

# Estratificación poblacional según necesidades sanitarias futuras: Comparativa de modelos predictivos para el País Vasco

Alexander Olza<sup>1</sup>   Eduardo Millán<sup>23</sup>  
María Xosé Rodríguez-Álvarez<sup>41</sup>

<sup>1</sup>BCAM- Basque Center for Applied Mathematics.

<sup>2</sup>Osakidetza Servicio Vasco de Salud, Departamento de Salud.

<sup>3</sup>Kronikgune, Instituto de Investigación en Servicios de Salud.

<sup>4</sup>Dept. Statistics and OR, Universidade de Vigo.

JEDE 2020 (Octubre de 2021)

# Índice

## 1. Objetivos

## 2. Metodología

- Datos
- Las tres fases
- Métodos

## 3. Resultados

- Discriminación
- Calibración

## 4. Conclusiones

# Objetivos

Proyecto en marcha desde 2014

## Objetivo principal

Facilitar una atención sanitaria **proactiva y eficiente** capaz de **prevenir el deterioro** de la salud.

Envejecimiento poblacional  $\Rightarrow$  Mayor tasa de multimorbilidad.

Modelos predictivos para:

- Detectar pacientes de riesgo
- Facilitar la labor en Atención Primaria
- Facilitar la coordinación asistencial
- Sugerir intervenciones preventivas

**Enfoque:** Predecir la probabilidad de Ingreso Hospitalario Urgente (IHU).

# Datos

Población del País Vasco (2016):  $N = 2240526$

- Variables independientes (2016, 190 categóricas):
  - **Sociodemográficas:** Edad, Sexo, Índice de privación
  - **Clínicas:** Extraídas de la historia clínica (Johns Hopkins ACG System)
    - ACG: Grupos de isomorbilidad (excluyentes)
    - EDC: Diagnósticos
    - Rx-Mg: Prescripciones de farmacia
- Variable dependiente: **Algún ingreso urgente** en 2017  $y \in \{0, 1\}$

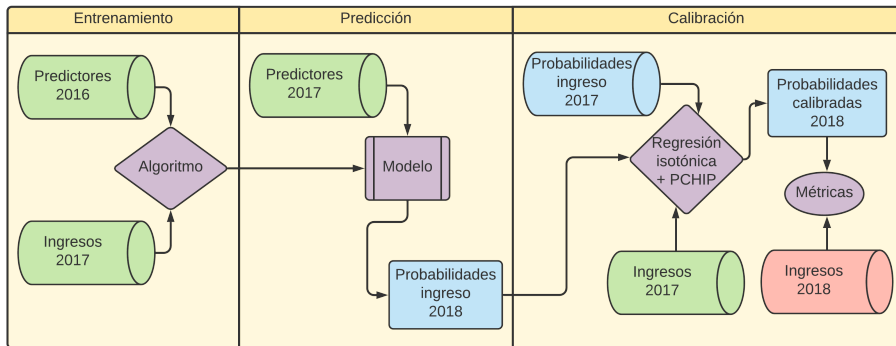
## Ejemplos

**ACG 4910:** 6-9 Diagnósticos, Edad > 34, 0-1 Diagnósticos graves

**EDC FRE03:** Endometriosis

**Rx-Mg PSYx030:** Psicosocial/Ansiedad

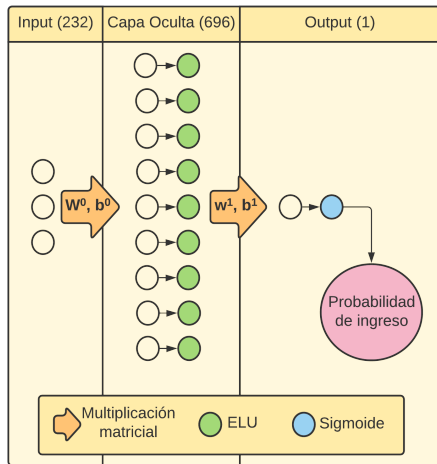
# Metodología en tres fases



# Métodos

Método	Búsqueda de hiperparámetros
Regresión logística	Sin regularización (Orueta et al., 2014)
Conjuntos de árboles	<div>3-fold CV</div> <ul style="list-style-type: none"> <li>• Random forest</li> <li>• Gradient Boosted Decision Trees</li> <li>• Adaboost</li> </ul>
Red neuronal	Experimentación manual

# Sobre la red neuronal



$$\hat{p}_i = \sigma(w^1 \cdot ELU(W^0 \cdot x_i + b^0) + b^1)$$

$$ELU(z) = \begin{cases} z & z \geq 0 \\ e^z - 1 & z < 0 \end{cases}$$

- $\sigma$ : Función sigmoide
- $x_i$ : Predictores
- $W^0, w^1, b^0$  y  $b^1$ : Parámetros

## Discriminación - Métricas

*Los médicos recibirán un **aviso** sobre los **top K** pacientes de riesgo.*

- Analizamos  $K = 20000$ : Aprox. 10 pacientes por médico
- Prevalencia IHU:  $|I| \sim 100000$  personas/año

$$\text{Sens@K} = \frac{|L_K \cap I|}{|I|} \times 100$$

$$\text{PPV@K} = \frac{|L_K \cap I|}{K} \times 100$$

$L_K$ : Listado de  $K$  pacientes con mayor  $p_i$ .

$I$ : Listado de pacientes con algún ingreso urgente.



## Discriminación - Resultados

**Cuadro:** Rendimiento con  $K = 20000$ . Cutoff@K: mínima  $p_i$  incluida en el top K.  $\Delta C$ : incremento de predicciones correctas respecto a regresión logística. Métricas evaluadas con datos de 2018.

	<b>AUC</b>	<b>Sens@K</b>	<b>PPV@K</b>	<b>Cutoff@K</b>	<b><math>\Delta C</math></b>
Red neuronal	0.815	10.77	54.92	0.44	1268
Gradient-Boosted trees	0.813	10.43	53.21	0.46	584
Random forest	0.811	10.26	52.34	0.45	236
Regresión logística	0.805	10.14	51.75	0.49	0
AdaBoost	0.805	9.46	48.25	0.47	-1400

# Calibración - Importancia

## Definición

**Clasificador calibrado:** Sus predicciones se corresponden con la verdadera distribución de probabilidad para cada clase.

- ▲ Fundamental para toma de decisiones
- ▲ Permite estratificación poblacional → Incluir  $p_i$  en historia clínica
- ▼ Algoritmos de ML no aseguran calibración → Necesitan paso adicional
  - Especialmente árboles
  - En menor medida regresión logística y red neuronal

## Calibración - Proceso

$$\left. \begin{array}{l} \text{Clase verdadera } y \in \{0, 1\} \\ \text{Probabilidad predicha } \hat{p} \end{array} \right\} \xrightarrow{f} p_i = f(\hat{p}_i)$$

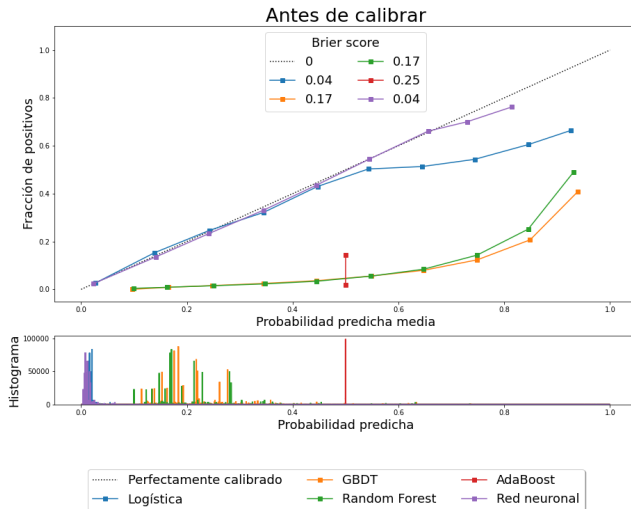
Requerimientos:

- **Monótona** para preservar discriminación
- **Suave** para permitir estratificación

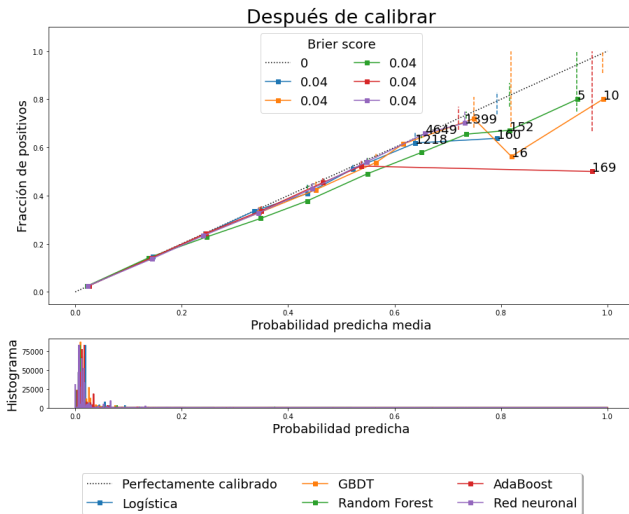
**Regresión Isotónica**  
+  
**PCHIP**

Piecewise  
Cubic  
Hermite  
Interpolating  
Polynomial

# Calibración - Resultados



# Calibración - Resultados



## Conclusiones

- La atención sanitaria proactiva requiere anticipar el deterioro de la salud.
- Los modelos deben calibrarse mediante función suave.
- Ingreso urgente en el País Vasco: Red neuronal, GBDT y random forest superan a la regresión logística.
- Sigüientes pasos: Inclusión de nuevas variables.

## Bibliografía



J. Bröcker y L. A. Smith. *Increasing the Reliability of Reliability Diagrams*. Vol. 22:3. Weather & Forecasting, AMS, 2001, págs. 651-661.



X. Jiang et al. *Smooth Isotonic Regression: A New Method to Calibrate Predictive Models*. AMIA Summits on Translational Science proceedings, 2011, págs. 16-20.



J. F. Orueta et al. *Desarrollo de un modelo de predicción de riesgo de hospitalizaciones no programadas en el País Vasco*. Vol. 88. Revista Española Salud Pública, 2014, págs. 251-260.

🐦 → @AlexOlza - @BCAMBilbao

## Apéndice - Diagrama de confiabilidad

- $B_k$ : Individuos con  $p_i$  en cierto intervalo
- $f_k$ : Fracción de positivos en  $B_k$
- $I_k$ : población en  $B_k$
- $r_k$ : Predicción media en  $B_k$

Bajo hipótesis de calibración,

$$f_k \sim \text{Bin}(I_k, r_k)$$

Bröcker y Smith (2001): Calculan las variaciones de  $f_k$  para un conjunto de respuestas binarias perfectamente calibradas, construidas artificialmente.



## Apéndice - Necesidad de suavizado

**Cuadro:** Rango del número de probabilidades predichas únicas después de la calibración para los distintos métodos.

	<b>Sólo reg. isotónica</b>	<b>Reg. isotónica + PCHIP</b>
Valores únicos	300-410	753000-841000

## Apéndice - Hiperparámetros

- **Random forest:** 800 árboles, máxima profundidad 100
- **GBDT:** 2500 iteraciones, máxima profundidad 20, tasa de aprendizaje 0.01, regularización L2 0.5.
- **AdaBoost:** 3315 árboles de profundidad 1, tasa de aprendizaje 0.1.