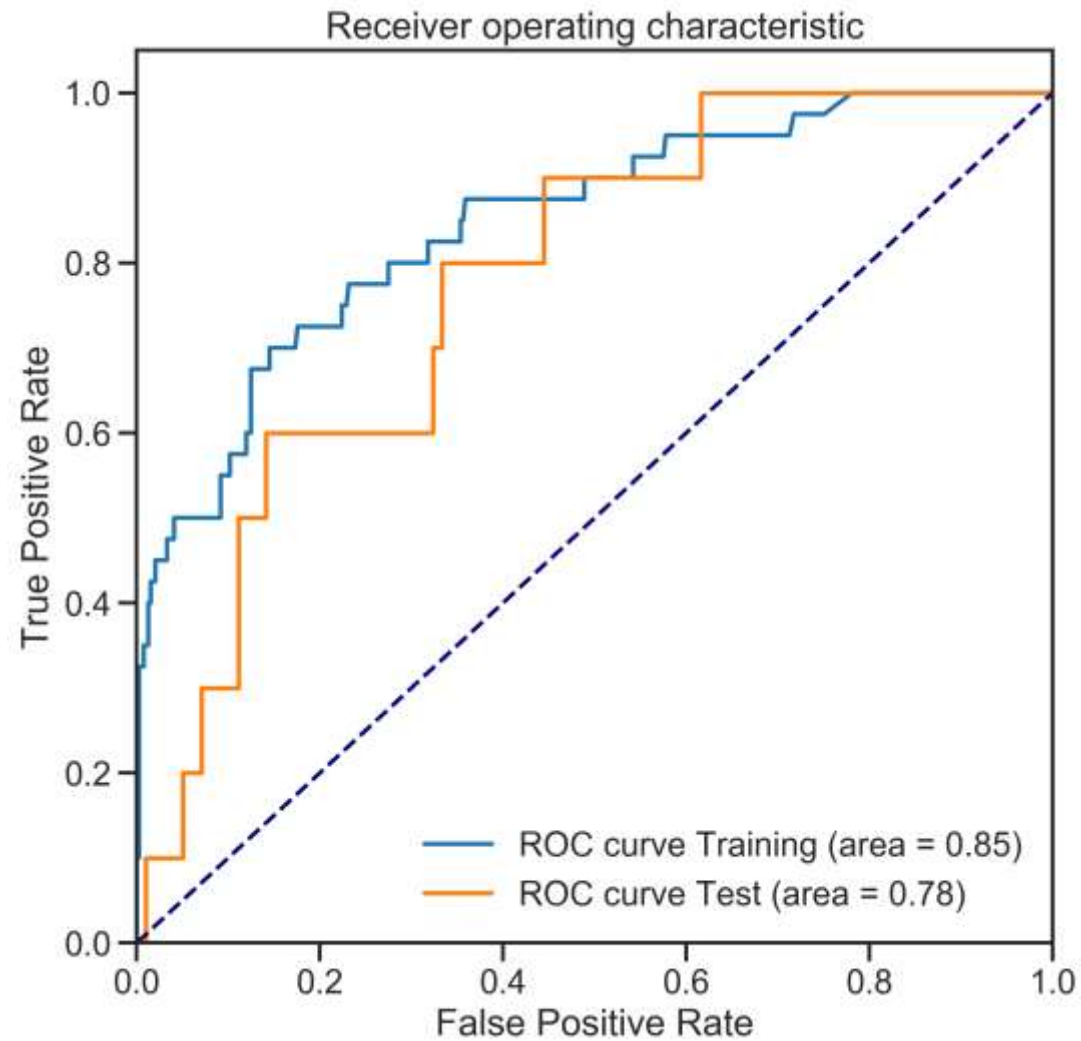- Outcome of the model
  - *"Profit"*

Shapley value in a nutshell

F1	F2	F3	F4
X			~
	X		~
		X	~
X	X		~
	X	X	~
X		X	~
X	X	X	~

F1, F2, F3   + F4

y   + Δy

Output = 0.4

+0.4 ← Age = 65

-0.3 ← Sex = F
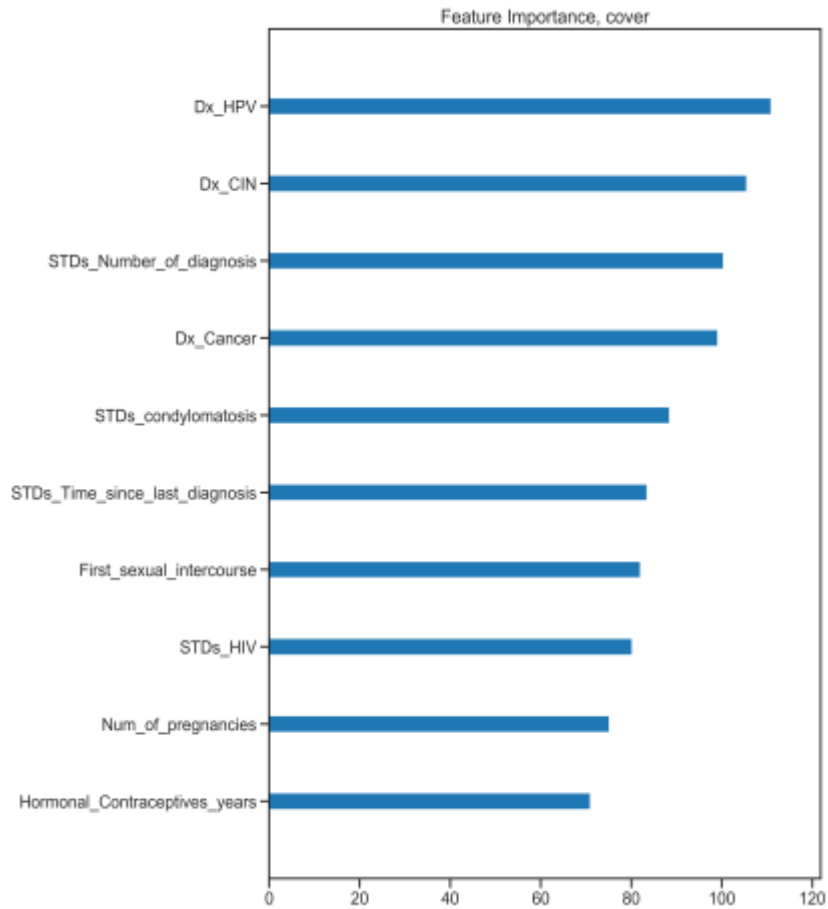
+.1 ← BP = 180

Base rate = 0.1

How much did the feature contribute to the models prediction in *all possible* combinations?

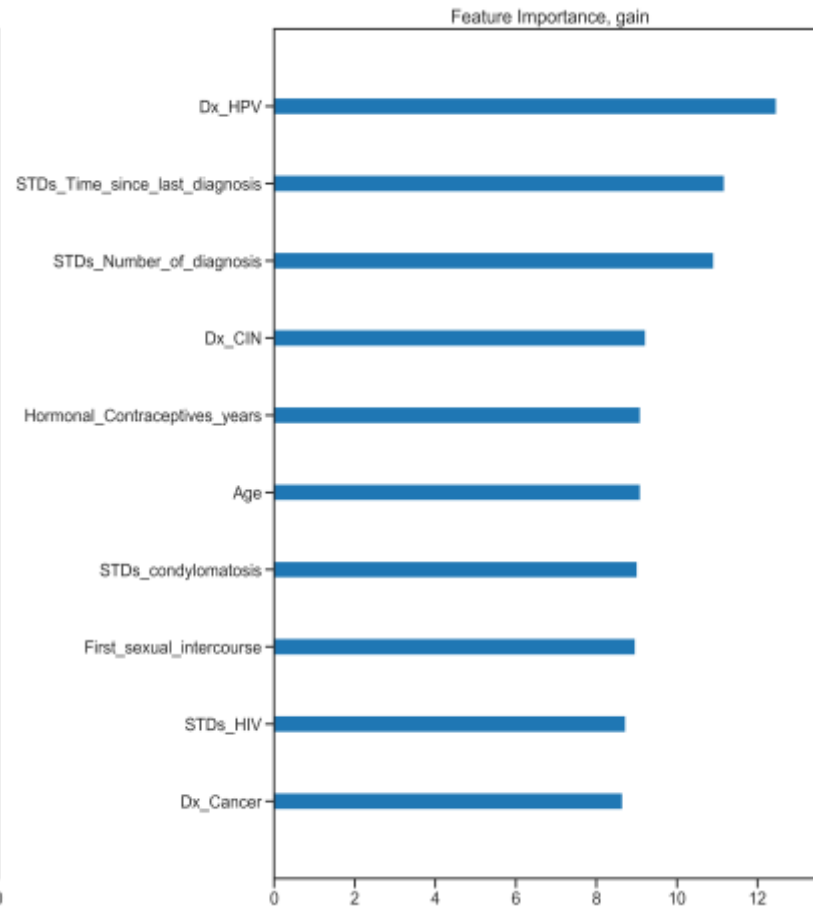# XGBoost performance
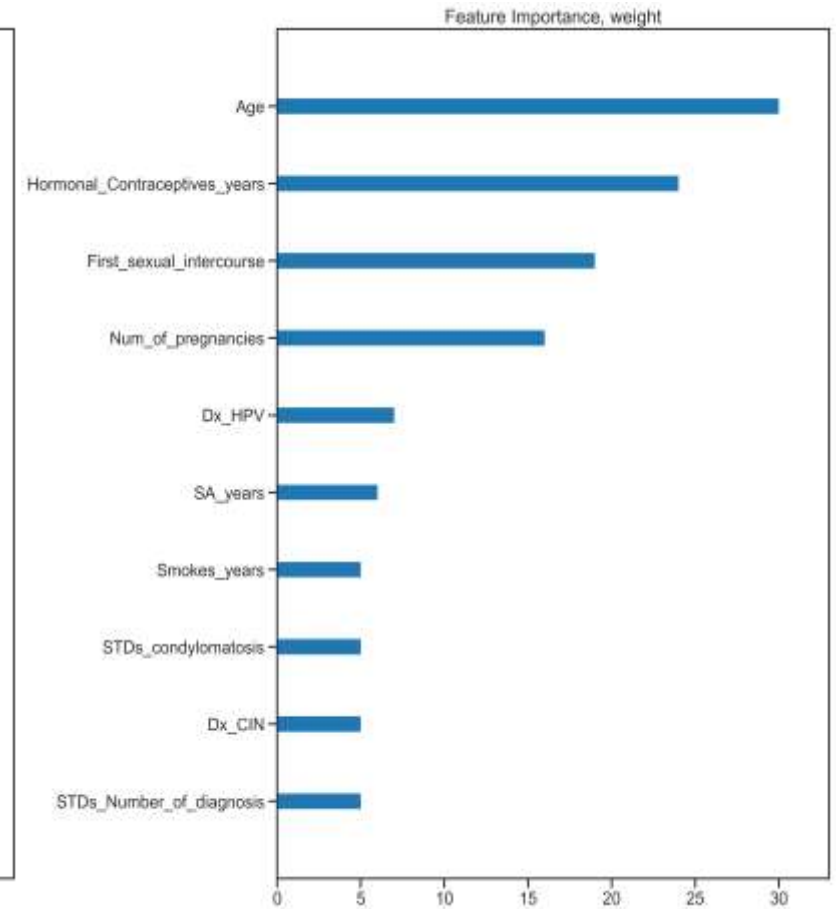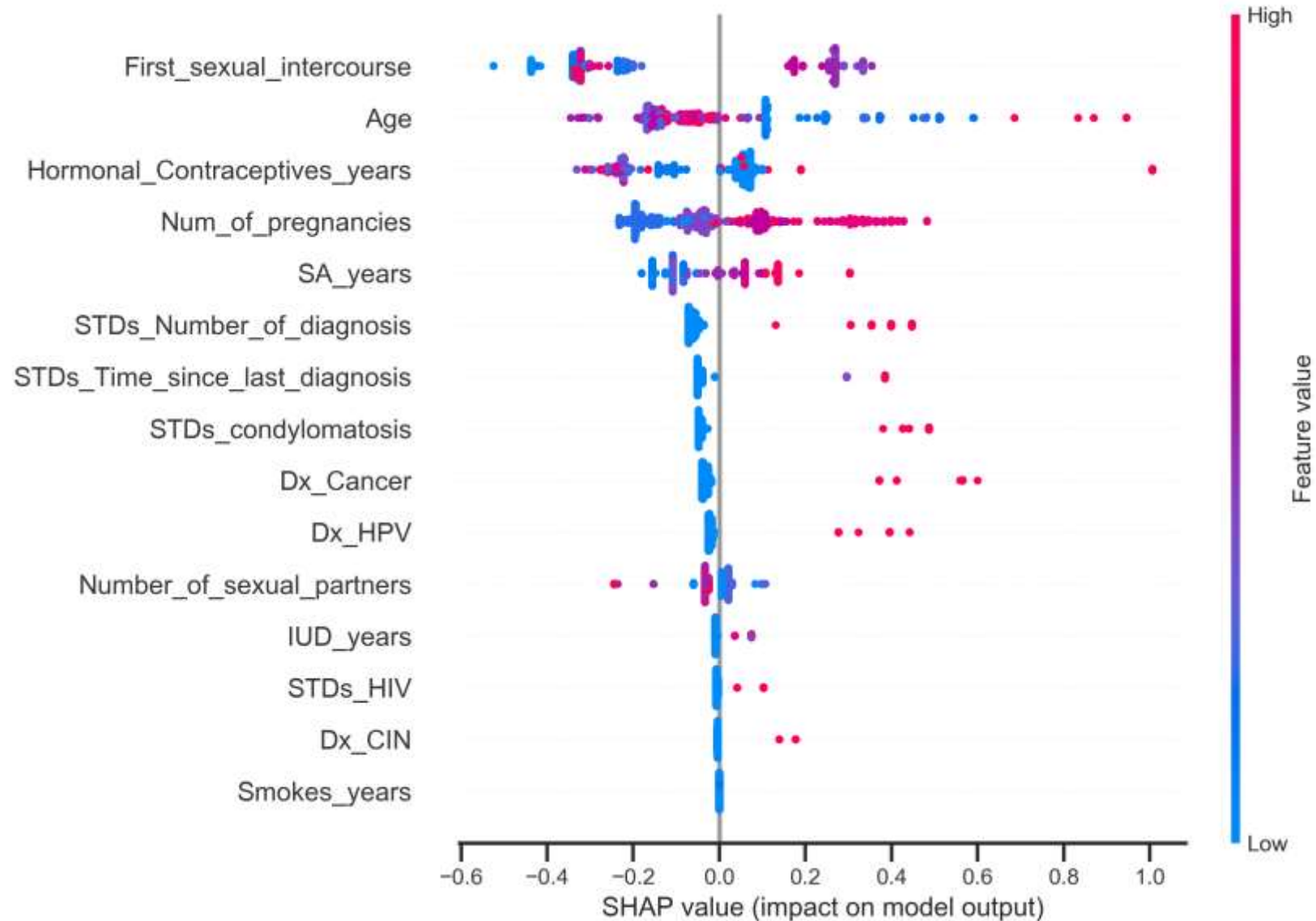
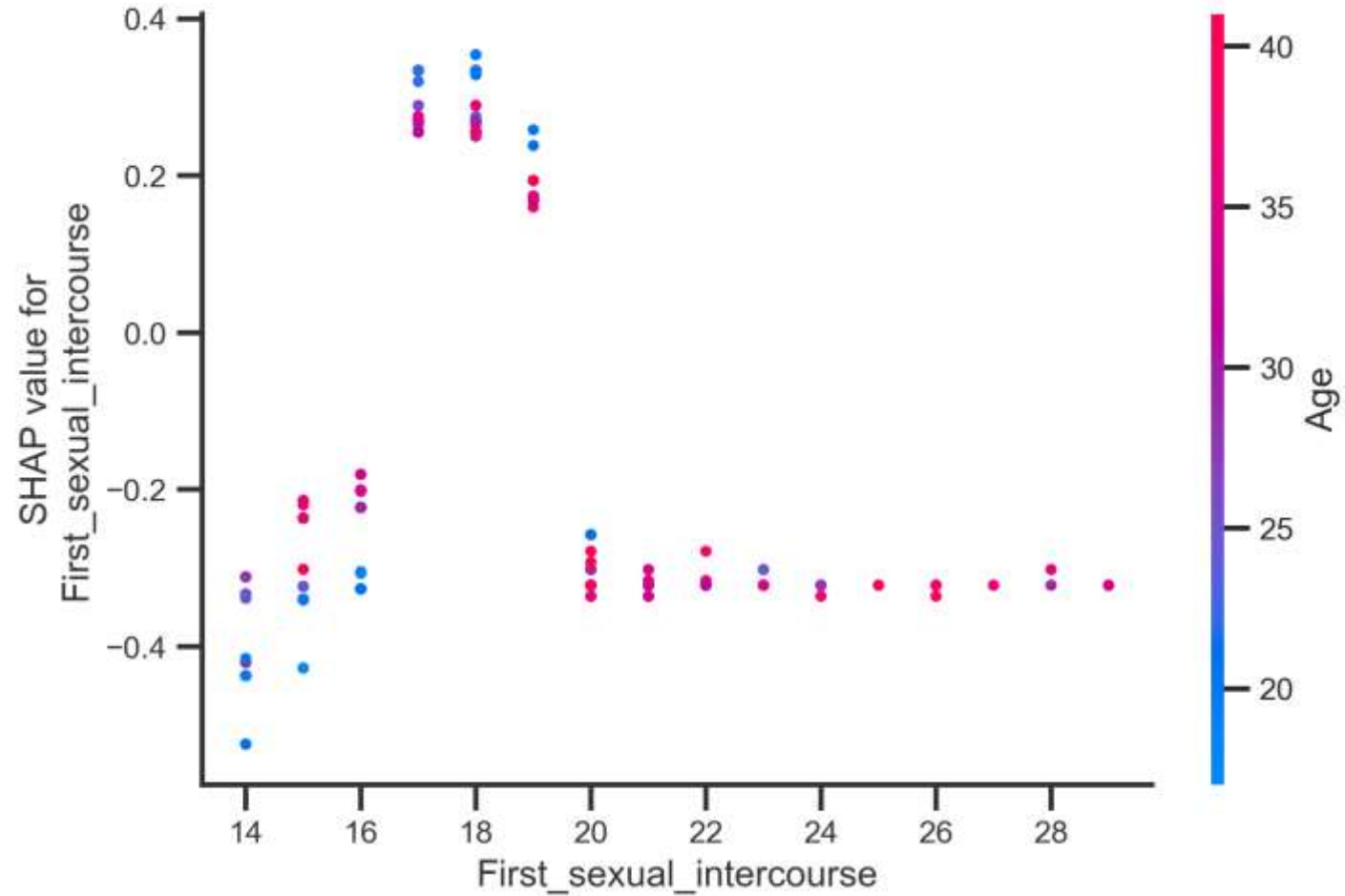# Feature importance inconsistency
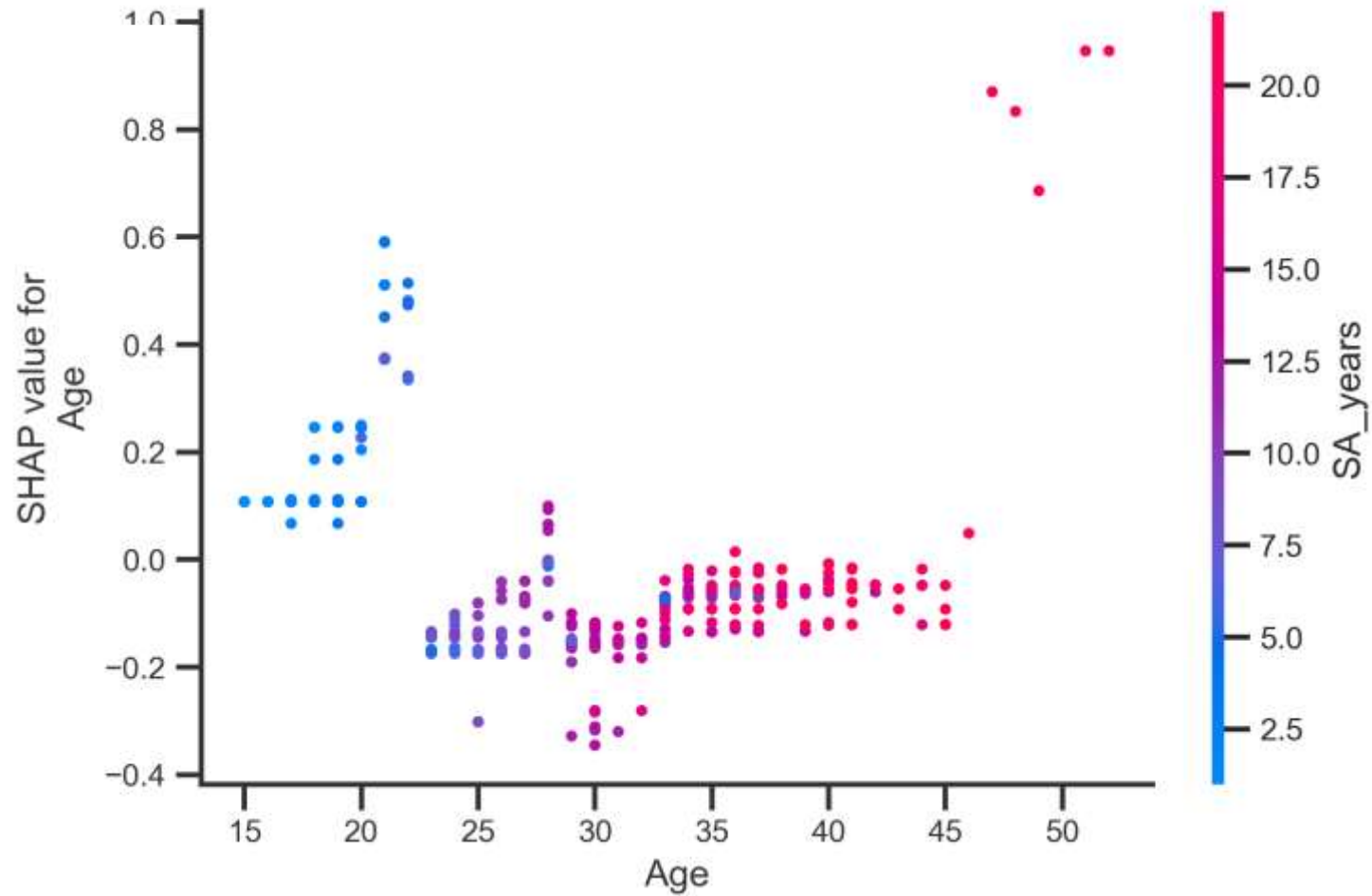

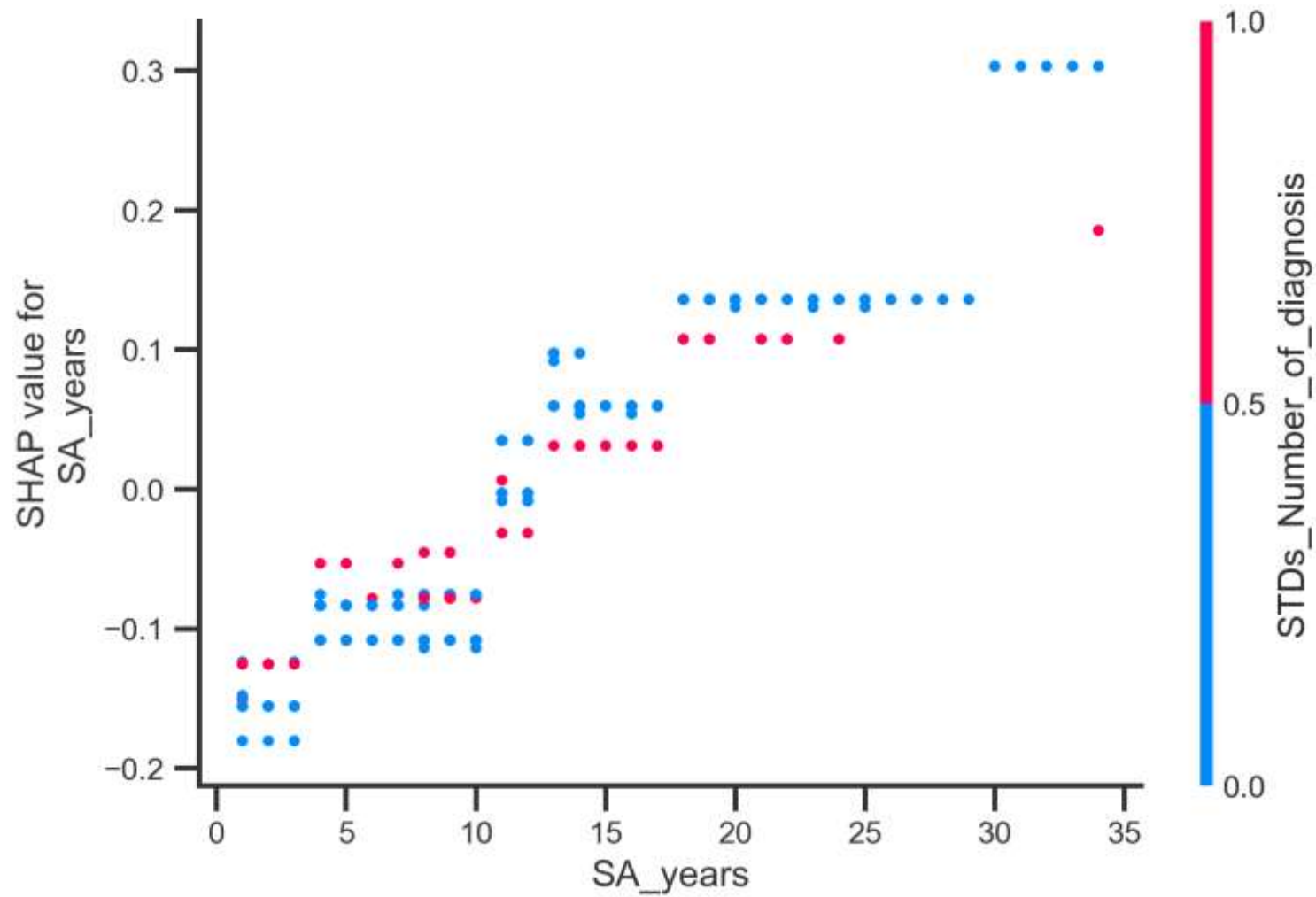Cover   Gain   Weight

SHAP output
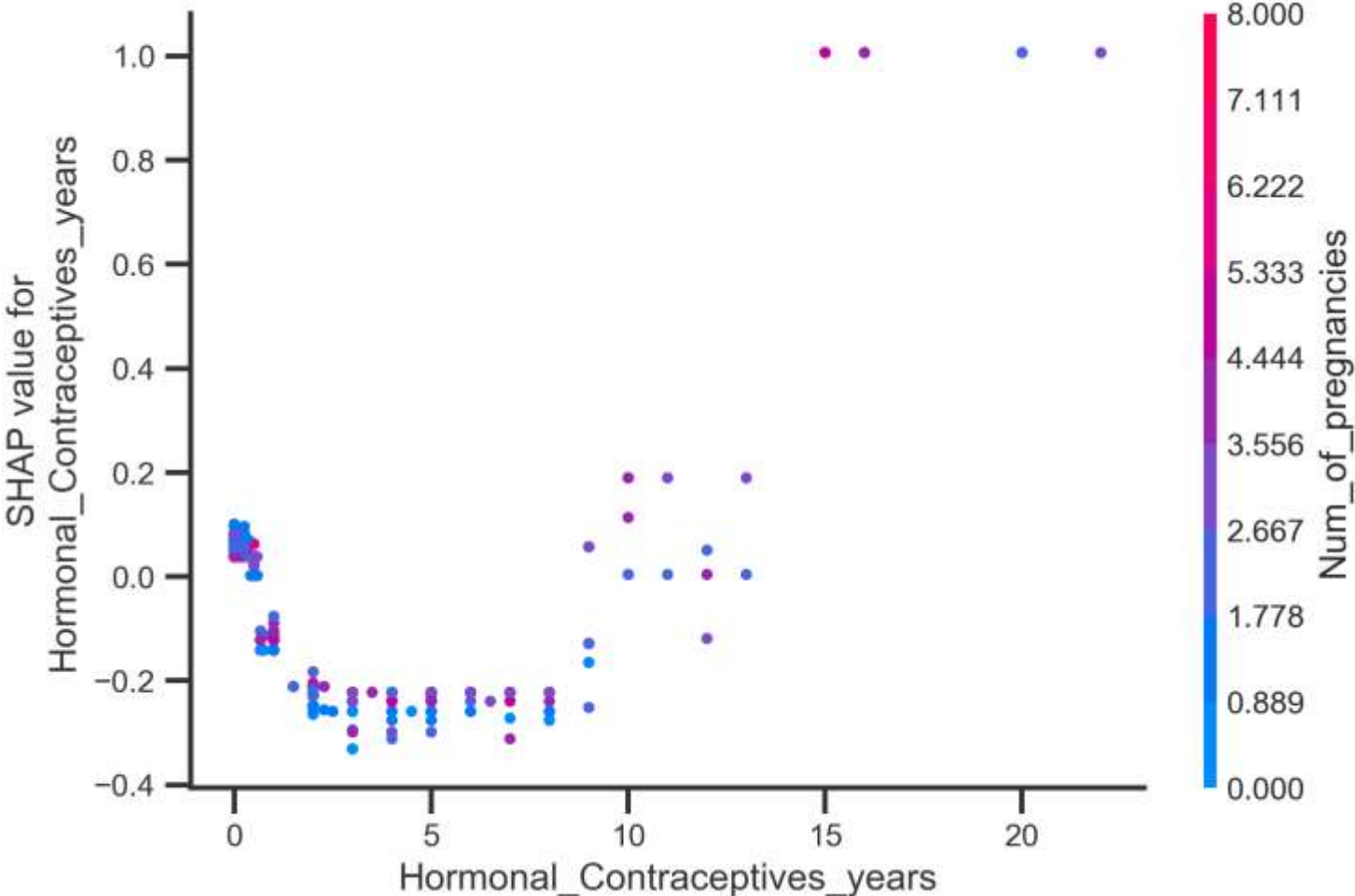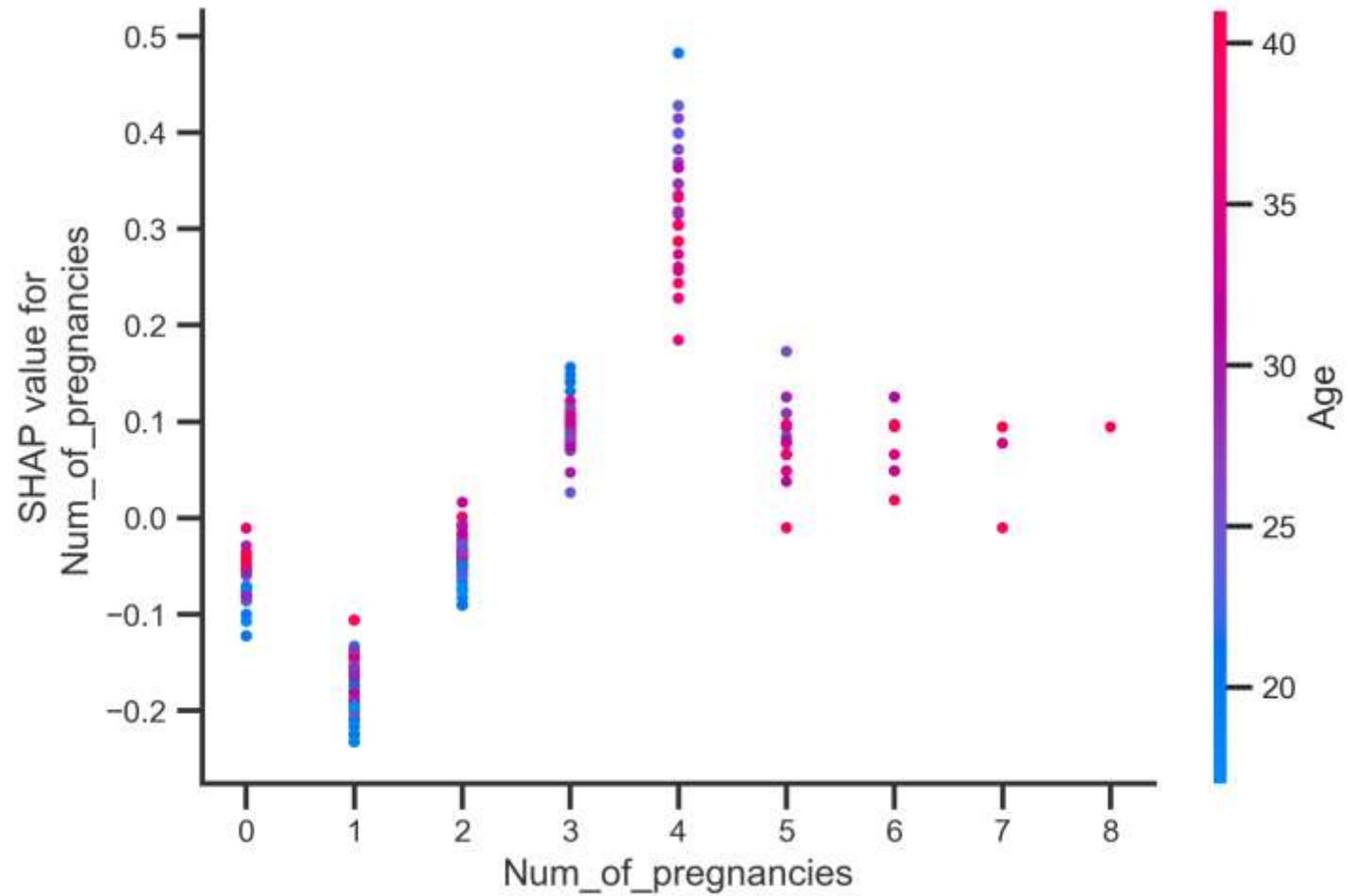
Sexual debut age

Biological age

Sexually active age

Hormonal contraception duration

Parity*

*Parity = the number of times that a woman has given birth to a foetus with a gestational age of 24 weeks or more, regardless of whether the child was born alive or was stillborn
 Gravidity = number of times that a woman has been pregnant

Examples of positive class predictions

Examples of negative class predictions

True Negative, P=0.25

False Negative, P=0.22
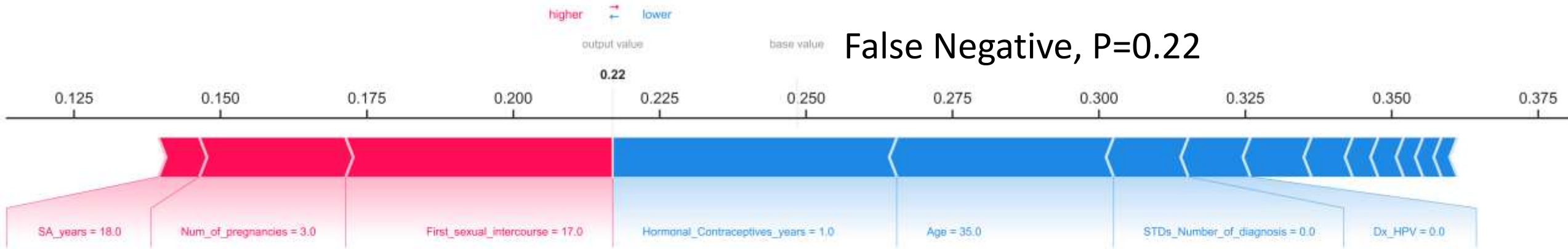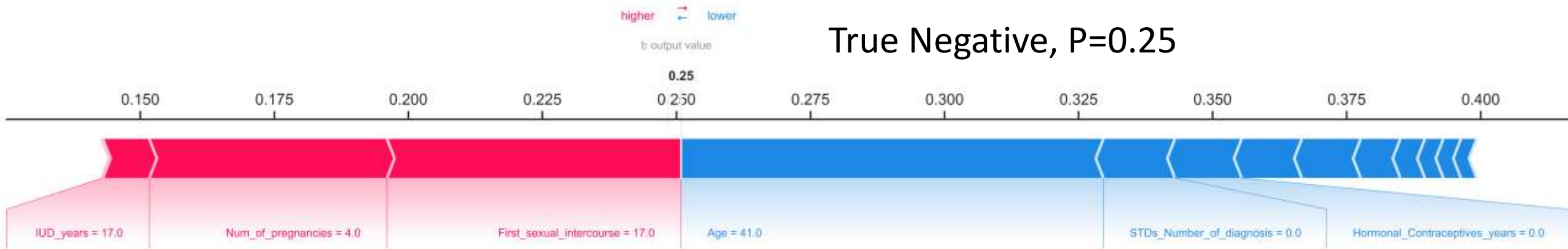
Conclusions

- Good predictive performance
  - Training ROC AUC = 0.85, Test ROC AUC = 0.78

- Explainable Machine Learning identified important features
  - Sexually active age
  - Hormonal contraception usage
  - Number of pregnancies
  - History of STDs

- Non-linear dependency between features and predicted cervical cancer risk

Discussion