UNIVERSITÀ DEGLI STUDI DI TRIESTE

# Data Preparation Strategies in a Machine Learning Application: A Case Study on Voting Intentions

Luca Pennella

December 12th, 2024

# Overview

1. **Introduction**
2. **Data Preparation**
3. **Choosing a good ‘model’**
4. **Results**

## What we will learn today (I hope)

- How to prepare data for statistical analysis.
- How to solve the most common problems in data preparation.
- How to try to do research in data analysis.
- The main characteristics of the voters of the different party coalitions (relative to 2017-2019).

## Who am I?

- PhD student in Applied Data Science and Artificial Intelligence (ADSAI) at the University of Trieste with the support of Rachael (Spinoff by SWG, the University of Trieste and SISSA).
- Guest Scholar at IMT of Lucca for projects on blockchain and Decentralized Finance.
- Bachelor in Economics, Master in Statistics from the University of Bologna and Master in Data Science and AI from the University of Florence and IMT of Lucca.
- A few years of experience as data analyst and data scientist in Jakala, Data Reply, Crif and Diennea.

## Goals

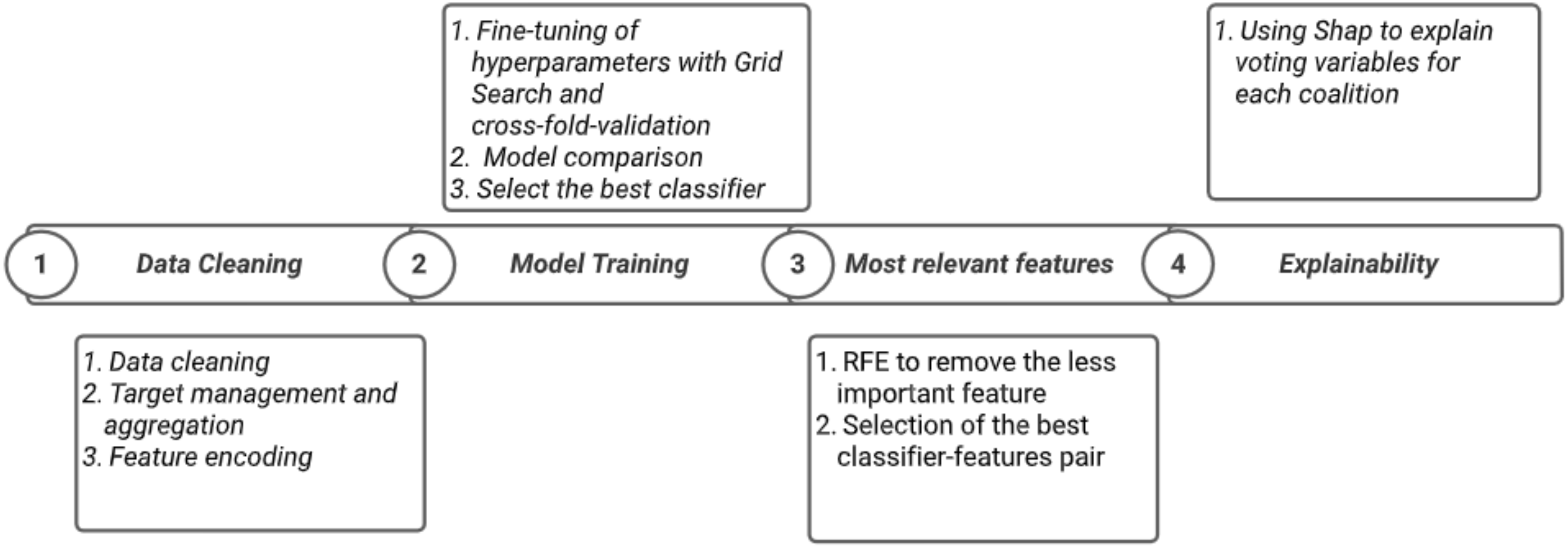
- **Understanding voting intentions** using demographic and values-based variables.
- **Classify party voting intentions** with high accuracy using machine learning methods.
- Apply eXplainable Artificial Intelligence (XAI) techniques to interpret the **impact of the most relevant features** for each class of target variable.

# Data Preparation

# Data

- Data collected annually from the **SWG and Rachael Monitoring survey** (sample of 1500 units each year from SWG Panel).
- We focus on the **interviews conducted from 2017 to 2019**.
- The data are totally anonymous.
- Initial dataset includes **4,504 users and 116 variables**.

# Pipeline


1. Fine-tuning of hyperparameters with Grid Search and cross-fold-validation
2. Model comparison
3. Select the best classifier
1. Using Shap to explain voting variables for each coalition
1
Data Cleaning
2
Model Training
3
Most relevant features
4
Explainability
1. Data cleaning
2. Target management and aggregation
3. Feature encoding
1. RFE to remove the less important feature
2. Selection of the best classifier-features pair

# Data preparation

1. Data normalisation for consistency across the survey year.
2. Data cleaning and handling missing data.
3. Apply **one-hot and ordinal encoding** for categorical variables.
4. Group each party into one of the three main coalitions (center-left (Sx/Csx), center-right (Dx/CDx) and Movimento 5 Stelle (M5S)).

# Respondent profile variables

| | |
|---|---|
| PERSONAL DATA | Gender<br>Age<br>Region<br>Education |
| FAMILY | Marital status<br>Children<br>Family unit composition<br>Socio-economic status |
| EMPLOYMENT | Employment status<br>Occupation<br>Sector of economic activity |
| CONSUMPTION HABITS | Decision makers<br>Purchases and consumption<br>Focus on automotive, financial/insurance services, power providers, holidays, food, household appliances, technology |
| LIFESTYLE | Travel<br>Sport<br>Volunteering<br>Reading habits<br>Internet use |

## Respondent opinions and attitudes

- What is your opinion on the **legalisation of soft drugs**?
- Do you think the **Islamic religion is a danger to society**?
- Will the **new generations be able to improve the world** in which they live?
- What is your **opinion on immigrants** in terms of job evaluation or crime?

Single Closed-ended questions:

- Single binary/categorical or Likert Scale answers.

# Data Description

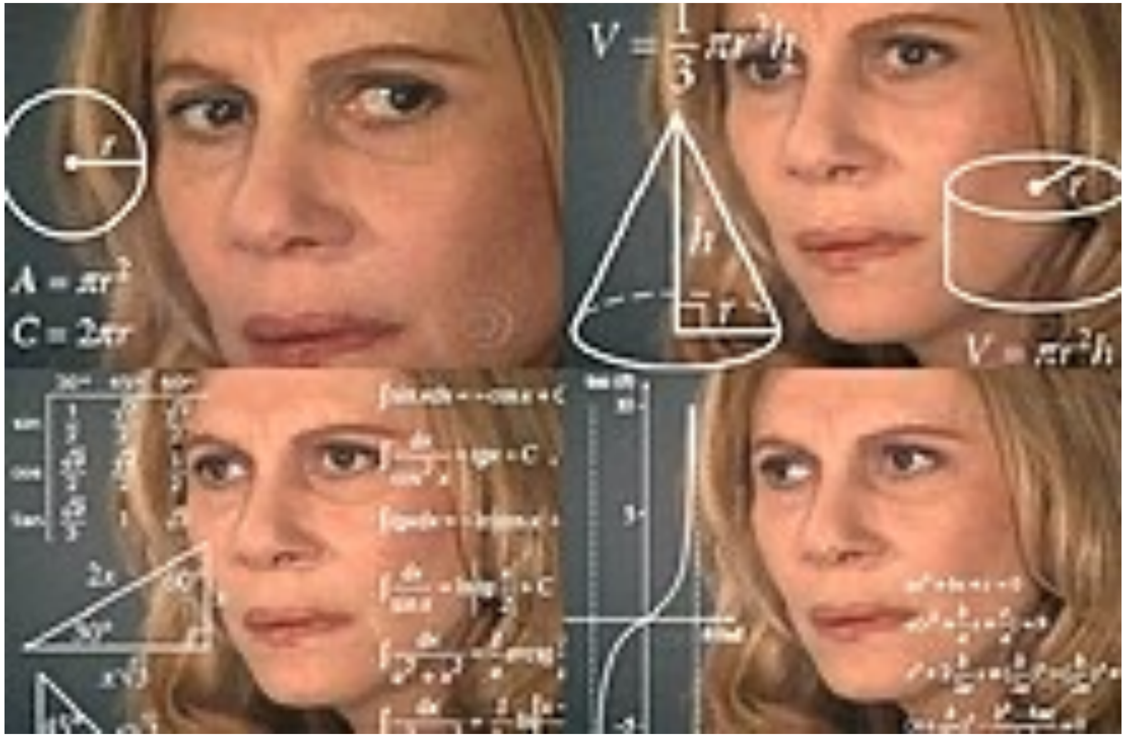
| Gender | Sample (%) | Population (%) |
| --- | --- | --- |
| Male | 50 | 49.7 |
| Female | 50 | 51.3 |

| Region | Sample (%) | Population (%) |
| --- | --- | --- |
| North-West | 27.0 | 26.8 |
| North-East | 20.0 | 19.4 |
| Center | 19.3 | 19.8 |
| South | 22.5 | 23.1 |
| Islands | 11.1 | 10.9 |

| Age Group | Sample (%) | Population (%) |
| --- | --- | --- |
| 15-24 | 7.8 | 11.4 |
| 25-34 | 15.1 | 12.7 |
| 35-44 | 18.9 | 15.7 |
| 45-54 | 20.0 | 18.9 |
| 55-64 | 16.5 | 16.0 |
| 65-90 | 21.2 | 22.6 |

| Education Level | Sample (%) | Population (%) |
| --- | --- | --- |
| Low (no title, primary, lower secondary) | 18.0 | 50.9 |
| Middle (upper secondary, post-secondary non-tertiary) | 47.4 | 35.1 |
| High (tertiary) | 35.0 | 14.1 |

## How do I convert a categorical variable into a number?




Any suggestions?

# Encoding

**One-hot Encoding**

| Gender | Location |
|---|---|
| Male | South |
| Female | North |
| Male | West |
| Male | East |

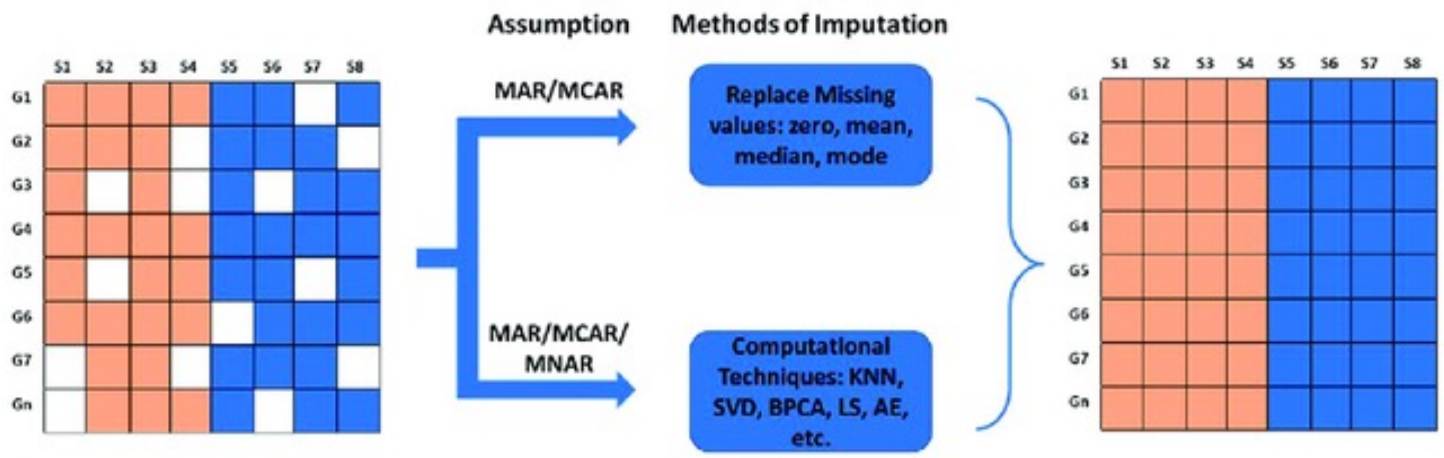
| Gender_Male | Location_North | Location_West | Location_East |
|---|---|---|---|
| 1 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 |
| 1 | 0 | 1 | 0 |
| 1 | 0 | 0 | 1 |

**Ordinal encoding**

| Original Encoding | Ordinal Encoding |
|---|---|
| Poor | 1 |
| Good | 2 |
| Very Good | 3 |
| Excellent | 4 |

Different variables names and modalities across the three years

# Imputation of missing values



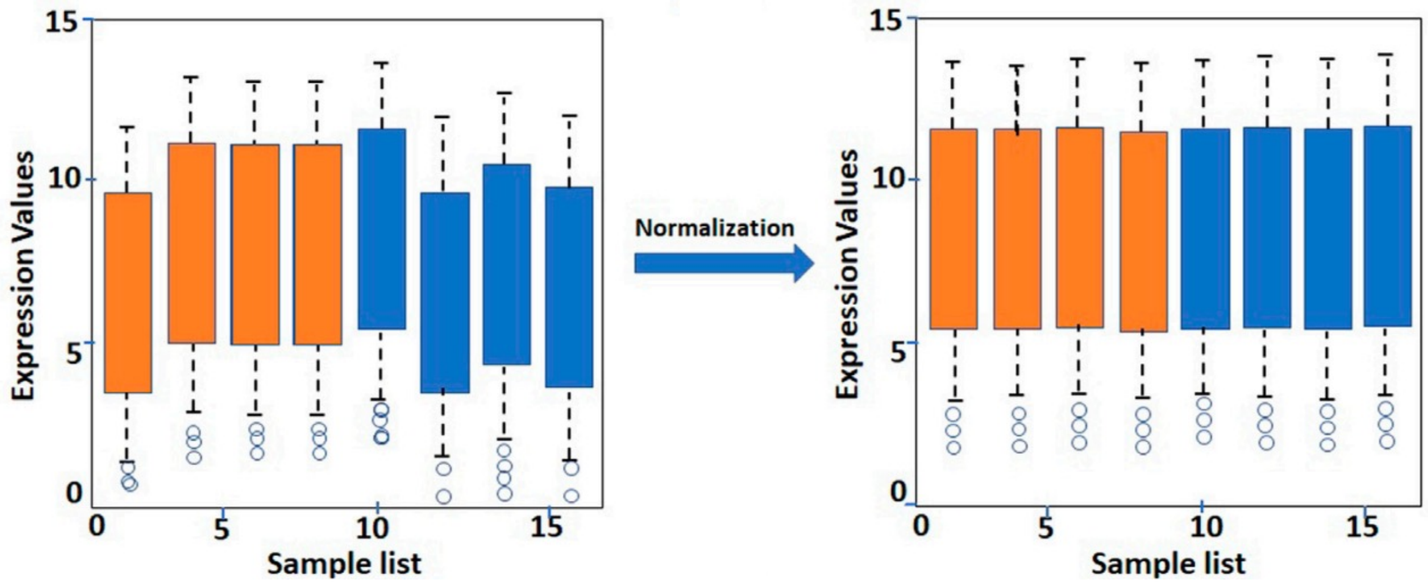

Missing data are typically grouped into three categories: **Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR).**

In **MCAR**, the missing data is **independent of their unobserved values and** independent of the **observed data**. In other words, the data is completely missing at random, independent of the nature of the investigation.

In **MAR**, the **missingness** of **data is random but conditionally dependent on observed and unobserved values**.

In **MNAR**, the **missingness depends on the observed and/or unobserved data**.

# Standardization




**Standardization** is crucial for classification tasks because it **ensures that all input features have a similar scale**. This **helps algorithms** (especially those based on distances, like K-Nearest Neighbors or SVM) work more effectively by **preventing larger values from dominating smaller ones.**

# Feature and rows selection


**Feature selection** in classification refers to the process of selecting the most relevant features from the dataset that contribute significantly to the target variable.
By removing irrelevant, redundant, or noisy features, feature selection improves usually model performance.

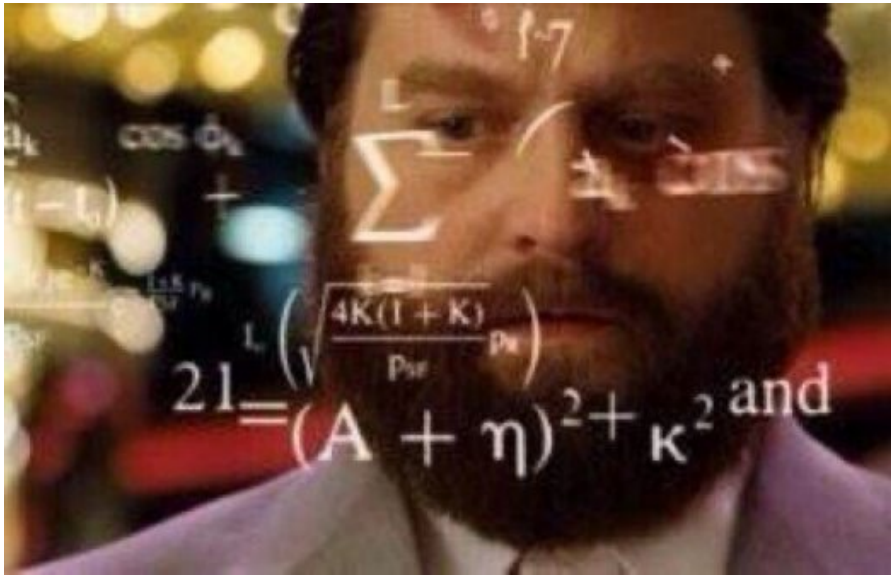
**Rows selection** how to deal with missing data

# The target variable – The voting intention!

| | |
|---|---|
| sono indeciso | 695 |
| Partito Democratico-PD | 608 |
| MoVimento 5 Stelle | 580 |
| Partito Democratico | 364 |
| Lega con Salvini | 299 |
| Movimento 5 stelle | 289 |
| Lega | 244 |
| Forza Italia | 214 |
| Lega Nord | 173 |
| non andrei a votare | 157 |
| preferisco non rispondere | 150 |
| Fratelli d'Italia | 112 |
| voterei scheda bianca / annullerei la scheda | 85 |
| Sinistra italiana (SEL + altri) | 67 |

| | |
|---|---|
| Liberi e Uguali | 56 |
| La Sinistra | 54 |
| Fratelli d'Italia-Alleanza Nazionale | 44 |
| +Europa | 44 |
| piu' Europa con Emma Bonino | 42 |
| Potere al Popolo | 39 |
| voterei scheda bianca / scheda nulla | 32 |
| Verdi | 30 |
| Rifondazione Comunista | 26 |
| Italia dei Valori | 9 |
| Noi con l'Italia UDC; | 8 |
| Nuovo Centro Destra con UDC e PPI | 8 |
| Scelta Civica; | 7 |
| altro partito di area di governo (SVP, Centro Democratico....) | 4 |
| Italia Unica di Corrado Passera | 2 |

# The target variable - Suggestion for managing the variables?




a small hint

## The target variable - behind the scenes

```
m_p_int_voto
Sx/CSx                          1348
Dx/CDx                          1102
M5S                              869
indecisi                         695
astensione/bianca/nulla          274
preferisco non rispondere        150
Altro partito                     66
```

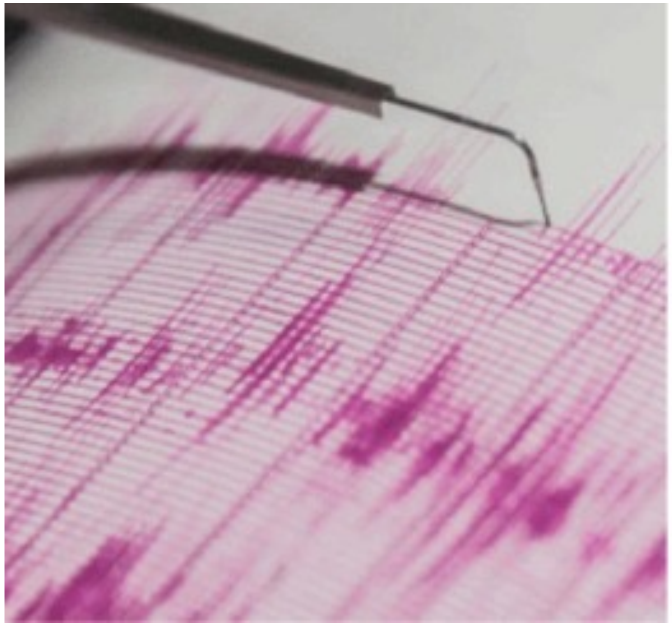
- Group each party into one of the **three main coalitions** (centre-left (Sx/Csx), centre-right (Dx/CDx) and Movimento 5 Stelle (M5S)) (inside the paper).

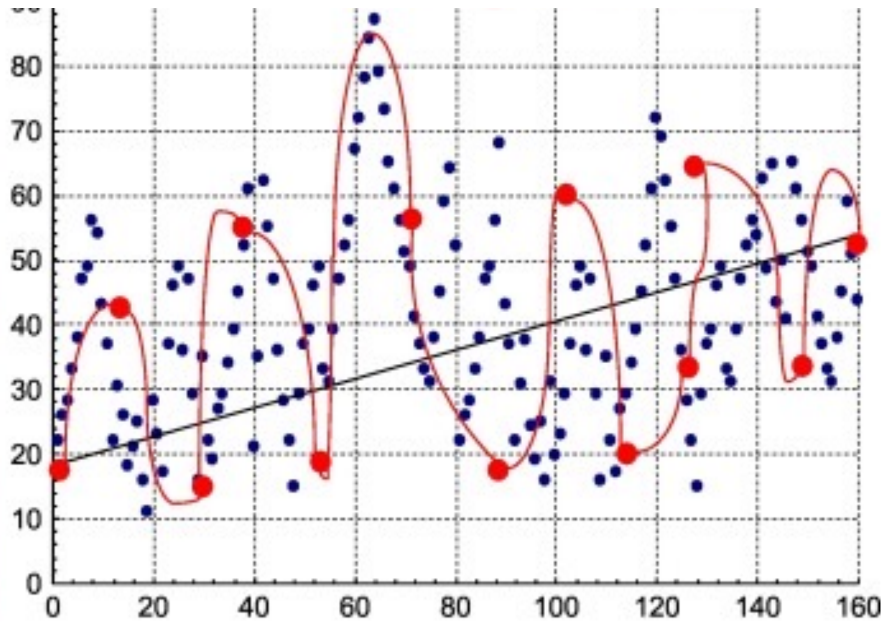
- We also try to classify and explain the absentees and undecideds, but it is very difficult, for this reason we decide to exclude these groups.

# Choosing a good ‘model’

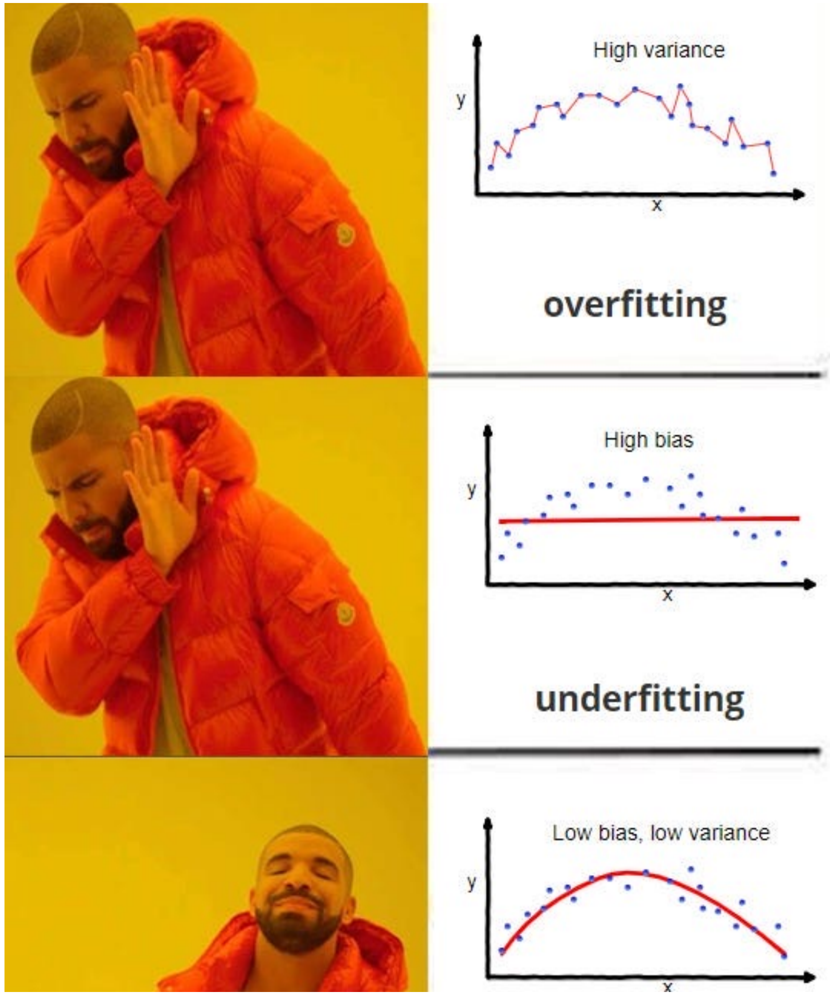
## Is a fitting line a good choice?

One needs to know priors on
the phenomenon studied

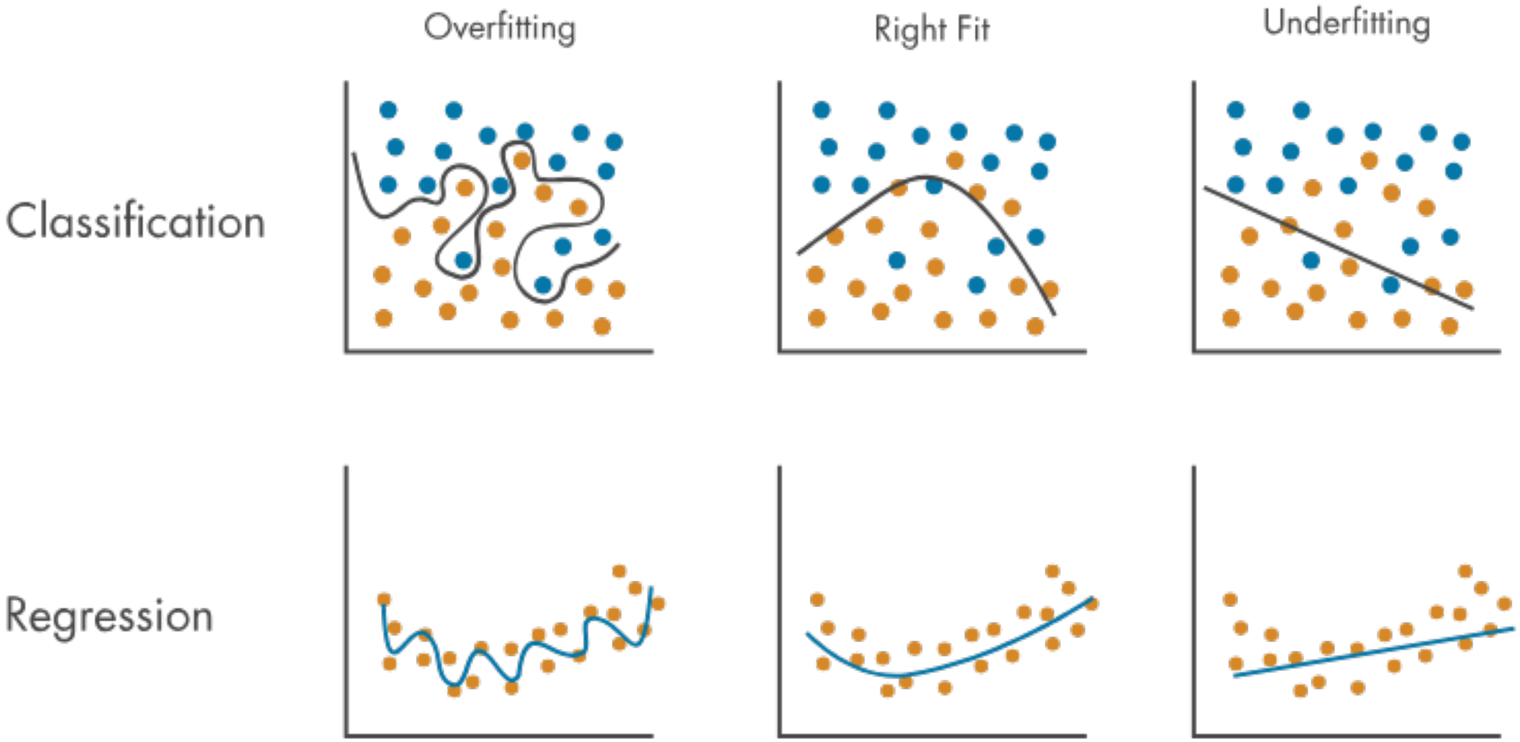




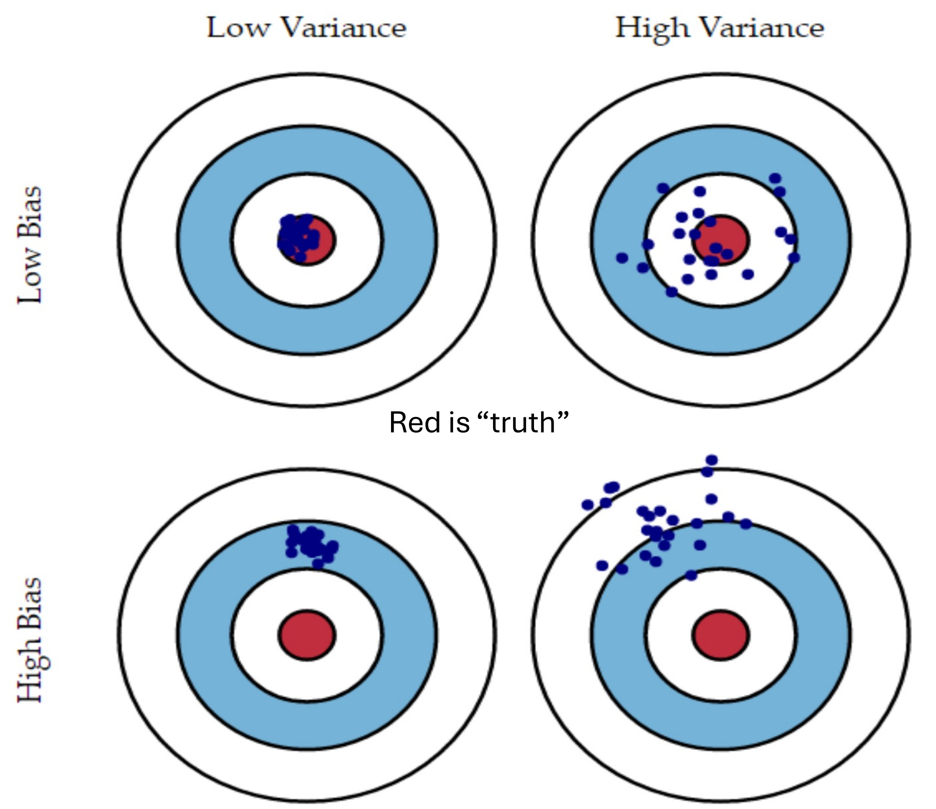
# Bias-variance tradeoff and model complexity



# Bias-variance tradeoff and model complexity

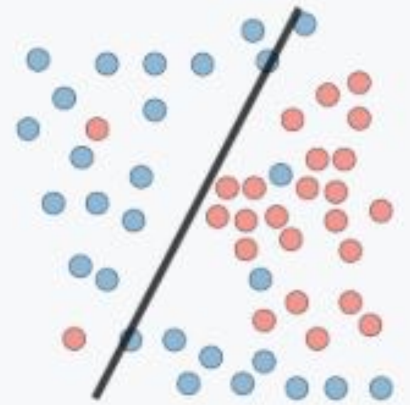
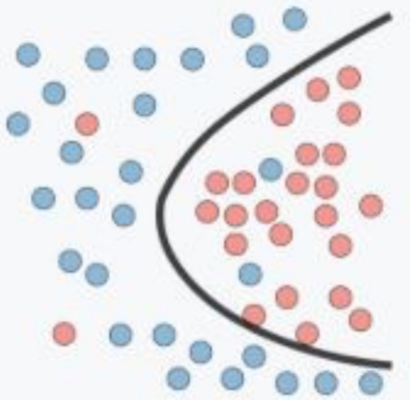
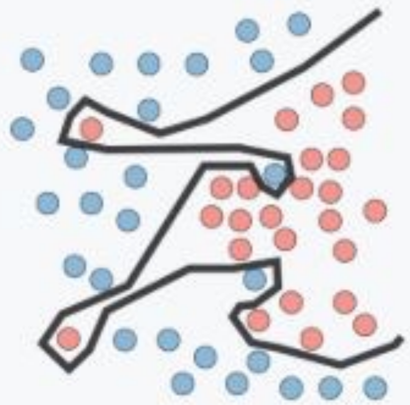


# Bias and Variance



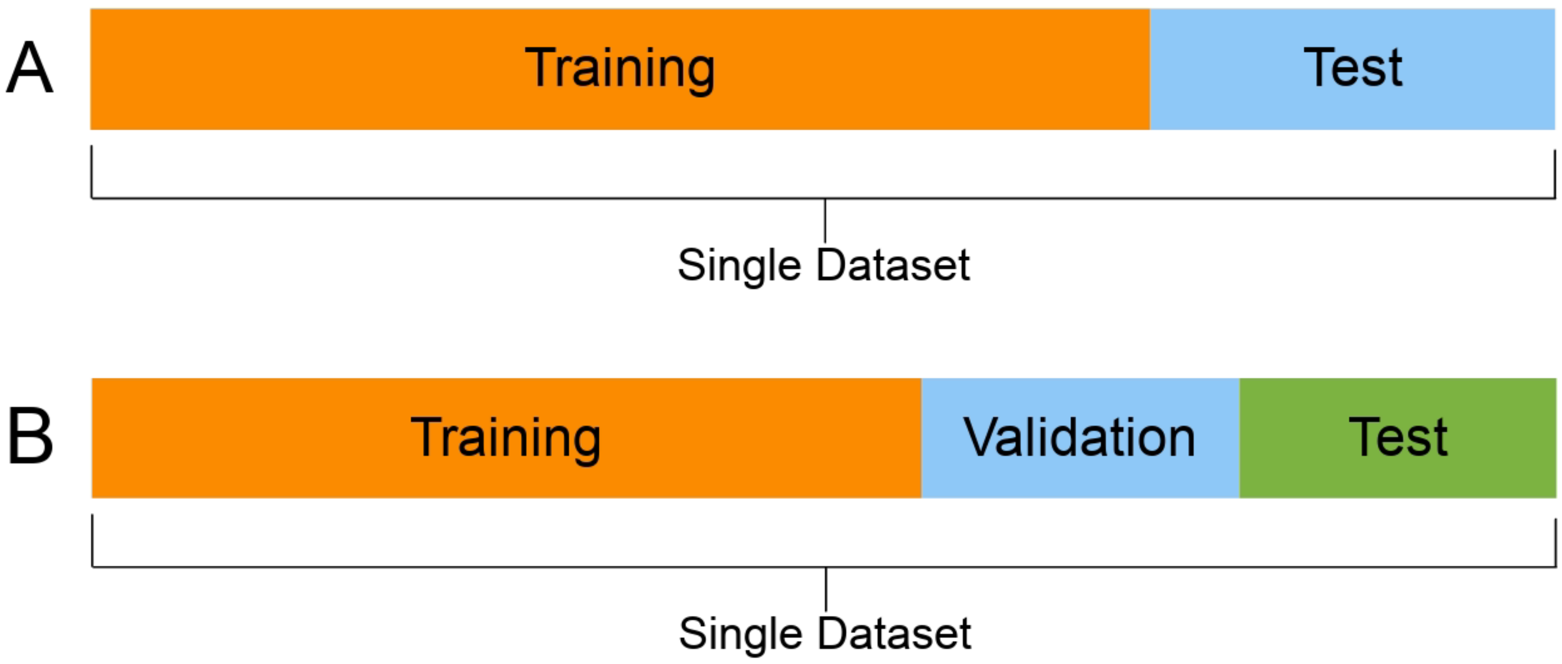

Red is “truth”

# Bias and Variance

| | Underfitting | Just right | Overfitting |
|---|---|---|---|
| **Symptoms** | • High training error<br>• Training error close to test error<br>• High bias | • Training error slightly lower than test error | • Very low training error<br>• Training error much lower than test error<br>• High variance |
| **Classification illustration** |  |  |  |
| **Possible remedies** | • Complexify model<br>• Add more features<br>• Train longer | | • Perform regularization<br>• Get more data |

# How ‘good’ is the model?

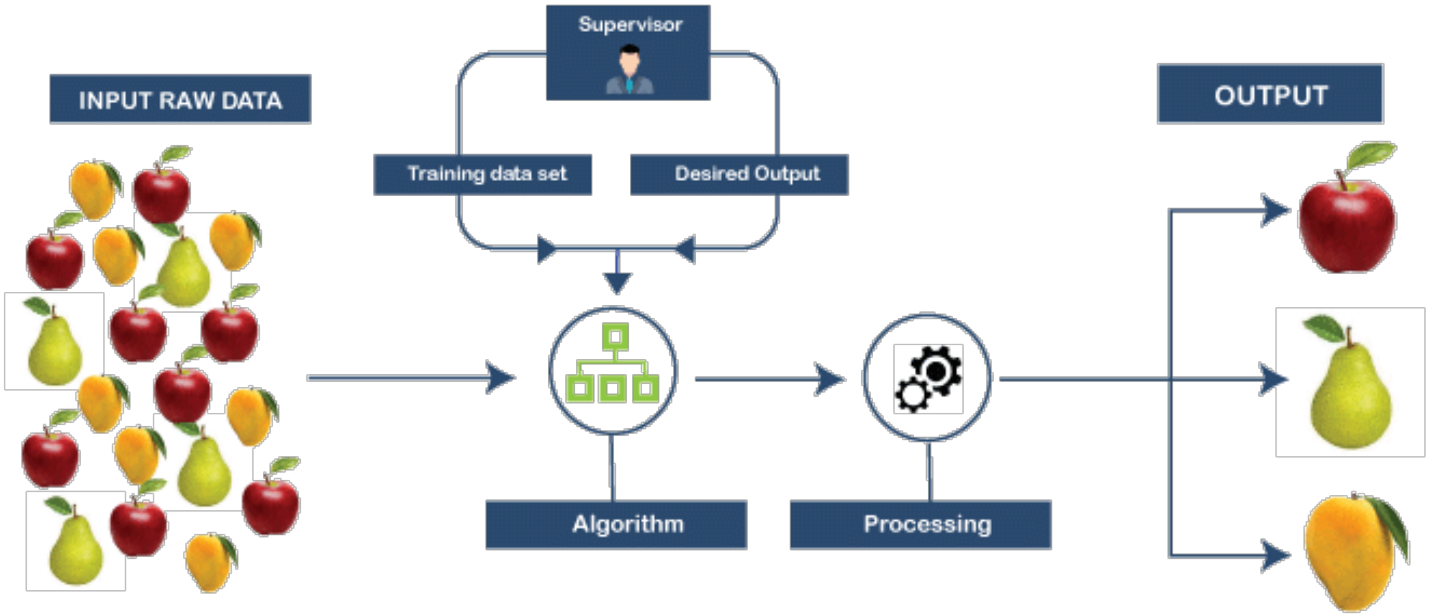
The ability to predict well is called **Generalization**



# Results

## Classifier

A **classifier** in machine learning is a model that takes data with features and predicts a label or category for each row.
For example, if you have a table of customer data, a classifier might predict whether each customer will "buy" or "not buy."



## Model Training

- We evaluated the performance of three classifiers: **Random Forest (RF), XGBoost and the Light Gradient Boosting Machine (LGBM).**
- The dataset is divided into an **80% training set and a 20% test set**, with stratification based on the target variable.
- **5-fold cross-validation** procedure.

# Result Model Training Step

A comparison of the result of the grid search of three classifiers

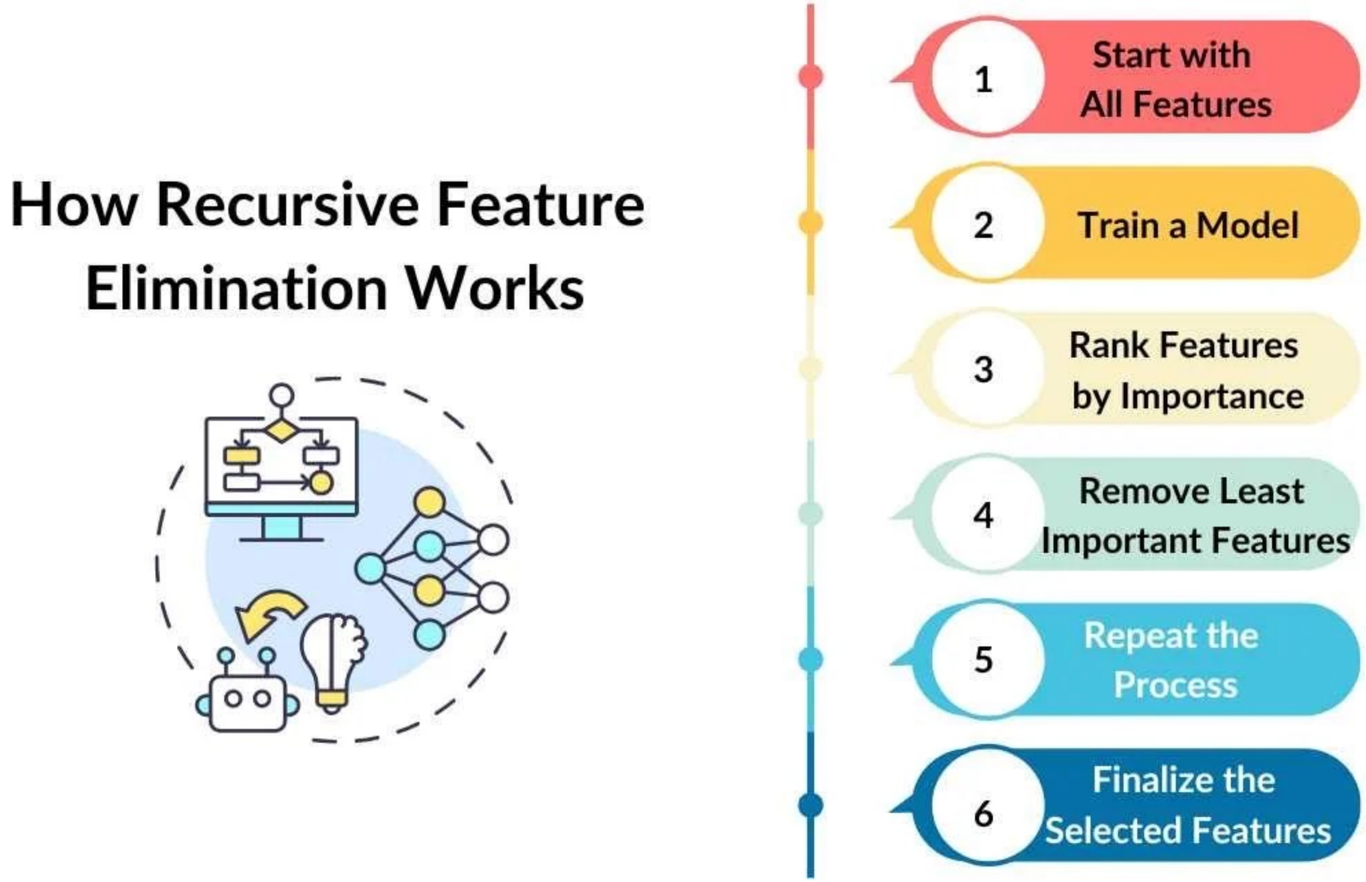
| Classifier | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| Random Forest | 0.68 | 0.68 | 0.68 | 0.66 |
| LGBM | 0.69 | 0.68 | 0.69 | 0.67 |
| XGBoost | 0.68 | 0.67 | 0.68 | 0.67 |

## Identification of the most relevant features

- We take **LGBM as the best classifier**.
- To select the best subset of variables, we used the **Recursive Feature Elimination (RFE) algorithm**.
- We obtain a final dataset with **71 variables**.

Best subset varaibles based on feature importance

# Features Selection



# Result

| | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|
| Best set | 0.70 | 0.70 | 0.70 | 0.69 |

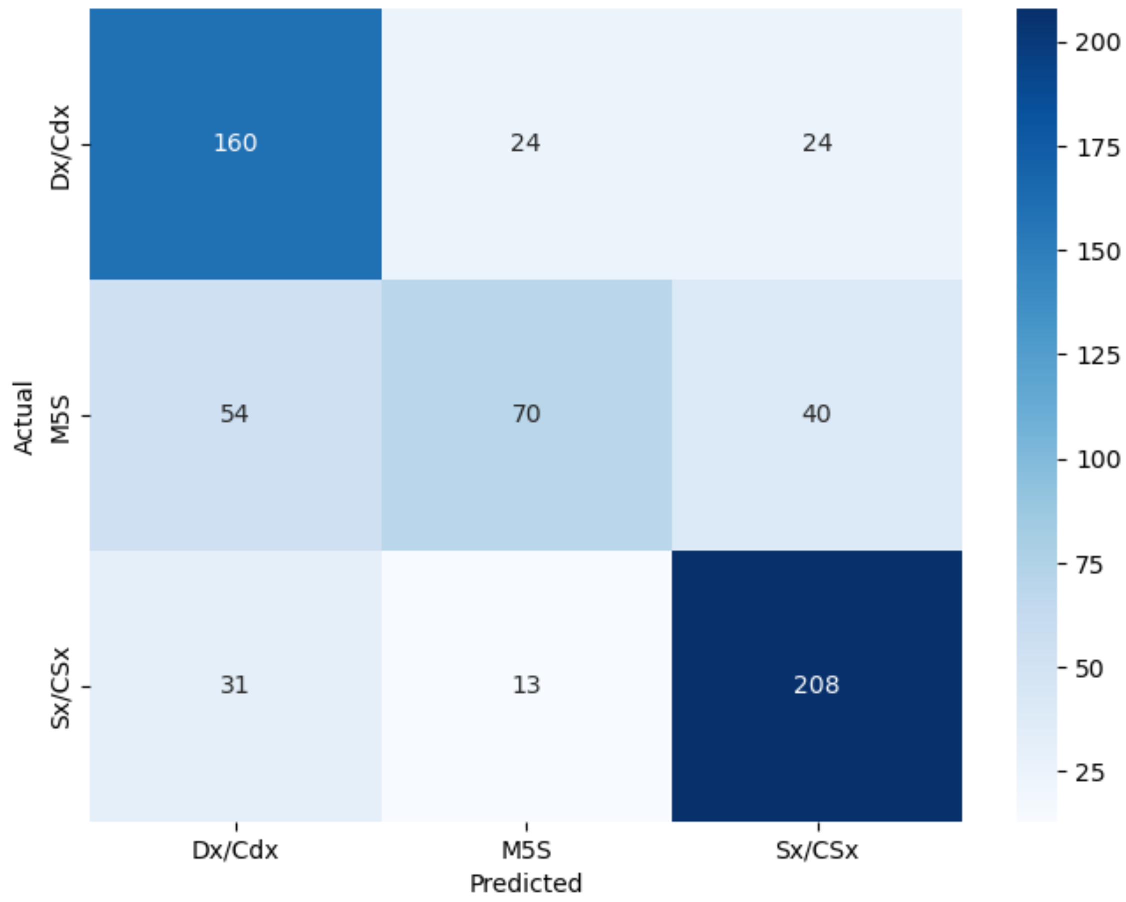
We can see how the best classifier-dataset pair performs better with only 71 features compared to the initial 120.

# Confusion Matrix

| | | Predicted | |
|---|---|---|---|
| | | Negative (**N**)<br>**-** | Positive (**P**)<br>**+** |
| **Actual** | Negative<br>**-** | True Negatives (T**N**) | False Positives (F**P**)<br>**Type I error** |
| | Positive<br>**+** | False Negatives (F**N**)<br>**Type II error** | True Positives (T**P**) |

Fonte: https://www.nbshare.io/

# Result



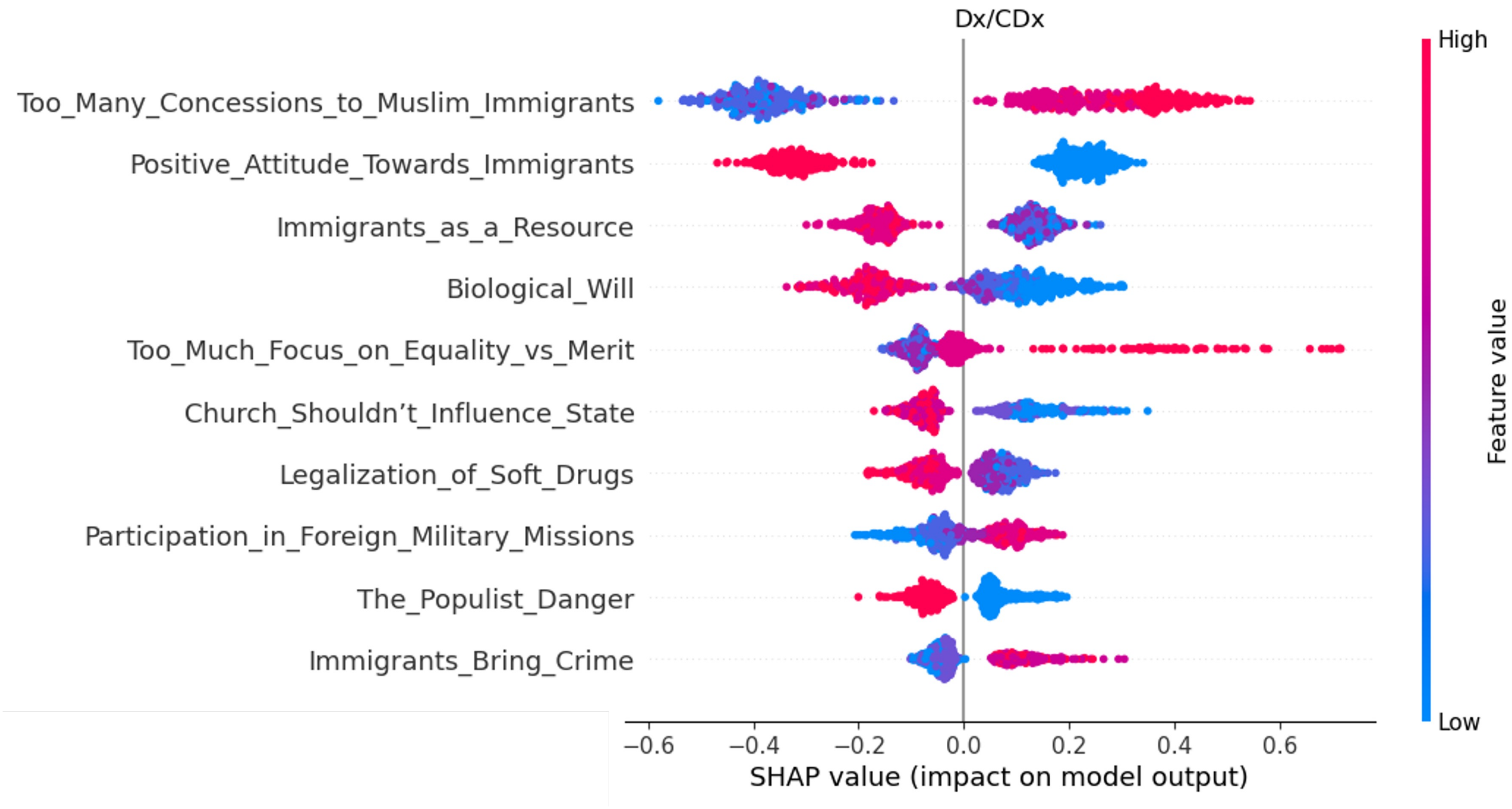
160
24
24
54
70
40
31
13
208
Actual
Dx/Cdx
M5S
Sx/CSx
Predicted
Dx/Cdx
M5S
Sx/CSx
200
175
150
125
100
75
50
25

## Shap Values

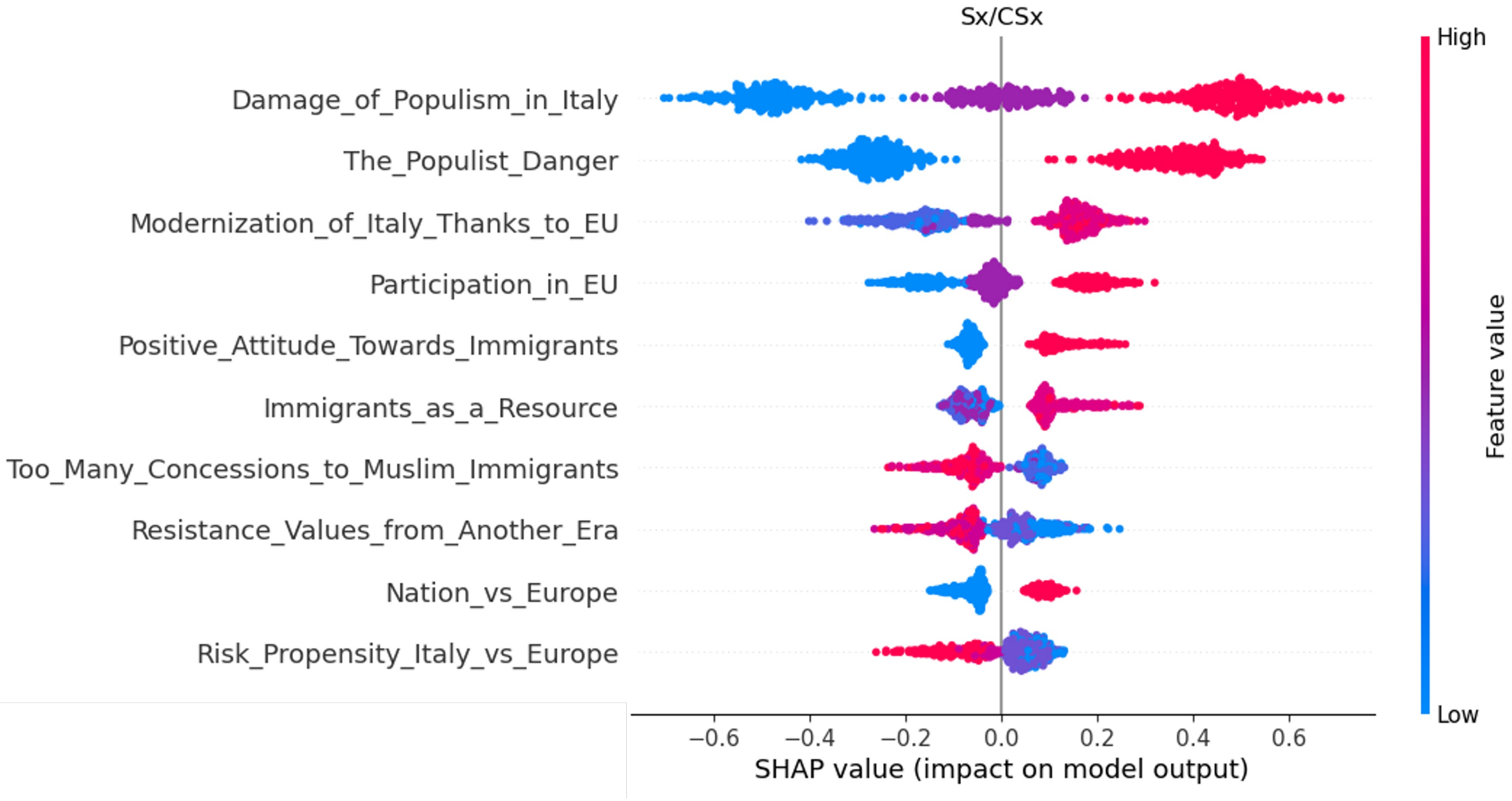
SHAP (SHapley Additive exPlanations) is a powerful tool in the machine learning world that draws its roots from game theory.
In simple terms, SHAP values allow you to break down a machine learning model's predictions by assigning each feature a "fair" contribution to the final output.

SHAP

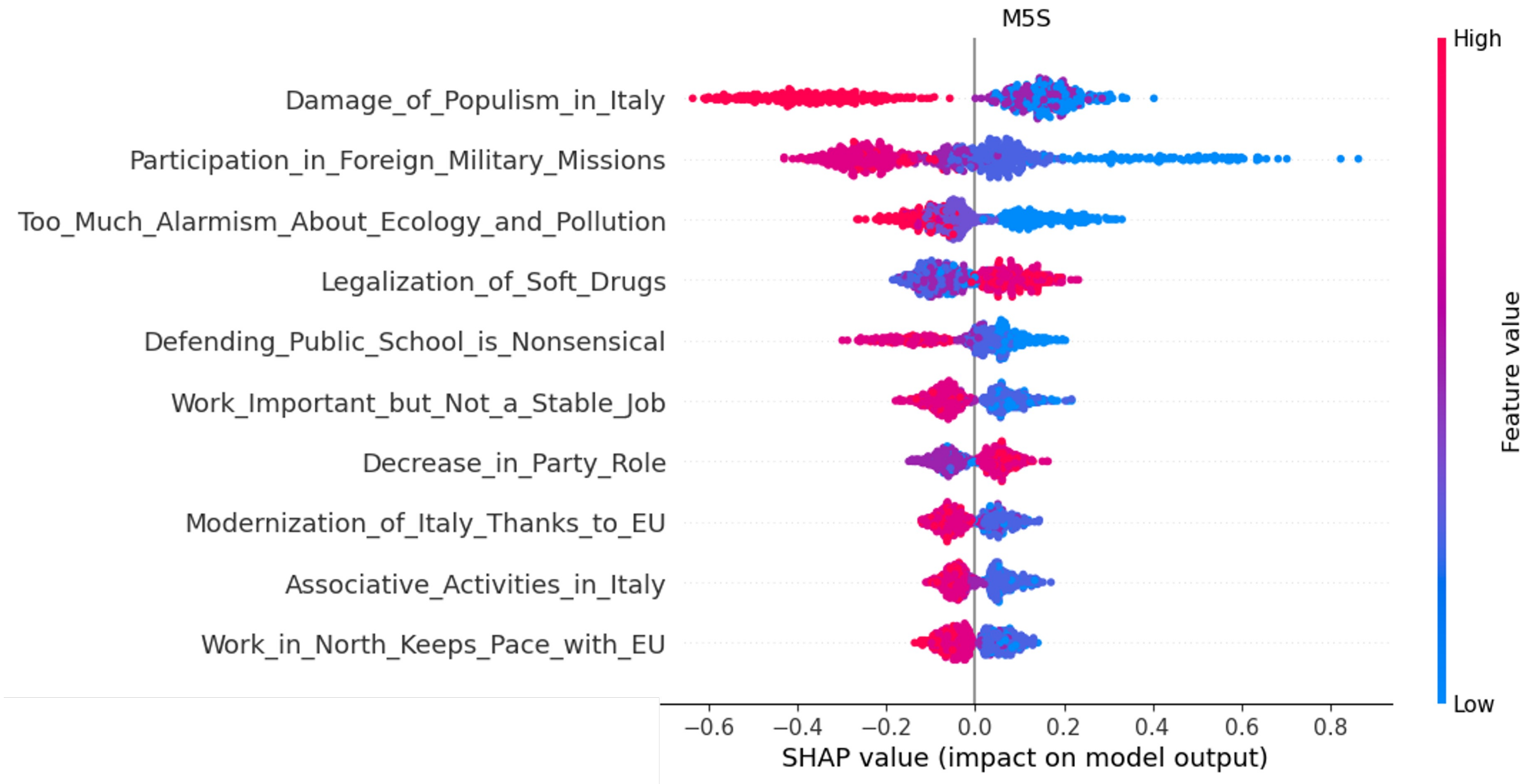
# Explainability- Dx/CDx



# Explainability- Sx/CSx



# Explainability- M5S



UNIVERSITÀ DEGLI STUDI DI TRIESTE

# Thank you for your attention!

Email: luca.pennella@phd.units.it
LinkedIn: https://www.linkedin.com/in/luca-pennella-4171a2192/