

Aprendizaje Automático Para La Clasificación De Pacientes Crónicos Con Comorbilidades

Autora: Paula Vecino Rodríguez
Tutora: Inmaculada Mora Jiménez
Co-Tutora: Cristina Soguero Ruíz



Universidad
Rey Juan Carlos



TESIS

Objetivo

Evaluación de métodos de aprendizaje automático multi-clase y multi-etiqueta para predecir el estado de salud de pacientes crónicos con comorbilidades

1. Introducción
2. Métodos de clasificación multi-clase y multi-etiqueta
3. Experimentos y resultados
4. Conclusiones y líneas futuras

Introducción

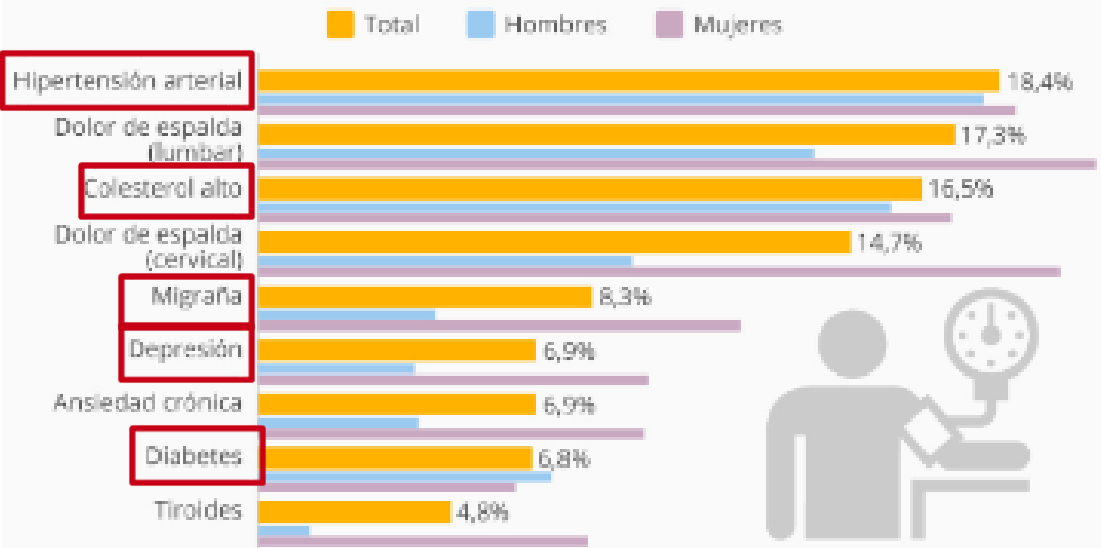
1. Pacientes crónicos. Situación en España
2. Sistema de clasificación poblacional
3. Análisis descriptivo de la base de datos

Pacientes crónicos. Situación en España

- Según la Organización Mundial de la Salud, se define una patología crónica como una *enfermedad de larga duración, normalmente de evolución lenta, y cuya curación no se puede prever.*
- Factores de riesgo: envejecimiento, elevado consumo de alcohol, tabaco y sal e inactividad física, entre otros.
- 42% de la población adulta de España, tiene una o varias patologías crónicas.

Enfermedades crónicas en España

Porcentaje de españoles que sufren distintos problemas de salud crónicos*

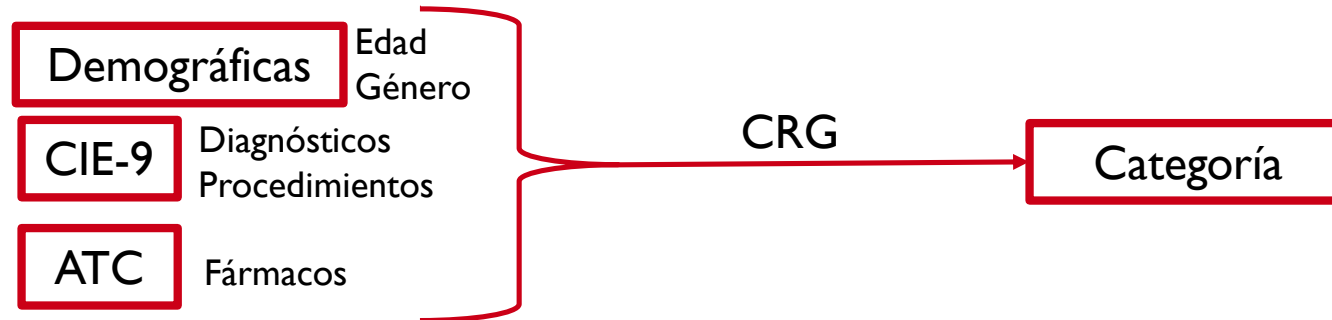


* Personas mayores de 14 años. Datos de 2014. Selección de problemas de salud.



Fuentes: INE, Ministerio de Sanidad, Servicios Sociales e Igualdad

Sistemas de clasificación poblacional



CRG base	Descripción del CRG	Pacientes HUF
1000	Sanos	46.835
5192	Hipertensión	12.447
5424	Diabetes	2.166
6144	Diabetes - Hipertensión	3.179
7071	Diabetes - Hipertensión - Otra Enfermedad Crónica Dominante	547

- Nuestra base de datos está formada por pacientes asignados al Hospital Universitario de Fuenlabrada en el periodo 2012.

Edad	Género	Diagnósticos				Fármacos
		001-999 Enfermedades	V01-V89 Códigos V	E000-E999 Códigos E	M888-M997 Códigos M	A01AA-V30ZZ ATC

- 2 características demográficas, 2263 características clínicas. La características clínicas pueden estar codificadas de forma binaria o numérica.

Análisis descriptivo de la base de datos

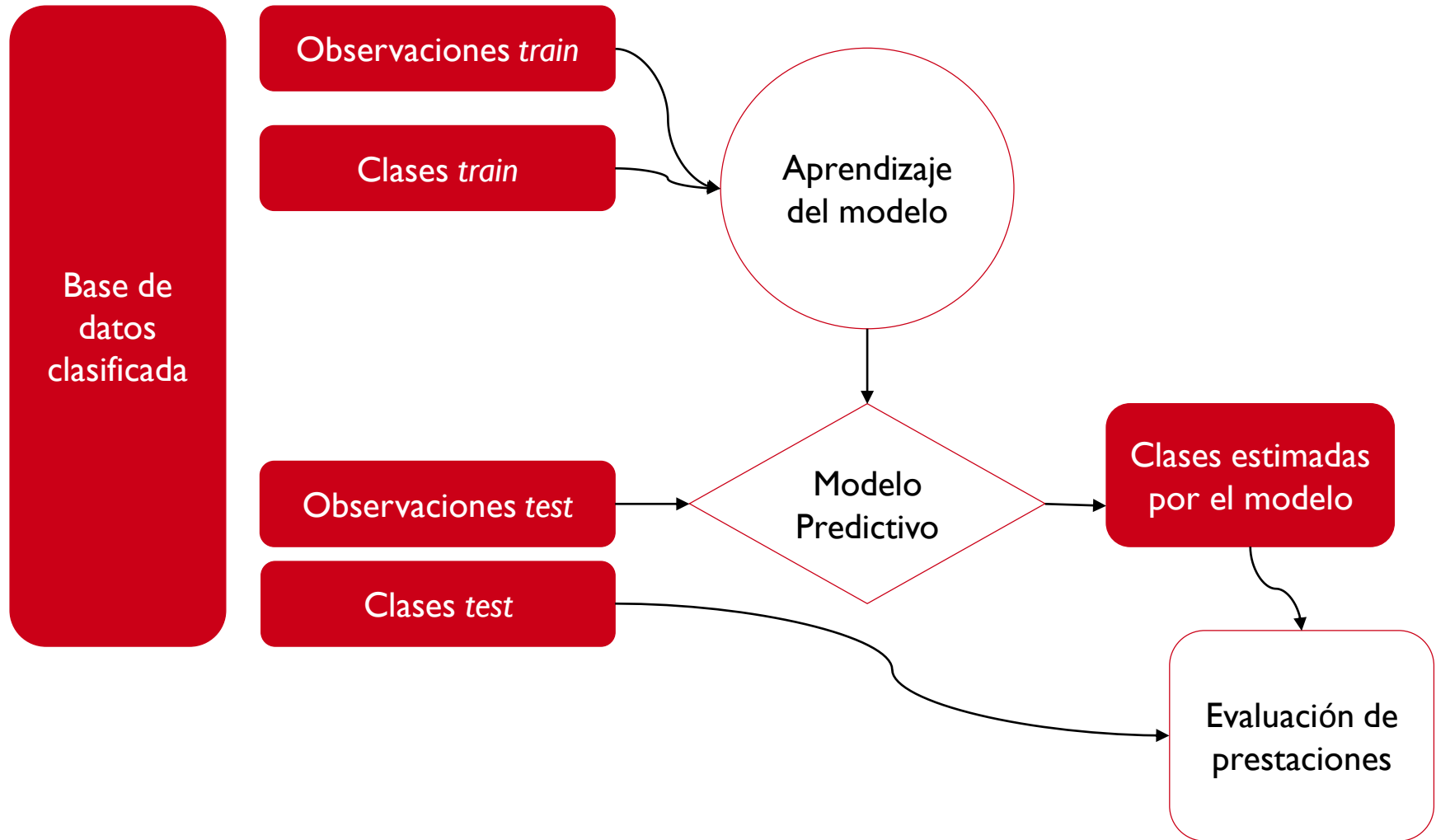
- Análisis descriptivo de los CRG utilizados en este trabajo:

	CRG1000	CRG5192	CRG5424	CRG6144	CRG7071
Pacientes en el CRG	46835	12447	2166	3179	547
Mujeres	24649	6464	743	1424	305
Hombres	22186	5983	743	1755	242
Códigos CIE-9 Totales	1517				
Códigos CIE-9 diferentes por paciente en cada CRG	2,70	4,84	4,31	6,01	9,52
Mujeres	2,90	5,57	4,97	6,98	9,93
Hombres	3,02	4,03	3,95	5,22	8,98
Valor medio de códigos CIE-9 por paciente en cada CRG	4,27	12,96	13,28	18,74	32,15
Mujeres	4,70	14,96	14,34	21,78	33,54
Hombres	4,79	10,79	12,28	16,27	30,39
Código ATC Totales	746				
Códigos ATC diferentes por paciente en cada CRG	2,12	5,71	5,30	9,11	14,94
Mujeres	2,16	6,11	5,65	10,42	15,65
Hombres	1,44	7,34	5,11	8,01	14,04
Valor medio de códigos ATC por paciente en cada CRG	2,95	22,01	21,81	50,07	86,44
Mujeres	3,01	23,59	23,25	57,29	90,57
Hombres	2,01	28,28	21,04	44,02	81,22

Métodos de clasificación multi-clase y multi-etiqueta

1. Introducción
 2. Tipos de clasificadores utilizados
 3. Evaluación de prestaciones
-

Introducción



Introducción

- Se desea que modelo predictivo construido tenga capacidad de clasificar adecuadamente casos no utilizados durante el aprendizaje.
- Se propone abordar el diseño del modelo bajo dos enfoques: multi-clase (MC) y multi-etiqueta (ME).
- Retos:
 - Desbalanceo en las categorías.
 - Gran número de características con respecto al conjunto de casos utilizados en el diseño.

Tipos de clasificadores utilizados

Lineales

- Regresión Logística Multinomial (MC)
- Máquinas de Vectores Soporte Lineal (MC y ME)

No lineales

- Máquinas de Vectores Soporte No Lineal (MC y ME)
- *K Nearest Neighbour* (MC y ME)
- Árboles de decisión (MC y ME)
- *Random Forest* (MC y ME)
- Perceptrón Multicapa (MC y ME)

Evaluación de prestaciones

- Tasa de acierto multi-clase.

$$accuracy = \frac{tp + tn}{tp + fp + tn + fn}$$

		Clase estimada		
		C1	C2	C3
Clase test	C1	<i>tp</i>	<i>fn</i>	<i>fn</i>
	C2	<i>fp</i>	<i>tn</i>	<i>fn</i>
	C3	<i>fp</i>	<i>fn</i>	<i>tn</i>

Considerando C1 en *test*.

- Tasa de acierto multi-etiqueta.

$$accuracy = \frac{1}{n} \sum_{i=1}^n \frac{Y_i \cap P_i}{Y_i \cup P_i}$$

n: número máximo de etiquetas un caso.

Clases de test – $Y = \{Y_1, Y_2, \dots, Y_i\} \ i = 1 \dots n$

Clases estimadas – $P = \{P_1, P_2, \dots, P_i\} \ i = 1 \dots n$

Experimentos y resultados

1. Diseño de los experimentos
 2. Procedimiento
 3. Resultados
-

Diseño de los experimentos

- Clasificación multi-etiqueta y multi-clase.

CRG	Multi-clase	Multi-etiqueta		
	Clase	Hipertensión	Diabetes	Comorbilidades
CRG1000	0	0	0	0
CRG5192	1	1	0	0
CRG5424	2	0	1	0
CRG6144	3	1	1	0
CRG7071	4	1	1	1

- Dos escenarios:
 - Escenario 1 - Pacientes sanos y crónicos
 - Escenario 2 - Pacientes crónicos.
- Balanceo de observaciones: submuestreo sobre las clases mayoritarias con los pacientes de a clase minoritaria (CRG7071 - 547) creando 50 subconjuntos.
 - Escenario 1 - 2735 pacientes
 - Escenario 2 - 2188 pacientes
- Conjunto de *train* (80%) y *test* (20%).
- Características clínicas binarias o numéricas.
- Selección de las 100 características clínicas.
 - FS1 - Frecuencia
 - FS2 - Prueba F-Fisher
 - FS3 - *Random Forest*

Procedimiento

Este procedimiento se lleva a cabo para cada enfoque: multi-clase y multi-etiqueta y considerando las características clínicas de forma binaria o numérica.

1. Carga 50 subconjuntos de *train* y *test*, considerando de manera independiente las características seleccionadas con FS1, FS2 y FS3.

Para cada subconjunto de *train* – FS y algoritmo de clasificación:

2. Filtrado de los casos duplicados y normalización de cada característica del conjunto de *train*.

3. Búsqueda de los mejores parámetros libres con validación cruzada.

4. Ajuste del modelo con los mejores parámetros.

5. Normalización de los 50 subconjuntos de *test*.

6. Evaluación de los 50 subconjuntos de *test* con el diseño obtenido en el paso **4**.

Resultados

- Mejores resultados en tasa de acierto (media y desviación típica).

Clasificación	Forma de las características	Selección de características	Pacientes sanos y crónicos		Pacientes crónicos	
			Algoritmo	Tasa de acierto	Algoritmo	Tasa de acierto
Multi-clase	Numéricas	Frecuencia	Random Forest	85,3(1,4)%	Random Forest	83,2(1,6)%
		F-Fisher	Regresión Logística Multinomial	88,8(1,1)%	Linear SVM	88,6(1,4)%
		Random Forest	Random Forest	86,0(1,4)%	Regresión Logística Multinomial	84,9(1,7)%
	Binarias	Frecuencia	No Linear SVM	85,8(1,4)%	No Linear SVM	83,7(1,6)%
		F-Fisher	No Linear SVM	91,6(1,2)%	No Linear SVM	90,9(1,2)%
		Random Forest	No Linear SVM	88,9(1,3)%	No Linear SVM	87,7(1,4)%
Multi-etiqueta	Numéricas	Frecuencia	Linear SVM OneVsRest	85,2(1,52)%	Linear SVM Label Powerset	82,7(1,9)%
			Random Forest	84,7(1,4)%	Random Forest	82,5(1,8)%
		F-Fisher	Linear SVM OneVsRest	89,5(1,3)%	Linear SVM Label Powerset	88,6(1,5)%
			Random Forest	85,7(1,2)%	Random Forest	84,2(1,7)%
		Random Forest	Linear SVM OneVsRest	87,7(1,4)%	Linear SVM Label Powerset	86,4(1,9)%
			Random Forest	85,5(1,3)%	Random Forest	82,7(1,7)%
	Binarias	Frecuencia	No Linear SVM Label Powerset	85,6(1,3)%	No Linear SVM Label Powerset	84,0(1,4)%
			MLP	84,4(1,4)%	MLP	81,7(1,7)%
		F-Fisher	Linear SVM OneVsRest	91,2(1,0)%	No Linear SVM Label Powerset	90,3(1,2)%
			MLP	90,3(1,1)%	MLP	88,3(1,5)%
		Random Forest	Linear SVM OneVsRest	89,1(1,1)%	No Linear SVM Label Powerset	88,6(1,0)%
			MLP	87,7(1,5)%	MLP	86,8(1,2)%

Resultados

- Resultados (media y desviación típica) considerando distintos tipos de características (individuales y agregado).
- Mejor configuración: características binarias y selección de características basadas en la prueba F-Fisher (FS2).

			Pacientes sanos y crónicos		Pacientes crónicos	
			Número características	Tasa de acierto	Número características	Tasa de acierto
Perceptrón Multicapa	Multi-clase	Género + Edad + CIE-9 + ATC	102	90,3(1,0)%	102	88,0(1,5)%
		CIE-9 + ATC	100	72,8(7,5)%	100	98,0(1,4)%
		CIE-9	31	54,0(9,4)%	30	32,8(11,2)%
		ATC	69	74,2(8,8)%	70	72,8(7,5)%
		Género + Edad	2	31,8(2,0)%	2	22,1(2,0)%
	Multi-etiqueta	Género + Edad + CIE-9 + ATC	102	90,3(1,1)%	102	88,3(1,5)%
		CIE-9 + ATC	100	89,5(1,2)%	100	88,1(1,5)%
		CIE-9	33	45,5(19,6)%	34	55,1(6,3)%
		ATC	67	82,0(3,5)%	66	82,3(1,5)%
		Género + Edad	2	25,1(5,3)%	2	30,4(2,5)%

Conclusiones y líneas futuras

1. Conclusiones

2. Líneas futuras

Conclusiones

- El paciente tiene asignados más códigos CIE-9 y ATC para los CRG con más de una cronicidad.
- Mejor método de selección de características: prueba F-Fisher.
- Mejores tasas de acierto cuando consideramos la base de datos de pacientes sanos y crónicos.
- Para características binarias, mejores prestaciones en modelos no lineales; para características basadas en ocurrencia, mejores prestaciones en modelos lineales.
- Para métodos de transformación de problema, resultados en multi-clase y multi-etiqueta similares.
- Para los modelos adaptados, resultados ligeramente mejores (0,3%) para multi-etiqueta. El número de pacientes considerados nos limita el aprendizaje de los modelos propuestos.

Líneas futuras

- Aumentar la población de estudio para que aumente la clase minoritaria y así poder tener subconjuntos de *train* con más casos para entrenar los algoritmos.
- Estudiar algoritmos que no beneficien a la clase mayoritaria para no realizar balanceo de observaciones.
- Trabajar la visualización para que el personal clínico no vea sólo una ‘caja negra’.
- Entrenar los algoritmos en algún servicio de *cluster* de servidores como *Amazon Web Services*.

**Muchas gracias
por su atención**



Universidad
Rey Juan Carlos



ESTRATEGIA