Linnæus University
UNIVERSIDAD DE MURCIA
Nantes Université
SEMMELWEIS UNIVERSITY 1769
euni well European University for Well-Being

ASTUTENESS PROJECT

AI-Driven Tools in Healthcare: A Visual Trustworthy Treatment Decision Support Systems

www.um.es/astuteness

SD for Machine Learning Explainability: Progress Report

Francisco José Mora Caselles
franciscojose.morac@um.es

February 2024

EUROPEAN UNIVERSITIES
Erasmus+
With the support of the Erasmus+ Programme of the European Union
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101035821.

Contents

Background

ANTIMICROBIAL RESISTANCE PROBLEM (ARP)
AR: the ability of microorganisms to become resistant to antibiotics.

EUROPE: ARP causes of 33,000 deaths/year and to be 1,500 M€ (ECDC)
Spain 2,500 deaths per annum, and an additional health expenditure of 150M€ /year [Spanish Agencey of Drugs report 2020]

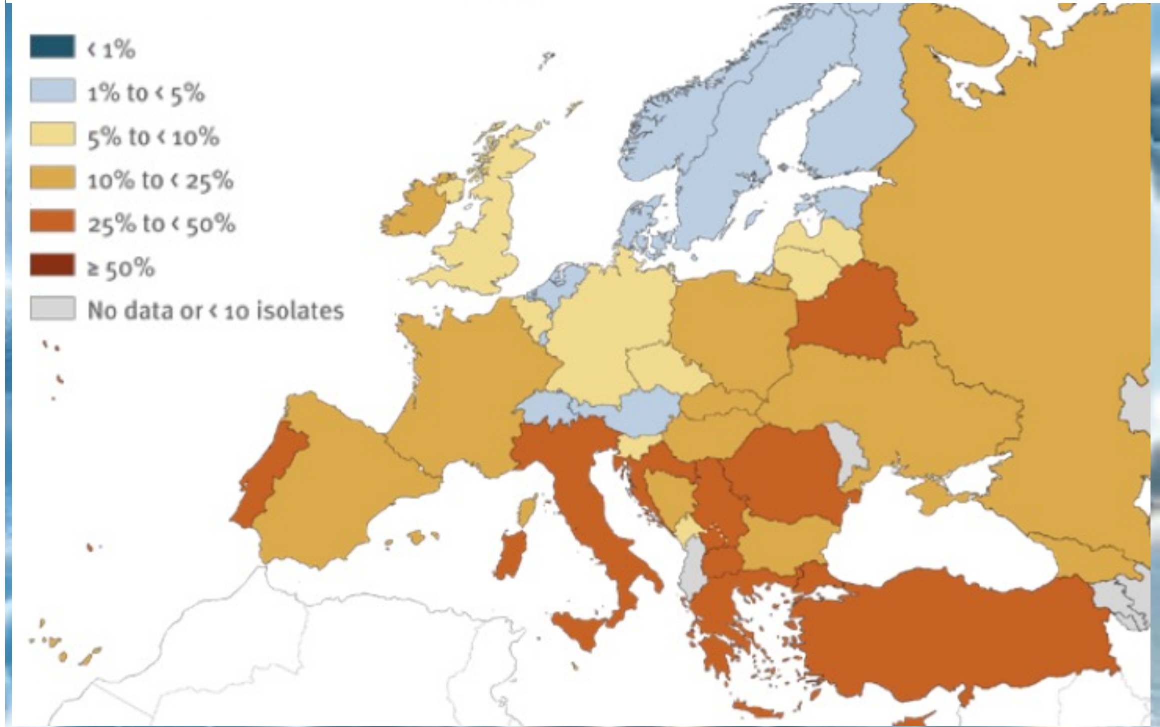
1 of the 6 priority strategic lines of the Spanish Plan against Antibiotic Resistance (PRAN): **surveillance of resistant bacteria** and the consumption of antibiotics in hospitals.

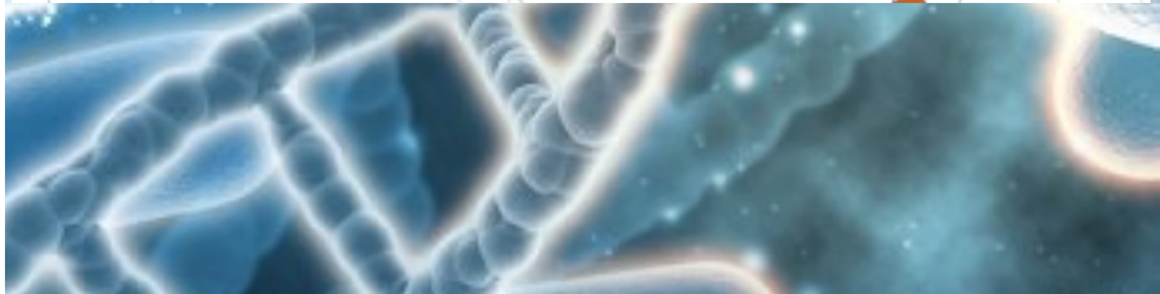
One of the main scenarios of this strategic line is the improvement to the **prescription of antibiotics.**

For this, we intend to predict the Minimum Inhibitory Concentration (**MIC**) of the bacteria to a given treatment using the data from their hospital stay.

Antimicrobial resistance surveillance in Europe 2022 by ECDC + WHO

S. aureus: percentage of invasive isolates resistant to methicillin (MRSA)





Picture (CC0) https://www.freepik.es/foto-gratis/adn-bacterias_923816.htm#query=BACTERIA&position=2&from_view=search&track=sph&uuid=cea4540f-c9b0-4bfe-974b-4839b2d99054

Dataset Open-Data MIMIC-III

MIMIC-III[1] ('Medical Information Mart for Intensive Care') is a large, single-center database of patients admitted to ICU at a large tertiary care hospital.

We use a smaller data **subset** containing information for cultures treated with **Vancomycin**.

Variables: patient gender and age, previous Vancomycin treatments, admission type and location, **culture_susceptibility,** etc.

[1] Johnson, A., Pollard, T., Shen, L. *et al.* MIMIC-III, a freely accessible critical care database. *Sci Data* **3**, 160035 (2016). https://doi.org/10.1038/sdata.2016.35

Our data

Our data consists of 531 instances of 26 variables.

We aim to predict the **culture susceptibility** (**R**esistant/**S**usceptible)

We observe that our data is highly unbalanced

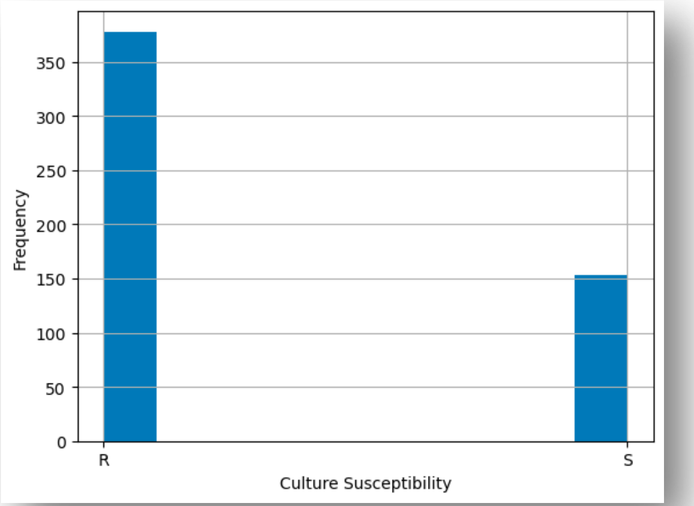
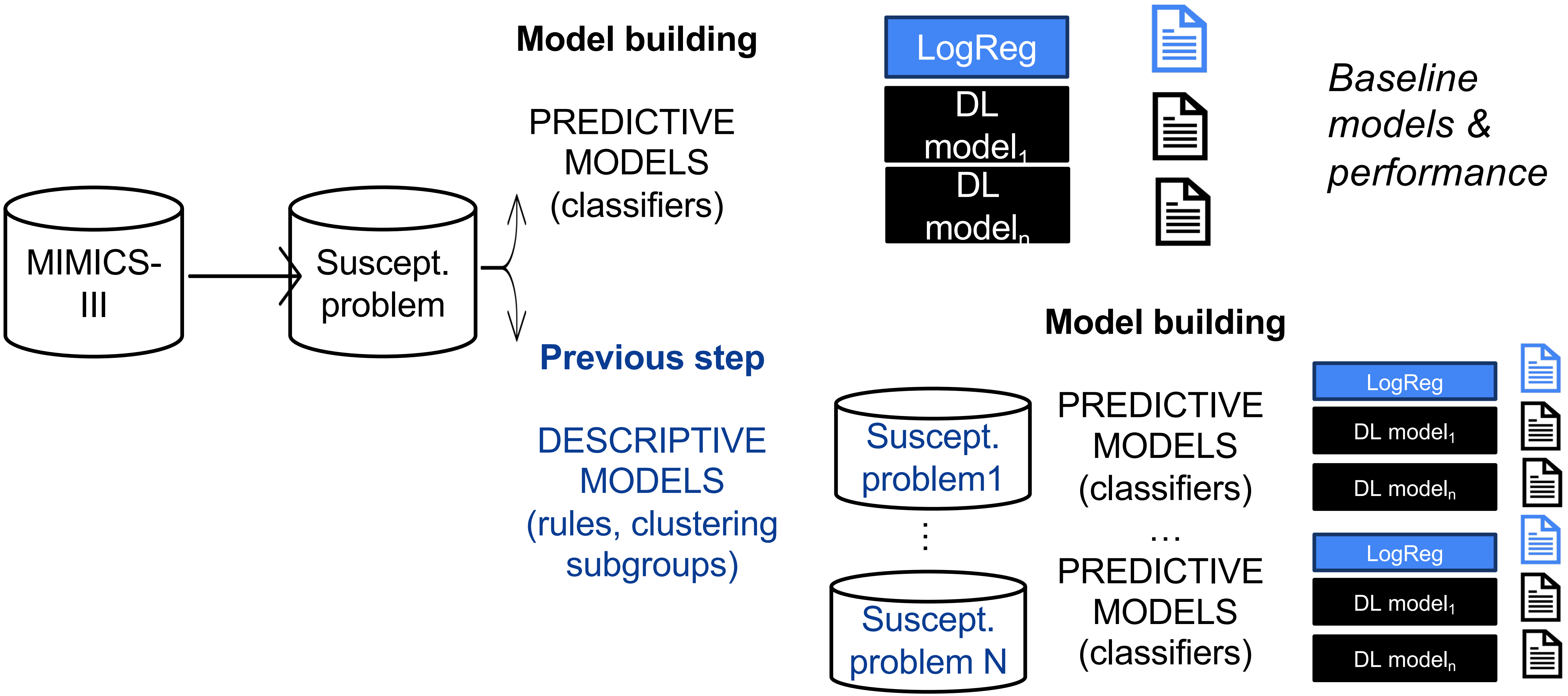
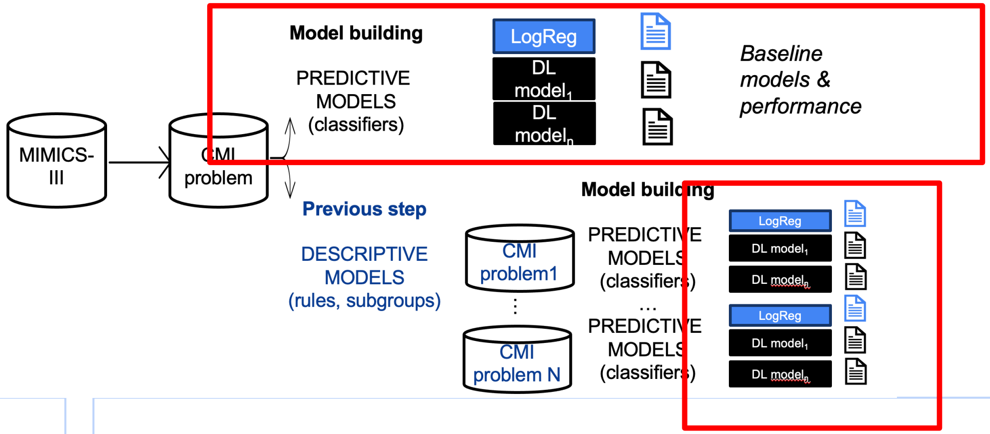


Figure 1: Culture susceptibility histogram

Methodology



Methodology. Predictive models



MODELS ANALYSIS

TESTING ML MODELS CAPACITY TO PRODUCE PREDICTORS

LOGIT

regularization applied

Tree-based & ensembles

RANDOM FOREST

deep trees (overfit individually)
- quality of split: gini
- samples to split node: 2
- depth: until leaves are pure
- samples to be leaf: 1

bootstrap to build trees

no class/individual weights

GRADIENT BOOSTING TREES

number estimators (stages) = 100

trees:
- samples to split node: 2 (quality split MSE)
- depth: max. 3 (weak classifiers)
- samples to be leaf: 1
- # leaf nodes: no limits

no class/individual weights

Kernel and NN-based

SVM CLASSIFIER

multi-class classification= 1vs1 scheme

kernel:
- rbf kernel
- max. degree function: 3
- coefficient gamma = 1 / (n_features * X.var())

no class weights

SIMPLE NNs

architecture

-1,2,3 layer

-16, 32, 64 neurons

-sigmoid, tanh, relu activation

-loss: binary cross entropy

grid search of hyperparameters

epochs=10

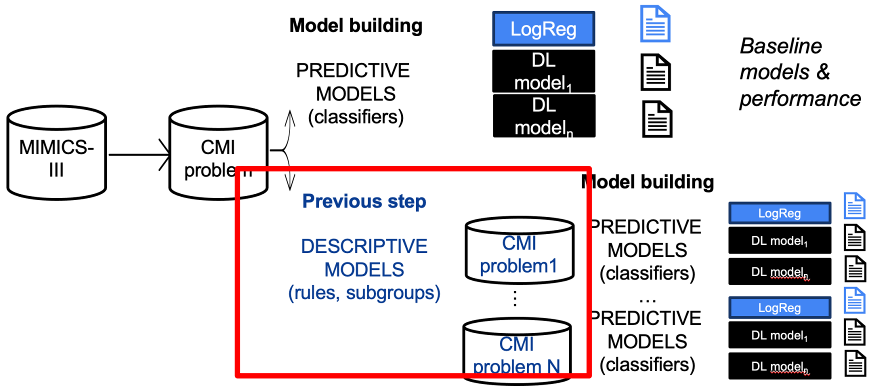
Results. Baseline models

- High accuracy for Resistant events, low accuracy for Susceptible events.
- All the models have similar behaviors.

	accuracy	specificity	sensitivity	f1	balanced_accuracy
LOGIT	0,8644	0,9974	0,5346	0,6853	0,766
R-forest	0,8494	0,9683	**0,5542**	0,6734	0,7612
GB	0,8531	0,9735	0,5538	0,6758	0,7636
SVM	0,8625	**1**	0,5221	0,6784	0,761
NN	**0,8682**	**1**	0,5412	**0,6937**	**0,7706**

Table 2: Baseline models results

Methodology. Descriptive Models



- Algorithm: BSD (define total amount of subgroups)
- Discretization of continuous variables using median as threshold
- Parameters:
  - Number of subgroups: 10
  - Quality measure: WRAcc and Qc
  - Minimum support: ~1/3 of total events in the target
  - Max depth: 6
  - Target: Response = Resistant and Response = Susceptible
- We generate a data subset using each rule (20 new dataset)

Results. Generated datasets

- For each rule, we generate a dataset containing only the instances that follow the rule.
- With this, we obtain 20 data subsets from the original dataset.
- We remove the categorical columns that appear in the rule (since they are now constant).
- The number of rows for each data subset is equal to the subgroup support (tp +fp)

Future work

- Deal with rows in multiple partitions
- Deal with rows not present in any partitions
- Apply conventional pattern mining techniques