# Diabetes prediction

## Supervised Learning

2nd Project

Duarte Assunção
Guilherme Matos
João Ferreira

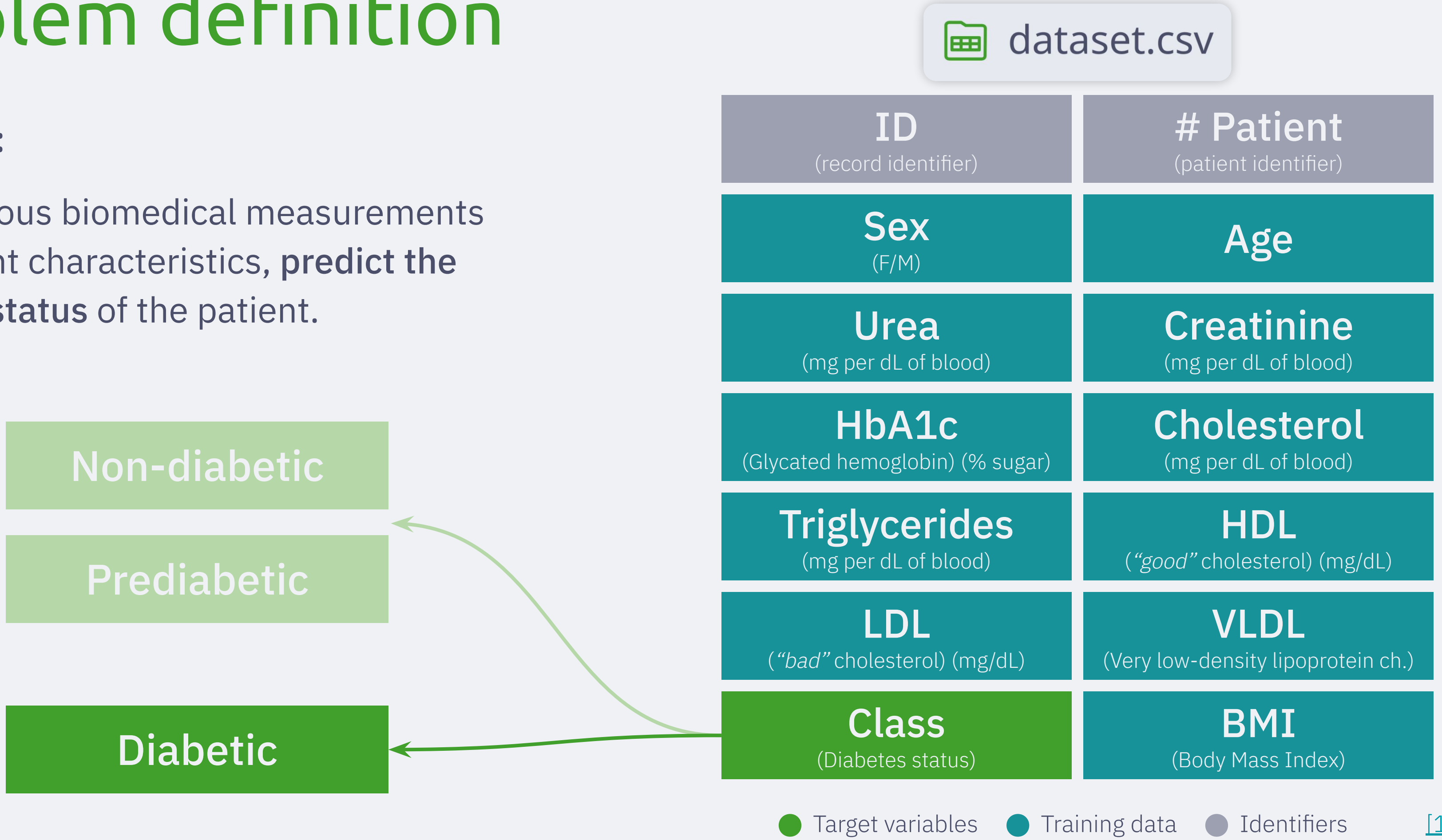
IA (L.EIC) 2024/2025

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

# Problem definition

**Objective:**

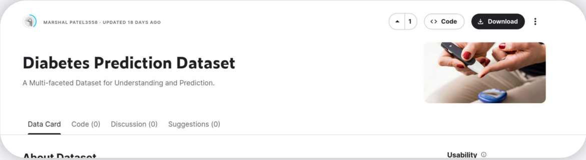
Given various biomedical measurements and patient characteristics, **predict the diabetes status** of the patient.




[1]

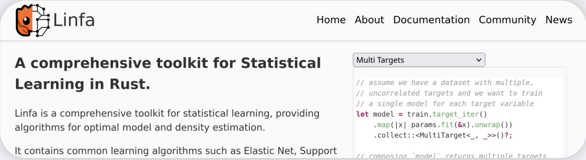
# Related work

Now with Python code examples!



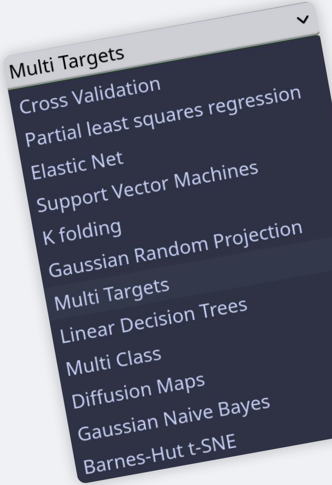

## [1] Diabetes Prediction Dataset @ kaggle

https://www.kaggle.com/datasets/marshalpatel3558/diabetes-prediction-dataset-legit-dataset/data




## [2] Algorithm examples with the Linfa crate

https://rust-ml.github.io/linfa/





## [3] Paper with the same goal and methodology

Afolabi, S., Ajadi, N., Jimoh, A., & Adenekan, I. (2025). Predicting diabetes using supervised machine learning algorithms on E-health records. Informatics and Health, 2(1), 9-16. https://doi.org/10.1016/j.infoh.2024.12.002

# Methodology

**Tools and libraries:**

- Rust programming language;
- Jupyter Notebook and EVCXR kernel;
- Plotters (~ Matplotlib);
- Ndarray (~ NumPy);
- Linfa (~ Scikit-learn).

**Evaluation metrics:**

- Accuracy: (TN + TP) / All
- Sensitivity or Recall: TP / (TP + FN)
- Specificity: TN / (TN + FP)
- Training time;
- Testing time.

# Data preprocessing

dataset.csv

**Data analysis:**

- 800 unique patients;
- 200 duplicate IDs with different genders (ignored);
- 56% male and 43% female;
- 70.6% of patients are between 50 and 61 years old;
- Urea, Cr, TG, HDL and VLDL may have outliers;
- 84% of patients have diabetes, i.e., the database is highly unbalanced!

**Data preprocessing:**

- Prediabetic patients will be labelled non-diabetic;
- Sex and Class will be encoded into a boolean;
- Errors in the database need to be fixed: Sex and Class have duplicated types;
- Oversampling the minority classes;
- 80% of the database will be used for training and 20% for testing;
- No normalization needed.

# Machine Learning Models

### Support Vector Machines

- Handles nonlinear feature interactions;
- Accuracy in high-dimensional spaces;

- Gaussian kernel
  for non-linear data;
- nu weight of 1%
  for only a few support vectors.

### Gaussian Naive Bayes

- Assumes feature independence;
- Foundation in conditional probabilities;

- Variable smoothing of `1e-9`
  for slow variance calculation stabilization;

### (Linear) Decision Trees

- Key-criteria subgroups (like a diagnosis);
- “Automatic” feature selection;

- Max depth of 100
  for efficiency;
- `Gini` algorithm of splitting
  for avoiding misclassifications;

# Results: Confusion Matrix



Predicted / Expected

**Support Vector Machines**

| | False | True |
|---|---|---|
| False | 165 | 0 |
| True | 6 | 166 |

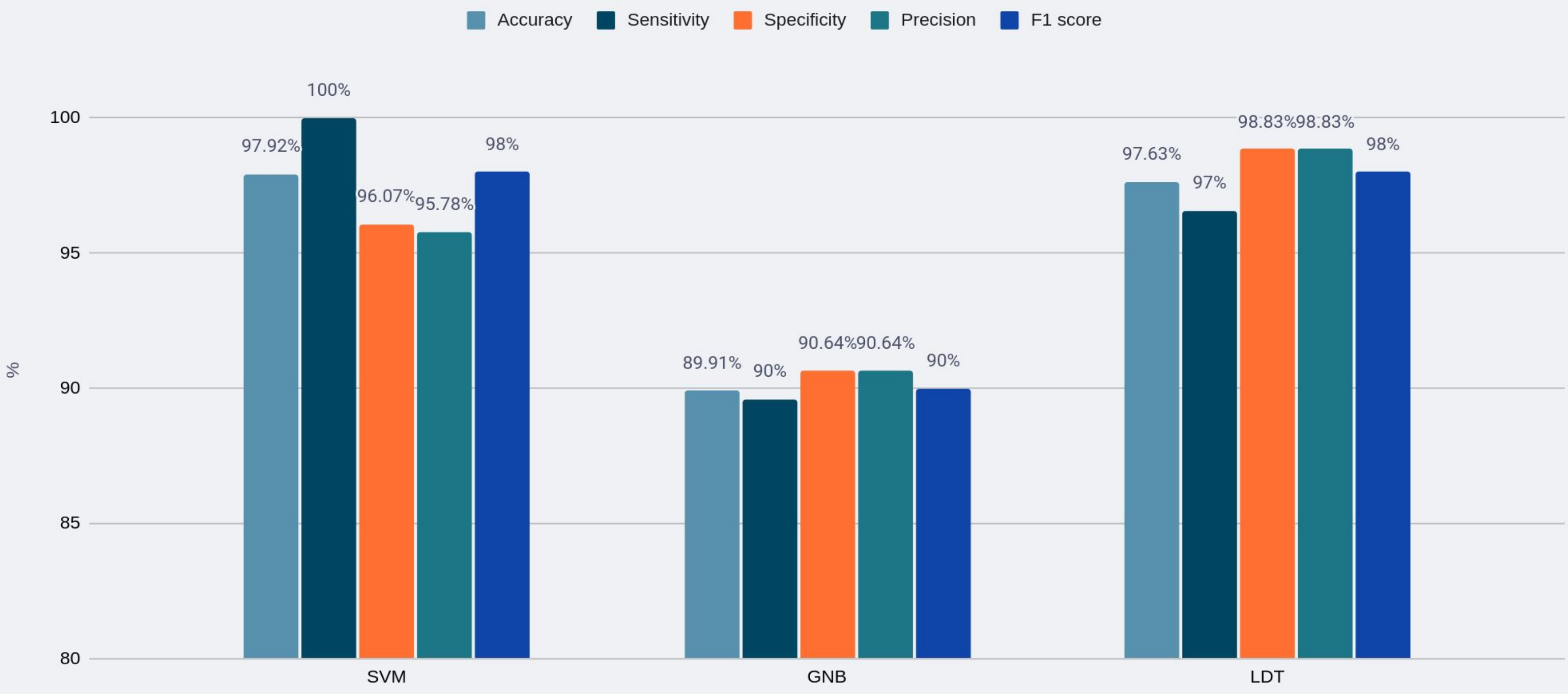
**Gaussian Naive Bayes**

| | False | True |
|---|---|---|
| False | 158 | 18 |
| True | 8 | 153 |

**(Linear) Decision Trees**

| | False | True |
|---|---|---|
| False | 165 | 8 |
| True | 1 | 163 |

# Results: Statistics



# Results: Time

*Tested with ~1600 entries (after preprocessing)