# DOCTOR DISEASE

*A machine learning based solution to predict syphilis and diabetes diagnoses from medical records*

T24 | Jorge Acevedo, Charic Farinango, Juan M Gómez, Santiago Garcia, Daniel Montes, Cristian Prieto & Steven Ruiz

## Highlights

- Identification and cleaning of mistakes in medical notes.
- XGBoost classifier with 88% of F1-score to distinguish between syphilis and diabetes diagnoses and types based on medical notes and patient info.
- Web app build for users to load notes and obtain prediction and insights.

## Background

"1 in 5 patients who read a note reported finding a mistake and 40% perceived the mistake as serious".

We address this issue analyzing Electronic Health Records (EHR) provided by IQVIA to identify and clean mistakes, extract relevant information, and predict syphilis and diabetes diagnoses.

## Data and challenges

Highly imbalanced anonymized data of patients diagnosed with **syphilis** and **diabetes** in Colombia.

**Patients info**
9,306 patients
**82%** males
**77%** mestizos
**96%** urban pop
**60%** ABO missing

**Laboratory tests**
189,643 entries
**178** unique tests
Time series
Sep/01 - Mar/22

**Medical notes**
140,227 EHR
**81%** syphilis - 6 types
**19%** diabetes - 3 types
Typographical errors

## Methodology and results

1

### Feature Engineering

Find average and maximum differences between laboratory tests dates, and find most performed lab test for each patient.

2

### Data Preprocessing

Normalize and scale numerical data, one-hot encode categories, and remove stop words, tokenize, and TDF-IF transform text.

3

### Model training and selection

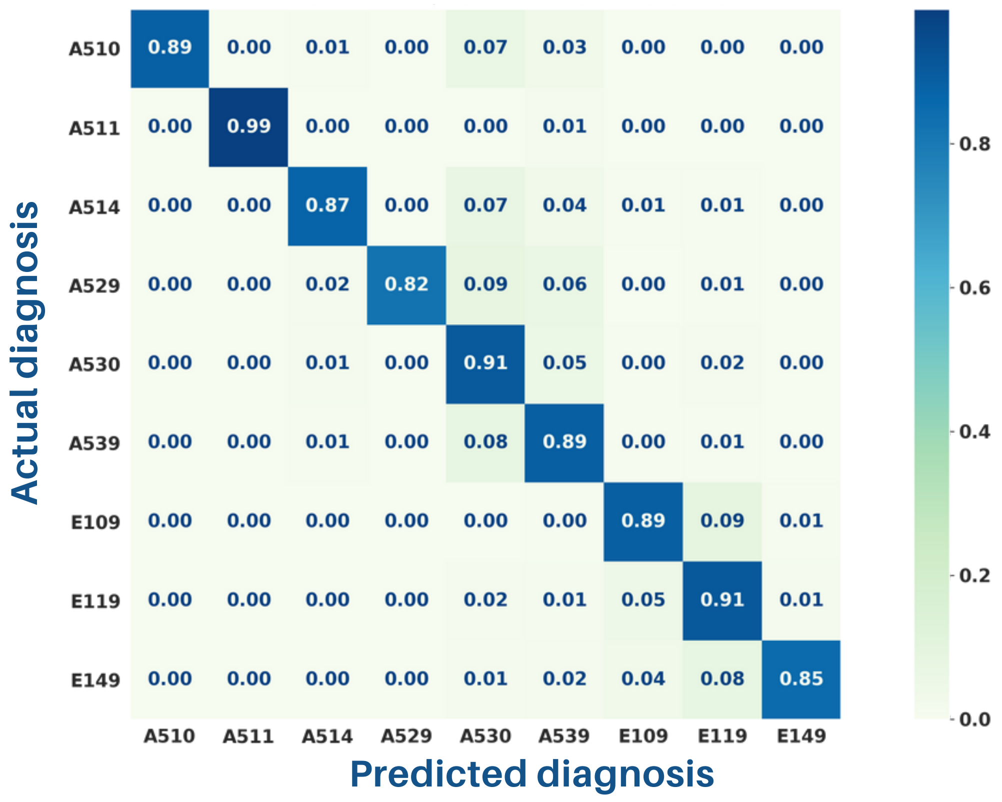
80-20% train-test split. Train different traditional machine learning classifiers and deep learning models.

4

### Optimize final model: XGBoost

Oversampling, feature selection using Elastic Net, and hyperparameter tuning using 5-fold CV.

**Confusion matrix, normalized on True labels**



| | A510 | A511 | A514 | A529 | A530 | A539 | E109 | E119 | E149 |
|---|---|---|---|---|---|---|---|---|---|
| A510 | 0.89 | 0.00 | 0.01 | 0.00 | 0.07 | 0.03 | 0.00 | 0.00 | 0.00 |
| A511 | 0.00 | 0.99 | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 |
| A514 | 0.00 | 0.00 | 0.87 | 0.00 | 0.07 | 0.04 | 0.01 | 0.01 | 0.00 |
| A529 | 0.00 | 0.00 | 0.02 | 0.82 | 0.09 | 0.06 | 0.00 | 0.01 | 0.00 |
| A530 | 0.00 | 0.00 | 0.01 | 0.00 | 0.91 | 0.05 | 0.00 | 0.02 | 0.00 |
| A539 | 0.00 | 0.00 | 0.01 | 0.00 | 0.08 | 0.89 | 0.00 | 0.01 | 0.00 |
| E109 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.89 | 0.09 | 0.01 |
| E119 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.01 | 0.05 | 0.91 | 0.01 |
| E149 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.02 | 0.04 | 0.08 | 0.85 |

**A5: Syphilis** (10 Primary genital, 11 Primary anal, 14 Other secondary, 29 Late, 30 Latent, 39 Unspecified)
**E1: Diabetes** (09 Type 1, 19 Type 2, 49 Unspecified)

# 88%

*of accuracy and F1 Score*

TRY OUR WEB APP

doctordisease.org