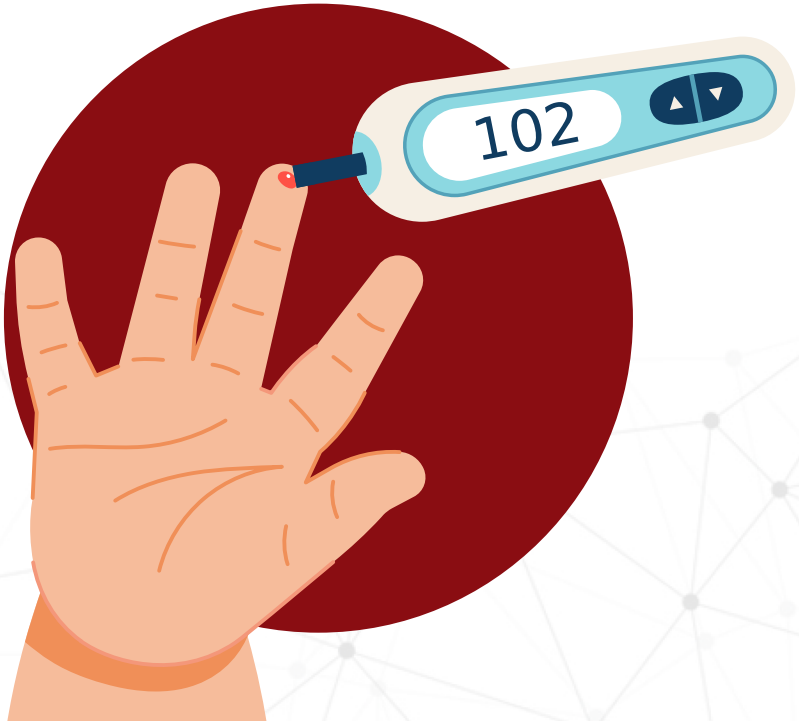
Primeras Experiencias en la Identificación de Personas con Riesgo de Diabetes en la Población Argentina utilizando Técnicas de Aprendizaje Automático



Gonzalo Tittarelli
Directores: Enzo Rucci, Franco Ronchetti

Octubre 2023

CENEXA
CONICET
U N L P
CEA: CICPBA

III-LIDI
Instituto de Investigación en Informática - LIDI

Facultad de INFORMÁTICA
UNIVERSIDAD NACIONAL DE LA PLATA

AGENDA

01
Introducción

02
conjunto de datos

05
cONCLUSIONES y
TRABAJOS FUTUROS

03
modelos y
experimentos

04
Resultados
obtenidos

0 1

Introducción

El problema, la motivación y el objetivo

diabetes

La enfermedad

La diabetes es una **enfermedad crónica** no transmisible caracterizada por **niveles elevados de glucemia**

Diabetes tipo 1

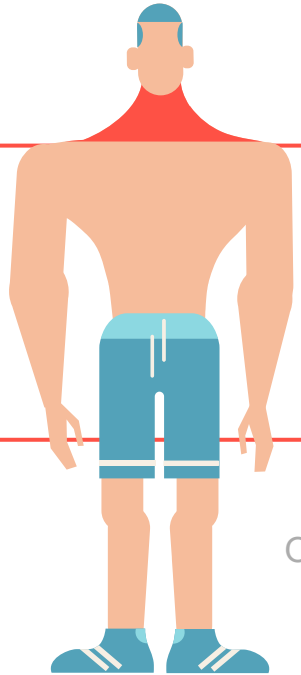
No prevenible - insulinodependientes

CASOS

5-10%

fACTORES DE RIESGO

Desconocidos con exactitud



hiperglucemia

Diabetes tipo 2

Prevenible - insulinoresistentes

CASOS

90 -95%

FACTORES DE RIESGO

Obesidad, sexo, sobrepeso, falta de actividad física, antecedentes familiares, tabaquismo

En la prediabetes (PDM) existe una elevación de glucemia pero no alcanza para diagnosticar diabetes.

complicaciones

retinopatía

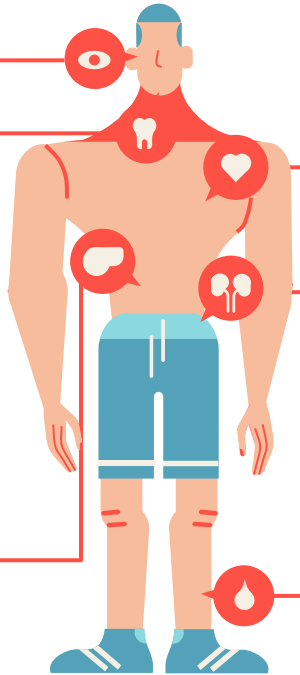
Pudiendo causar pérdida de visión o ceguera

problemas en salud bucal

Periodontitis, infecciones hasta pérdida de piezas

hígado

Acumulación de grasa e inflamación o insuficiencia



enfermedades cardiovasculares

Infartos, ACV, insuficiencia cardíaca

nefropatía

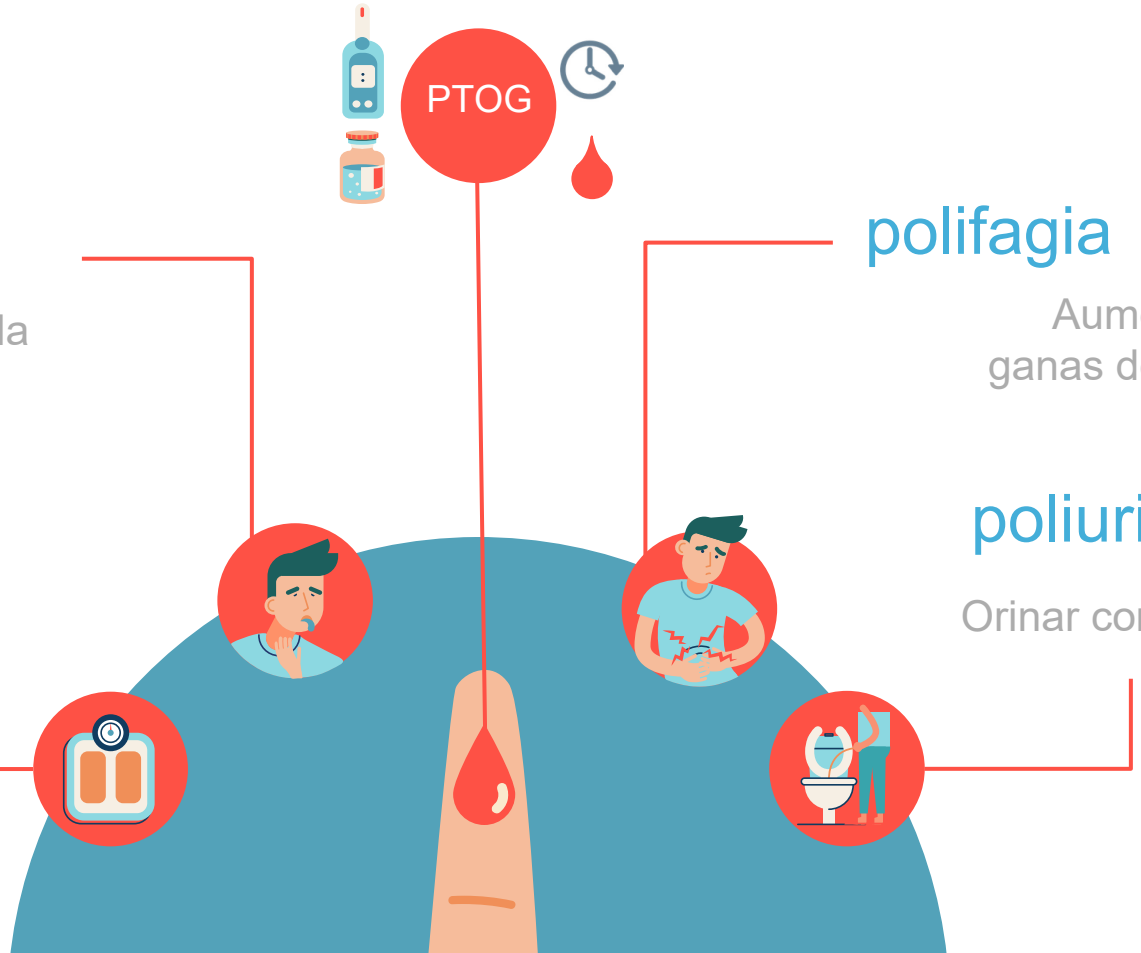
Riñones menos eficientes o fallen por completo

pie diabético

Riesgo de ulceración, infección y amputación

Costos económicos y sociales, su elevada tasa de mortalidad así como el impacto sobre el entorno social, familiar y laboral de los pacientes.

síntomas y diagnóstico



polidipsia

Necesidad exagerada de beber agua

polifagia

Aumento de ganas de comer

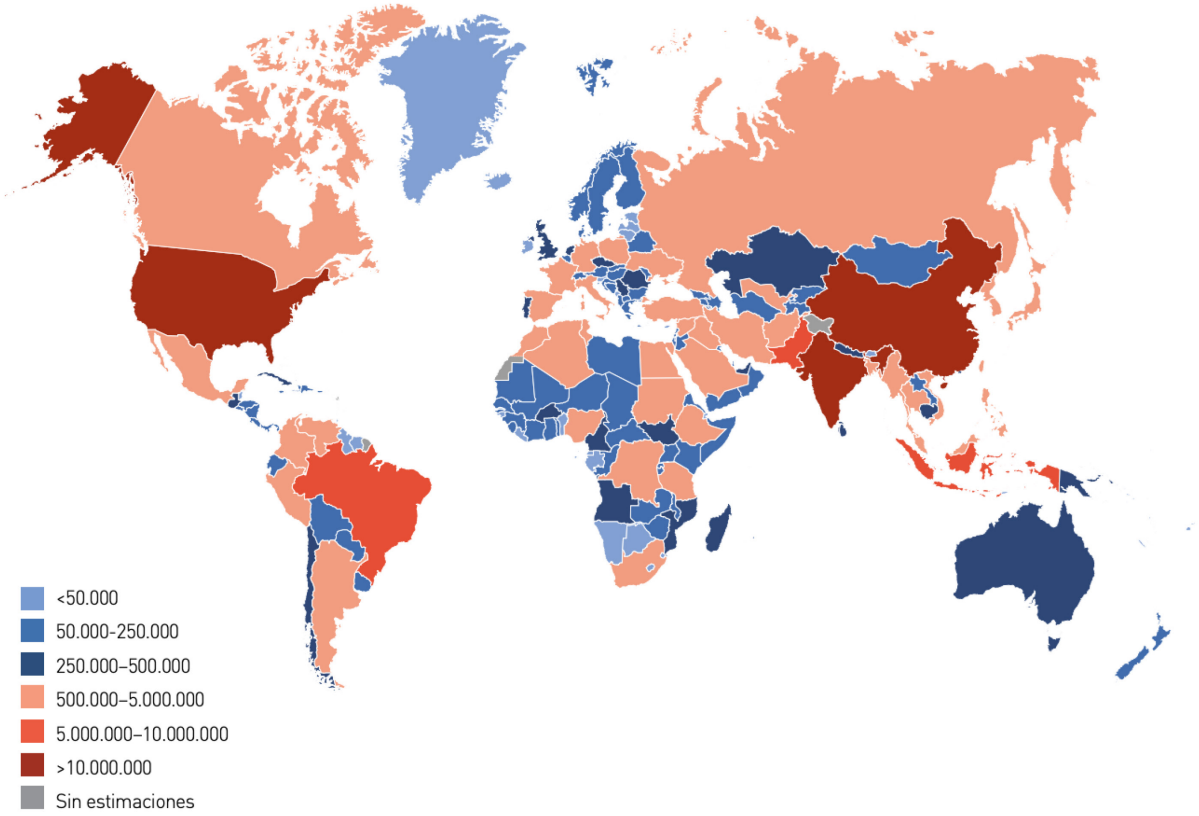
pérdida de peso

poliuria

Orinar con frecuencia

prevalencia de DM

- En los últimos años, ha habido un notorio aumento, tanto que se le ha denominado la “epidemia del siglo XXI”.
- En 2019, **500 millones** de personas padecían esta enfermedad
- Para 2045 se estimó un crecimiento a **700 millones**.
- La creciente prevalencia es un **serio problema de salud mundial**.

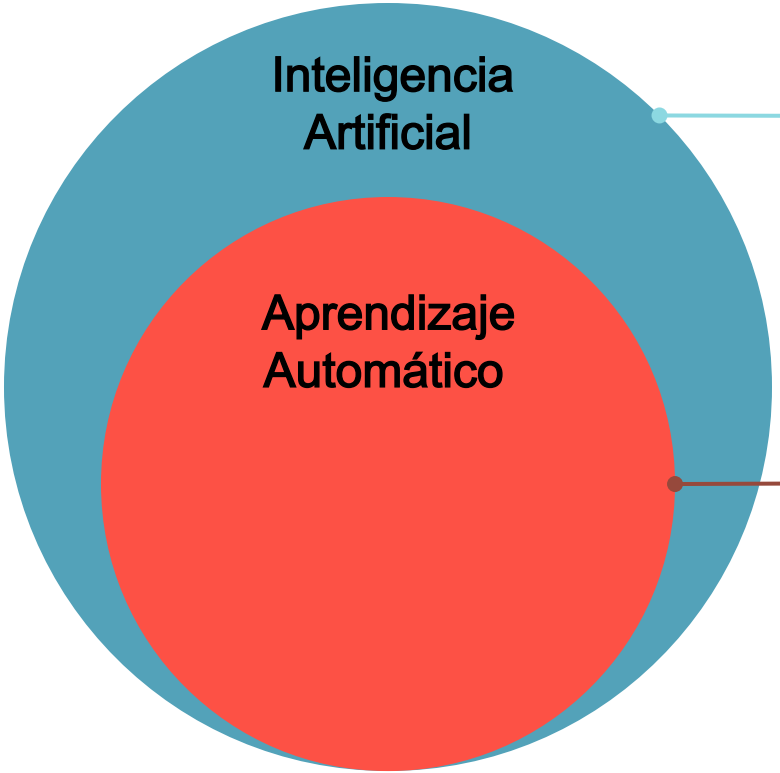


Cantidad de adultos (20–79 años) con diabetes sin diagnosticar en 2019. Extraído de *Atlas de la Diabetes FID*.

aprendizaje automático

El cerebro

Aprendizaje automático (ml)



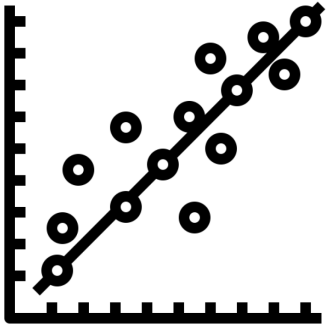
Disciplina dedicada al estudio y diseño de sistemas que puedan imitar comportamientos inteligentes.

Campo de estudio que brinda a las computadoras la capacidadde aprendera clasificar o predecir en baseа experiencias

Campo de estudio que brinda a las computadoras la capacidad de aprender sin ser explícitamente programadas
Arthur Samuel 1959

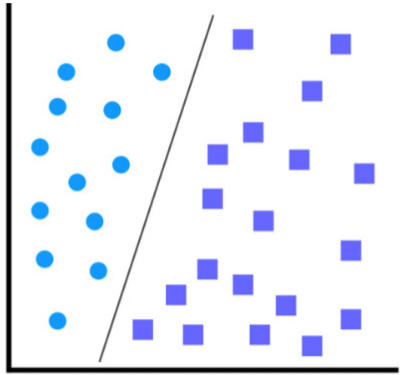
Tipo de problema

regresión

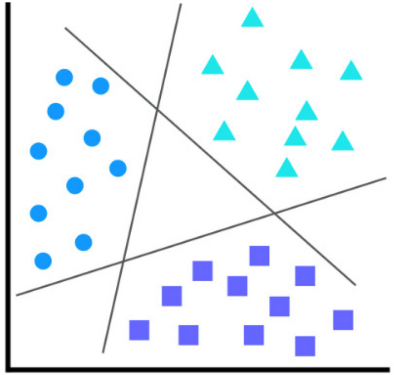


clasificación

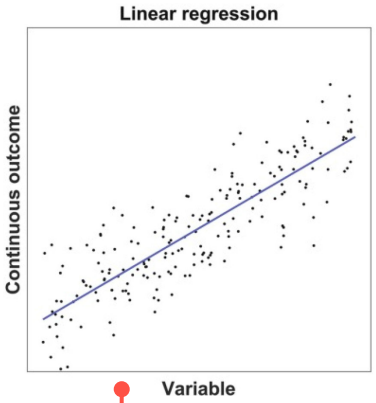
binaria



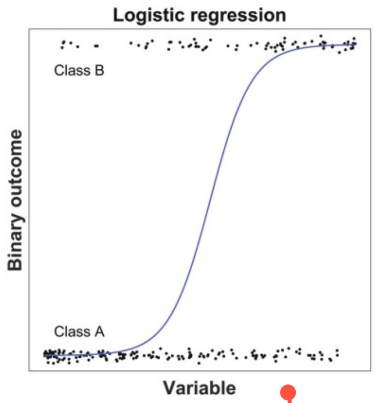
multiclase



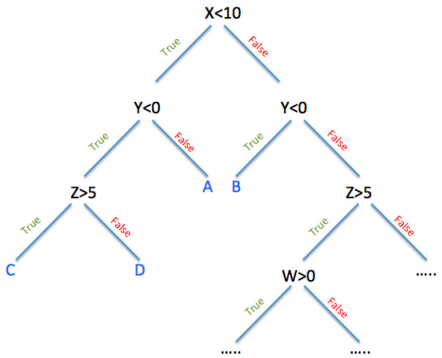
modelos



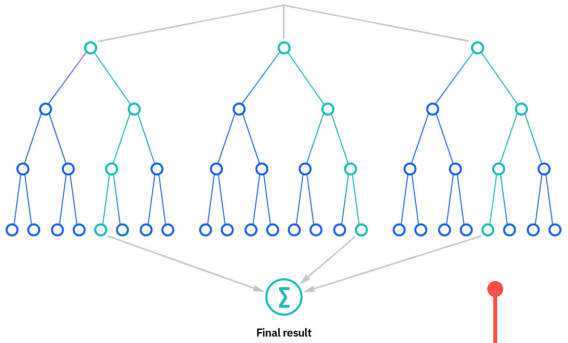
Regresión Lineal



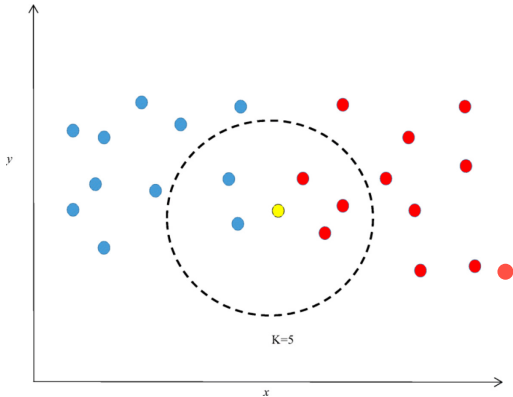
Regresión Logística



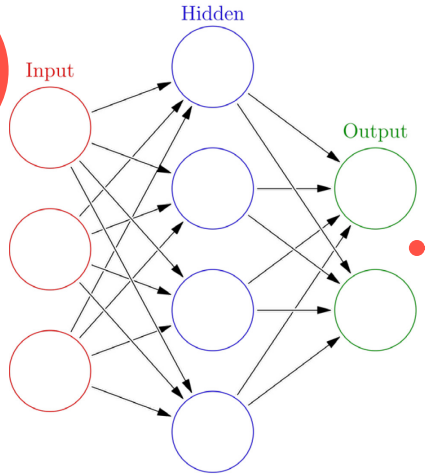
Árbol de Decisión



Random Forest



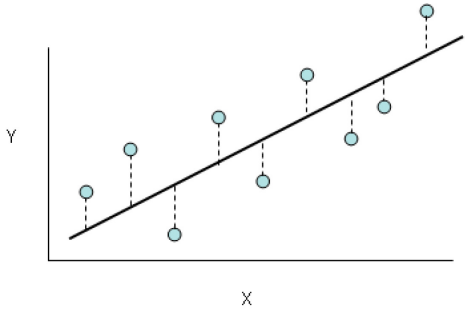
k vecinos más cercanos



Redes Neuronales

métricas de evaluación

regresión



$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

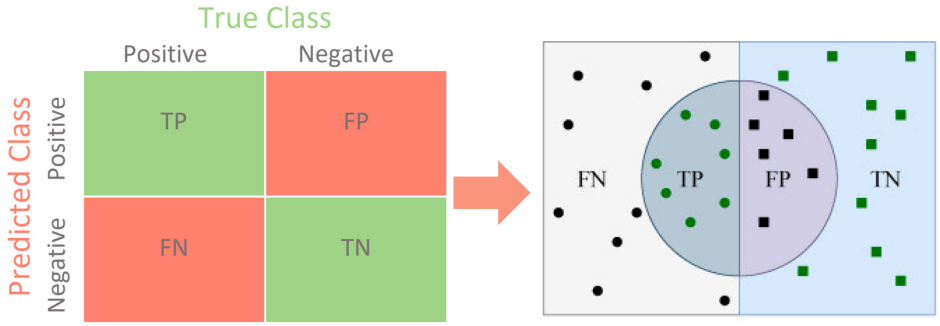
$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}}$$

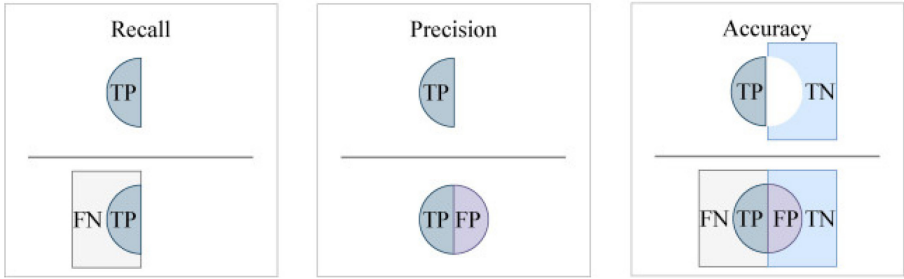
$$R^2 = 1 - \frac{\text{MSE(model)}}{\text{MSE(baseline)}}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

$$\text{MSE(baseline)} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2$$

clasificación binaria





$$F_1 score = 2 \cdot \frac{(precision \cdot recall)}{(precision + recall)}$$

motivación

- El amplio uso de ML en salud y medicina para el **diagnóstico** y **predicción de riesgos** de diversas enfermedades resulta ser prometedor.
- La detección de DT2 y PDM representa un desafío debido a la **ausencia de síntomas** y a la **falta de conocimiento de los factores de riesgo asociados**
- Las características de esta enfermedad hacen que sea **particular por población**
- Aunque existan modelos de ML orientados a su detección, esta particularidad los exime poblacionalmente

PPDBA es un programa desarrollado por el CENEXA (UNLP-CONICET) el cual nos disponibiliza su conjunto de datos

+

No existen modelos de ML para identificar personas con riesgo de DM y PDM en Argentina

=

objetivo

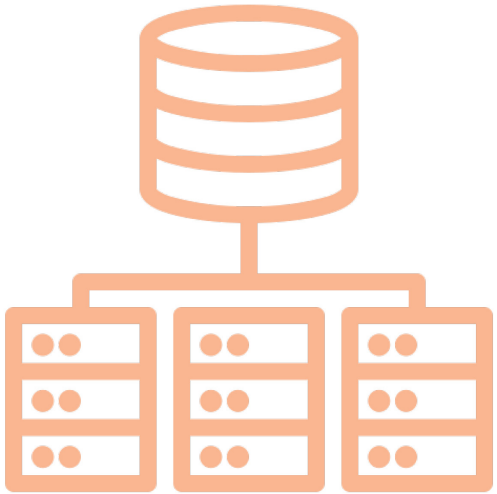
Desarrollar y evaluar modelos basados en ML que permitan identificar personas con riesgo de diabetes y PDM en la población Argentina utilizando el conjunto de datos correspondiente al programa PPDBA

Los modelos no reemplazan el diagnostico, más bien identifican personas con alta probabilidad de padecer la enfermedad

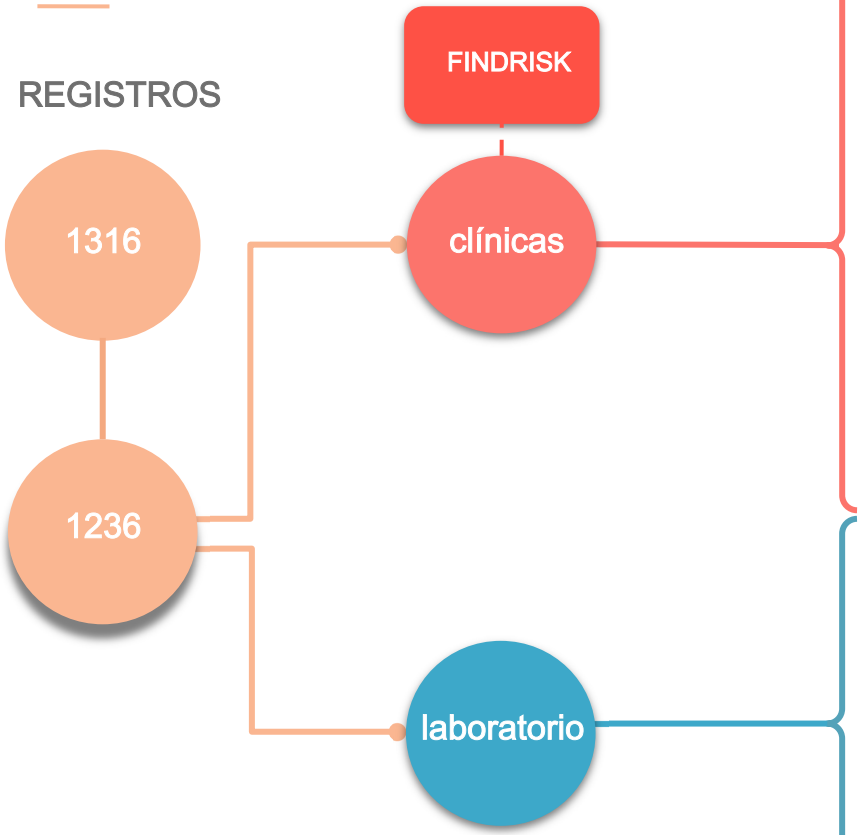
02

Conjunto de datos

El corazón



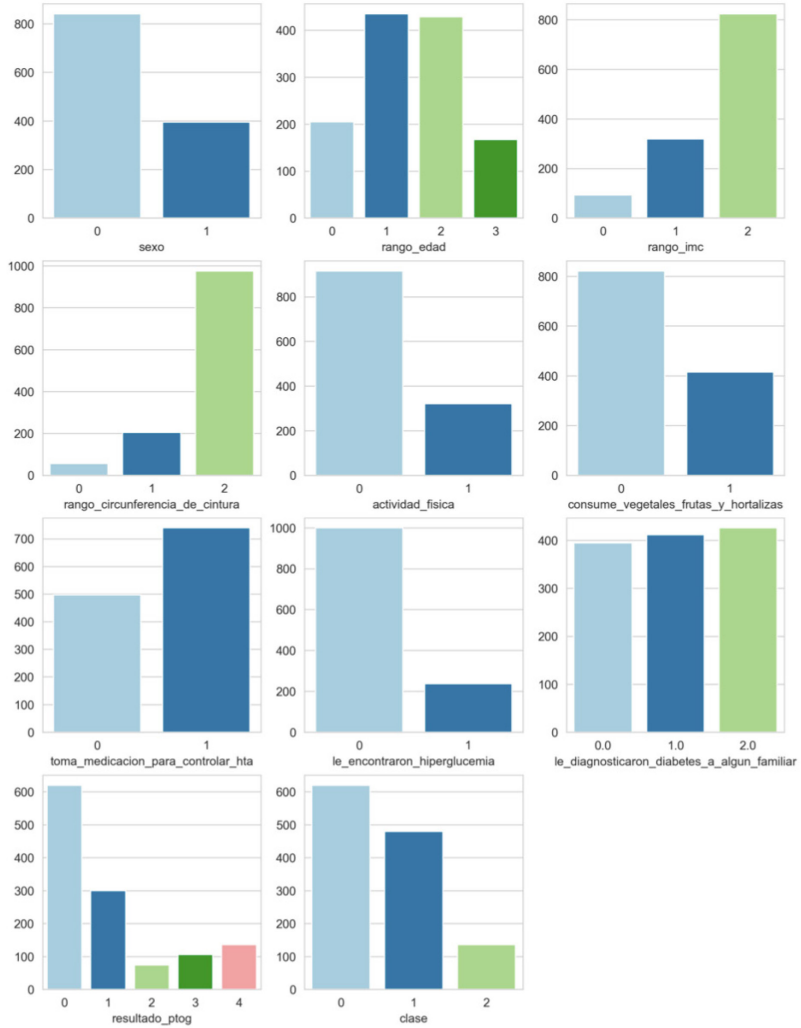
CONJUNTO DE DATOS



Feature	Significado	Categoría médica
id	Identificador del registro	-
sexo	Sexo de la persona	Clínica
edad	Edad de la persona	Clínica
rango_edad	Agrupación de la edad en rangos	Clínica
imc	Índice de Masa Corporal calculado como $\frac{peso\_en\_kg}{altura\_en\_mts^2}$	Clínica
rango_imc	Agrupación del IMC en rangos	Clínica
circunferencia _de_cintura	Circunferencia de cintura expresada en centímetros	Clínica
rango_circunferencia _de_cintura	Agrupación de la circunferencia de cintura en rangos	Clínica
actividad_fisica	Práctica regular de actividad física	Clínica
consume_vegetales _frutas_y_hortalizas	Incluye frutas y verduras en su alimentación diaria	Clínica
toma_medicacion _para_controlar_hta	Toma medicación para controlar hipertensión	Clínica
le_encontraron _hiperglucemia	Detección de hiperglucemia en un examen médico, durante embarazo o estudio	Clínica
le_diagnosticaron_diabetes _a_algun_familiar	Antecedentes familiares de DM	Clínica
glucemia_basal	Valor de glucemia en ayunas expresado en mg/dL	Laboratorio
glucemia_postprandial	Valor de glucemia, expresado en mg/dL, 2 horas después de haber ingerido una solución azucarada (75g de glucosa disuelta en 375ml de agua)	Laboratorio
colesterol_ldl	Colesterol LDL expresado en mg/dL	Laboratorio
colesterol_total	Colesterol total expresado en mg/dL	Laboratorio
colesterol_hdl	Colesterol HDL expresado en mg/dL	Laboratorio
trigliceridos	Triglicéridos expresados en mg/dL	Laboratorio
creatinina_basal	Creatinina basal expresada en mg/dL	Laboratorio
hemoglobina_glucosilada	Hemoglobina glucosilada expresada en porcentaje	Laboratorio
resultado_ptog	Resultado de la PTOG	Laboratorio
clase	Resultado agrupado de la PTOG	-

caracterización

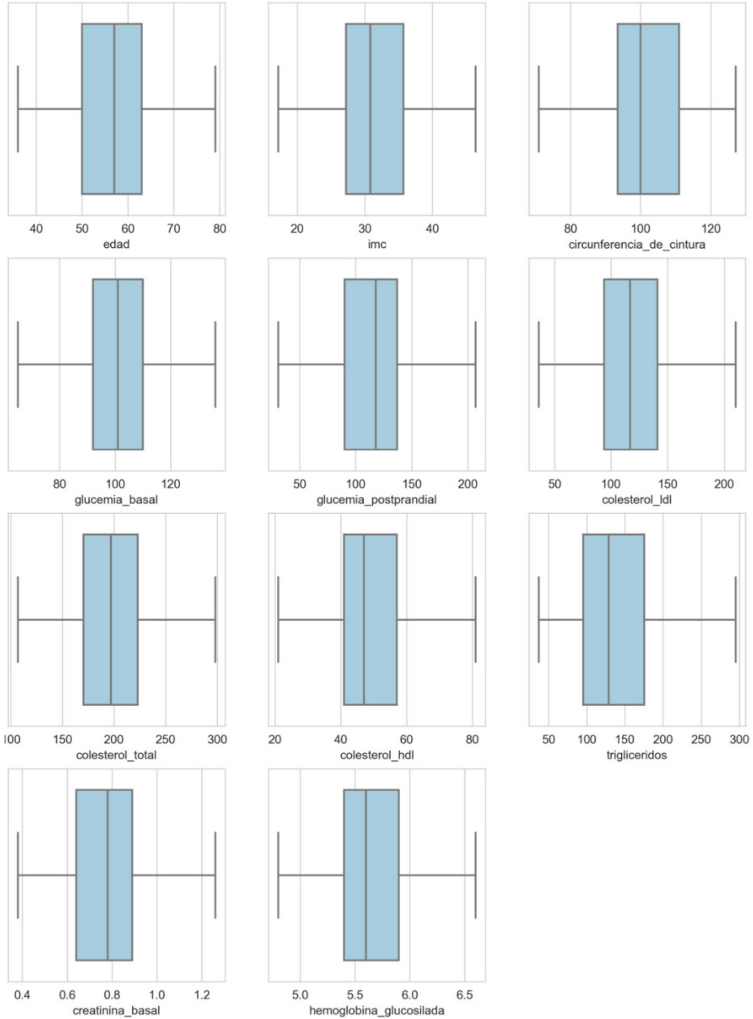
- Más personas de sexo femenino que masculino.
- No realizan actividad física regular.
- No consumen vegetales, frutas y hortalizas.
- Toman medicación para controlar hipertensión.
- No han tenido hiperglucemia previamente.
- Tienen antecedentes familiares que padecen DM.
- La mitad de las personas no están en riesgo de padecer PDM o DM.
- El porcentaje de personas con PDM es mayor a aquellas con DM.



Histogramas de los features categóricos

caracterización

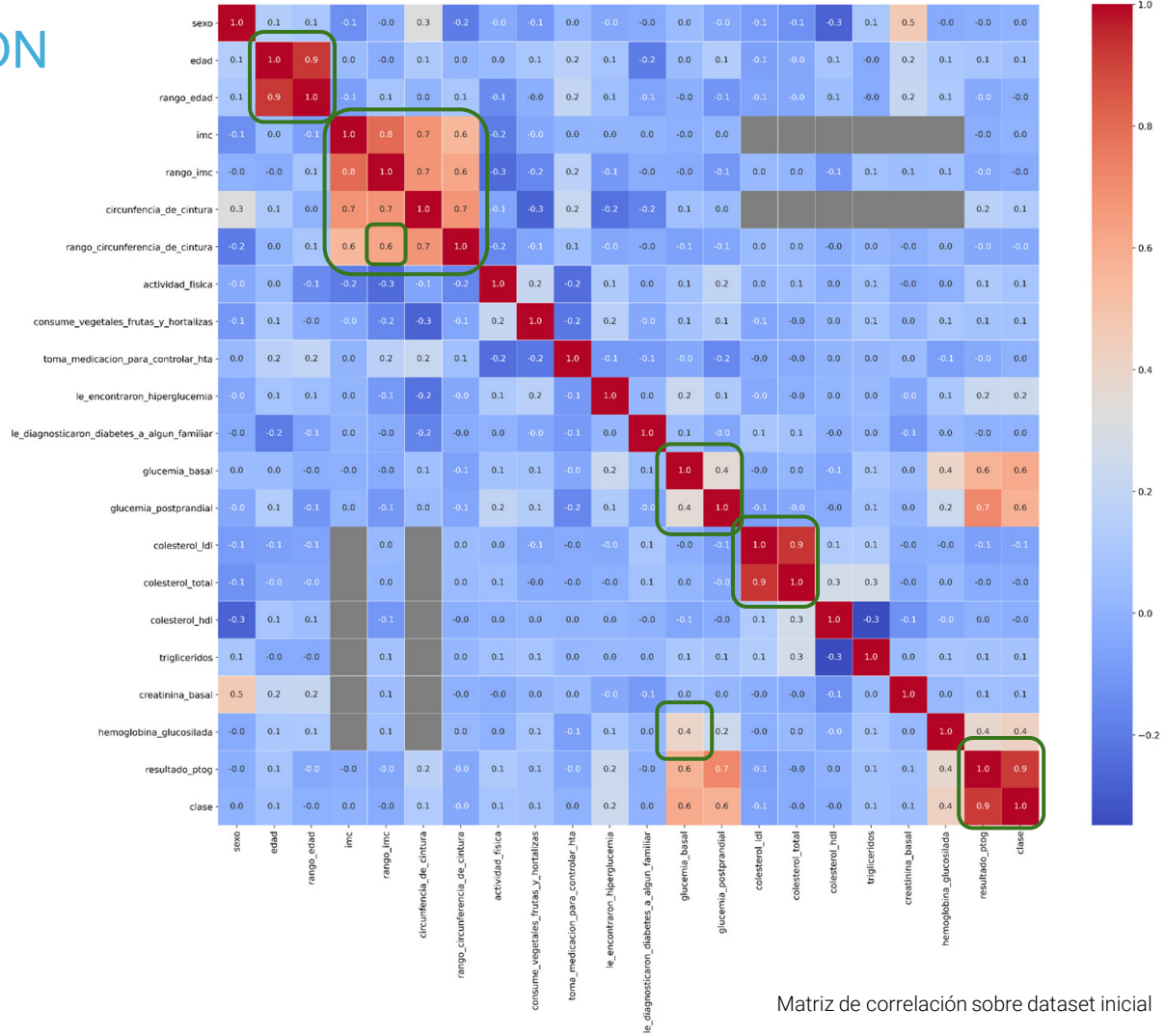
- Grupoetario de 45-64 años.
- IMCmayor a 30 kg/m2.
- Circunferenciade cintura mayor a 102cm para el sexomasculinoy mayor a 88cm para el sexofemenino.
- La glucemia basal, colesterol HDL, triglicéridos y creatinina basal parecerían tener una amplia dispersión.



Diagramas de cajas de los features continuos

matriz de CORRELACIÓN

- Correlaciónlineal débil entre rango de circunferencia de cintura y rango de IMC.
- Correlaciones débiles entre los rangos de edad, IMC y circunferencia de cintura y las variables `edad`, `imc` y `circ___de_cintura`.
- Correlación fuerte entre `colesterol_total` y `colesterol_ldl`.
- La relación entre `glucemia_basal` y `glucemia_pprandial` tiene sentido.
- La `glucemia_basal` y `hemoglobina_glucosilada` tienen una correlación débil.
- La `clase` es una agrupación de `resultado_ptog`.



Matriz de correlación sobre dataset inicial

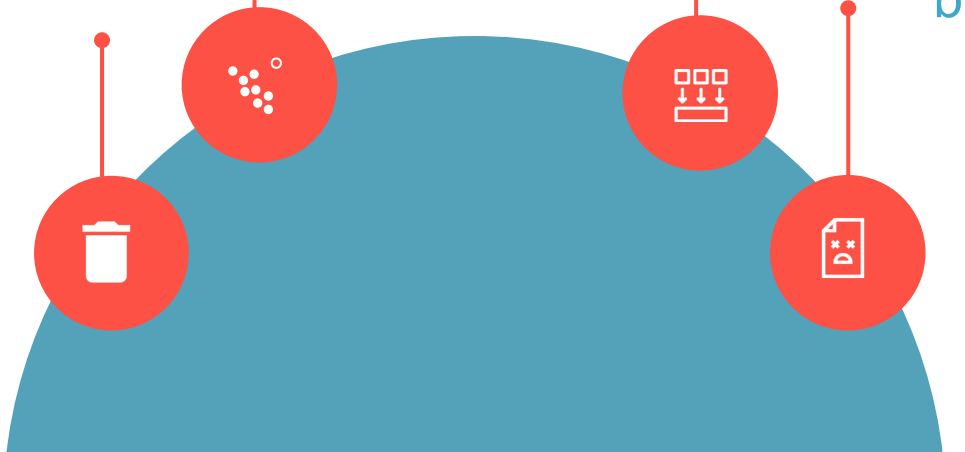
preprocesamiento

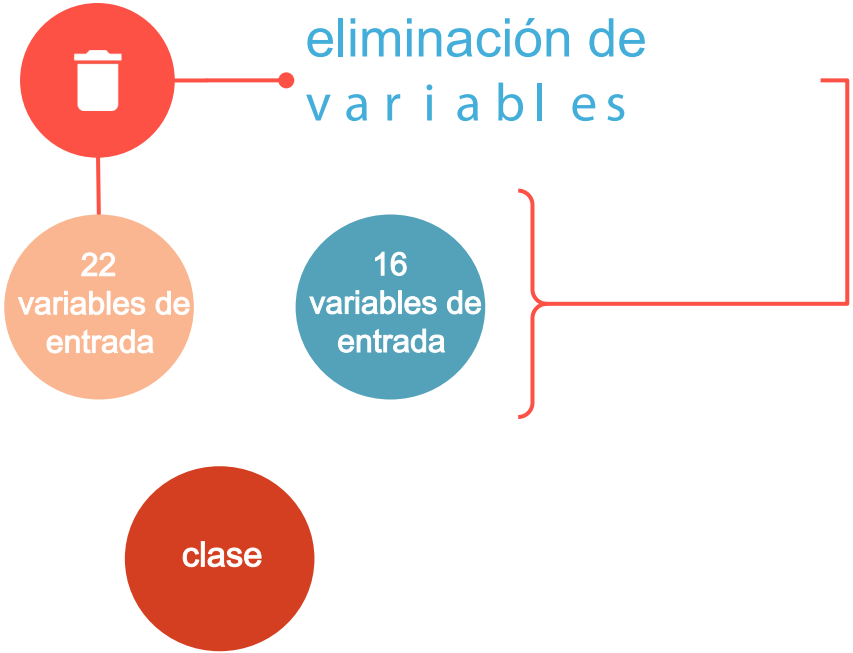
VALORES ATÍPICOS

valores nulos

Eliminación de variables

binarización





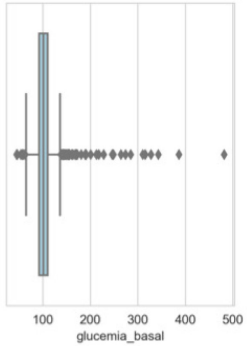
Feature	Significado	Categoría médica
sexo	Sexo de la persona	Clínica
rango_edad	Agrupación de la edad en rangos	Clínica
rango_imc	Agrupación del IMC en rangos	Clínica
rango_circunferencia_de_cintura	Agrupación de la circunferencia de cintura en rangos	Clínica
actividad_fisica	Práctica regular de actividad física	Clínica
consume_vegetales_frutas_y_hortalizas	Incluye frutas y verduras en su alimentación diaria	Clínica
toma_medicacion_para_controlar_hta	Toma medicación para controlar hipertensión	Clínica
le_encontraron_hiperglucemia	Detección de hiperglucemia en un examen médico, durante embarazo o estudio	Clínica
le_diagnosticaron_diabetes_a_algun_familiar	Antecedentes familiares de DM	Clínica
glucemia_basal	Valor de glucemia en ayunas expresado en mg/dL	Laboratorio
colesterol_ldl	Colesterol LDL expresado en mg/dL	Laboratorio
colesterol_total	Colesterol total expresado en mg/dL	Laboratorio
colesterol_hdl	Colesterol HDL expresado en mg/dL	Laboratorio
trigliceridos	Triglicéridos expresados en mg/dL	Laboratorio
creatinina_basal	Creatinina basal expresada en mg/dL	Laboratorio
hemoglobina_glucosilada	Hemoglobina glucosilada expresada en porcentaje	Laboratorio
clase	Resultado agrupado de la PTOG	-

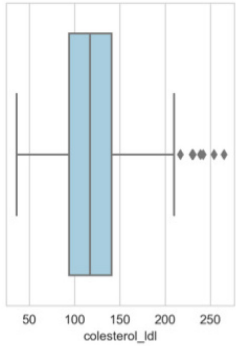
valores atípicos

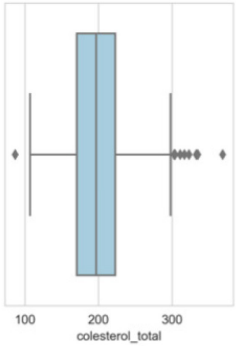
- Se utilizó el método de **análisis de rango intercuartil** para identificar los intervalos de **valores atípicos leves y extremos**.
- Existen **4 variables** con valores atípicos extremos.
- Únicamente se eliminaron dos valores de la `creatinina_basal`.

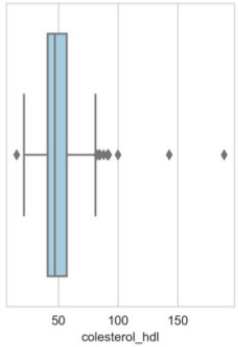
Feature	Intervalos de valores típicos leves	Atípicos leves	Atípicos extremos
glucemia_basal	[38; 65; 137; 164]	36 registros	24 registros
colesterol_ldl	[-47; 23,5; 211,5; 282]	[239; 230; 242; 265; 217; 254; 231]	-
colesterol_total	[13; 91,75; 301,75; 380,5]	[323; 317; 304; 369; 335; 303; 333; 87; 311]	-
colesterol_hdl	[-7; 17; 81; 105]	[100; 91; 92; 82; 88; 85; 84; 15; 92]	[143; 189]
trigliceridos	[-146,5; -25,75; 296,25; 417]	21 registros	12 registros
creatinina_basal	[-0,11; 0,26; 1,27; 1,64]	[1,56; 1,37; 0,21; 1,29; 1,4; 1,32; 1,49; 1,3]	[2,57; 1,77; **118**; 1,73; **85**; 2,17]
hem___glucosilada	[3,9; 4,65; 6,65; 7,4]	23 registros	-

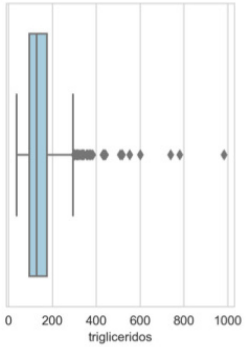
Valores atípicos leves y extremos para los features continuos

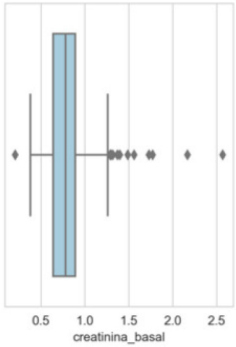


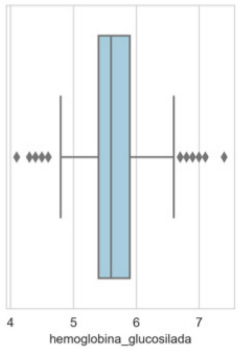












valores nulos variables categóricas

Feature	Tipo	N (no nulo)	Porcentaje de nulos (N)	Valor	N	Porcentaje
sexo	Cualitativo (Nominal)	1236	0%	0 ⇒ Femenino (f)	841	68%
				1 ⇒ Masculino (m)	395	32%
rango_edad	Cualitativo (Ordinal)	1236	0%	0 ⇒ menor de 45 años	205	17%
				1 ⇒ entre 45 y 54 años	435	35%
				2 ⇒ entre 55 y 64 años	429	35%
				3 ⇒ mayor de 64 años	167	14%
rango_imc	Cualitativo (Ordinal)	1236	0%	0 ⇒ menor de 25 IMC	93	8%
				1 ⇒ entre 25 y 30 IMC	319	26%
				2 ⇒ mayor de 30 IMC	824	67%
rango___cintura	Cualitativo (Ordinal)	1236	0%	0 ⇒ m: menos de 94cm	56	5%
				0 ⇒ f: menos de 80cm		
				1 ⇒ m: entre 94cm y 102cm	205	17%
				1 ⇒ f: entre 80cm y 88 cm		
				2 ⇒ m: mayor de 102 cm	975	79%
				2 ⇒ f: mayor de 88cm		
actividad_fisica	Cualitativo (Nominal)	1236	0%	0 ⇒ No	915	74%
				1 ⇒ Sí	321	26%
cons___hortalizas	Cualitativo (Nominal)	1236	0%	0 ⇒ No todos los días	821	66%
				1 ⇒ Todos los días	415	34%
toma___hta	Cualitativo (Nominal)	1236	0%	0 ⇒ No	497	40%
				1 ⇒ Sí	739	60%
le___hiperglucemia	Cualitativo (Nominal)	1236	0%	0 ⇒ No	999	81%
				1 ⇒ Sí	237	19%
le_diag___familiar	Cualitativo (Nominal)	1233	0,2% (3)	0 ⇒ No	395	32%
				1 ⇒ Sí: abuelo, tío, tía, o primo	412	33%
				2 ⇒ Sí: padre, hermano/a e hijo/a	426	35%
resultado_ptog	Cualitativo (Nominal)	1236	0%	0 ⇒ Normal	620	50%
				1 ⇒ GAA	300	24%
				2 ⇒ TGA	74	6%
				3 ⇒ GAA+TGA	106	9%
				4 ⇒ DM	136	11%
clase	-	1236	0%	0 ⇒ Normal	620	50%
				1 ⇒ PDM (GAA, TGA o GAA+TGA)	480	39%
				2 ⇒ DM	136	11%

valores nulos variables continuas

- El cuestionario FINDRISKrefiere a rangos de edad, IMC y circunferencia de cintura; algunas personas completaron con valores exactos en lugar de un rango.
- La mayoría de los datos de laboratorio presentan nulos.

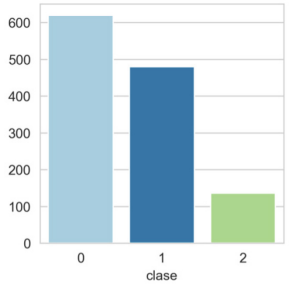
Feature	Tipo	N (no nulo)	Porcentaje de nulos (N)	Media	Desviación estándar	Mínimo	Máximo
edad	Cuantitativo (Discreto)	672	45,6% (564)	57,237	8,849	24	99
imc	Cuantitativo (Continuo)	617	50,1% (619)	31,651	6,337	17,2	54,9
cir___cintura	Cuantitativo (Discreto)	55	95,6% (1181)	101,309	13,684	62	127
glucemia_basal	Cuantitativo (Continuo)	1236	0%	104,362	27,298	45	482
glucemia_postprandial	Cuantitativo (Continuo)	1181	4,4% (55)	119,594	42,532	15	343
colesterol_ldl	Cuantitativo (Continuo)	521	57,8% (715)	119,798	36,851	0	265
colesterol_total	Cuantitativo (Continuo)	531	57% (705)	198,271	41,148	87	369
colesterol_hdl	Cuantitativo (Continuo)	530	57,1% (706)	49,826	14,419	15	189
trigliceridos	Cuantitativo (Continuo)	531	57% (705)	151,409	95,687	37	983
creatinina_basal	Cuantitativo (Continuo)	617	50,1% (619)	1,117	5,809	0,21	118
hem___glucosilada	Cuantitativo (Continuo)	601	51,4% (635)	5,614	0,440	4,1	7,4

Eliminar registros completos

Eliminación por columna

Reemplazar con algún valor especial

Utilizar técnicas de ML su completitud



resultado_ptog	Cualitativo (Nominal)	1236	0%	0 ⇒ Normal	620	50%
				1 ⇒ GAA	300	24%
				2 ⇒ TGA	74	6%
				3 ⇒ GAA+TGA	106	9%
				4 ⇒ DM	136	11%
clase	-	1236	0%	0 ⇒ Normal	620	50%
				1 ⇒ PDM (GAA, TGA o GAA+TGA)	480	39%
				2 ⇒ DM	136	11%

Binarización

Feature	Tipo	N (no nulo)	Porcentaje de nulos (N)	Valor	N	Porcentaje
clase	-	1236	0%	0 ⇒ Sin riesgo	620	50%
				1 ⇒ Con riesgo	616	50%

Variable de clase creada transformando el problema en uno de clasificación binaria

SEGMENTACIONES

DCL-bin
DCL

CLÍNICA + LABORATORIO

- Eliminación de registros completos
- 16 variables de entrada
- 503 ejemplos
- DCL-bin: 229 sin riesgo / 274 con riesgo
- DCL: 229 normal / 251 PDM / 23 DM
- Más difícil de llevar a la práctica

DCG-bin
DCG

CLÍNICA + glucemia basal

- Eliminación de variables de laboratorio excepto la glucemia basal
- 10 variables de entrada
- 1233 ejemplos
- DCG-bin: 620 sin riesgo / 613 con riesgo
- DCG: 620 normal / 479 PDM / 134 DM
- Más fácil de llevar a la práctica

DC-bin
dc

CLÍNICA

- Eliminación de variables de laboratorio
- 9 variables de entrada
- 1233 ejemplos
- DC-bin: 620 sin riesgo / 613 con riesgo
- DC: 620 normal / 479 PDM / 134 DM
- Más sencillo, sin costo y factible de realizar en cualquier momento

03

MODELOS y experimentos

El planteo

ENTRENAMIENTO

validación cruzada

muestreo estratificado aleatorio

normalización

min-max
estándar para ANN

70 %

entrenamiento

30 %

evaluación

experimentaciones

modelos de clasificación

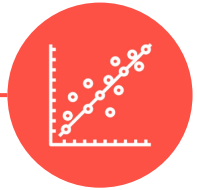
A fin de identificar el riesgo de diabetes, se desarrollaron y evaluaron **modelos de clasificación** utilizando las segmentaciones **DCL-bin**, **DCG-bin** y **DG-bin**

Selección de características

A modo exploratorio

Modelos de regresión

A fin de identificar el estadio particular de diabetes, se desarrollaron y evaluaron **modelos de regresión** utilizando la segmentación **DCL**



MODELOS de clasificación

Logistic Regression (LOR)

class_weight="balanced"

Random Forest (RF)

max_depth=2

Decision Tree (DT)

max_depth=5
class_weight="balanced"

Artificial Neural Networks (ANN)

1 capa oculta con 100 neuronas
Función de activación ReLU
Capa de salida con 2 neuronas (sigmoidea)
Regularización L2 de 0,1
Optimizador Adamax (lr=0,001)
Función de pérdida=BCE
epochs=60
batch_size=16

k-Nearest Neighbor (kNN)

n_neighbors=7

selección de características

Identificar y elegir un subconjunto relevante de variables del conjunto de datos original a fin de mejorar el rendimiento de los modelos

- Se realizó una selecciónde característicasmanual a partir de la segmentaciónDCL-bin (16 variablesde entrada).
- En cada iteración del proceso, se eliminó un atributo del conjunto de datos.
- Finalmente se llegó a un conjunto de datos con 503 ejemplos y 9 variables de entrada.

	Dataset (D)	N	in
D_0	DCL-bin	503	16
D_1	D_0 - hem___glucosilada	503	15
D_2	D_1 - creatinina_basal	503	14
D_3	D_2 - colesterol_ldl	503	13
D_4	D_3 - trigliceridos	503	12
D_5	D_4 - colesterol_hdl	503	11
D_6	D_5 - colesterol_total	503	10
D_7	D_6 - glucemia_basal	503	9

mODELOS DE REGRESIÓN

Linear Regression (LR)

class_weight=”balanced”

Random Forest (RF)

max_depth=2

Decision Tree (DT)

max_depth=5
class_weight=”balanced”

Artificial Neural Networks (ANN

1 capa oculta con 100 neuronas
Función de activación ReLU
Capa de salida con 1 neurona
Regularización L2 de 0,1
Optimizador Adamax (lr=0,001)
Función de pérdida=MSE
epochs=60
batch_size=16

k-Nearest Neighbor (kNN)

n_neighbors=7

04

resultados obtenidos

La contribución

DCL-bin · 16 variables de entrada - 503 ejemplos

Modelos aplicados en evaluación (clase positiva = “Con riesgo”)

modelo	F-SCORE	precision	RECALL	accuracy
RF	94.77 ± 1.60	98.73 ± 1.46	91.17 ± 2.77	94.56 ± 1.60
DT	93.84 ± 1.50	95.97 ± 2.24	91.88 ± 2.49	93.46 ± 1.58
ANN	91.87 ± 1.90	92.47 ± 2.85	91.37 ± 2.73	91.22 ± 2.06
LOR	86.01 ± 3.3	90.5 ± 3.11	82.1 ± 4.81	85.56 ± 3.21
kNN	71.97 ± 4	75.53 ± 3.87	70.83 ± 5.75	71.62 ± 3.82

DCG-bin · 10 variables de entrada - 1233 ejemplos

Modelos aplicados en evaluación (clase positiva = “Con riesgo”)

modelo	F-SCORE	precision	RECALL	accuracy
RF	92.68 ± 1.3	100 ± 0	86.38 ± 2.25	93.23 ± 1.12
DT	91.39 ± 1.73	96.68 ± 2.75	86.71 ± 2.17	91.87 ± 1.67
ANN	90.65 ± 1.76	94.35 ± 2.34	87.29 ± 2.45	91.05 ± 1.67
LOR	73.05 ± 5.3	80.96 ± 4.44	66.7 ± 6.55	75.66 ± 4.37
kNN	69.2 ± 2.23	70.63 ± 2.32	67.92 ± 3.29	69.97 ± 2.01

DC-bin · 9 variables de entrada - 1233 ejemplos

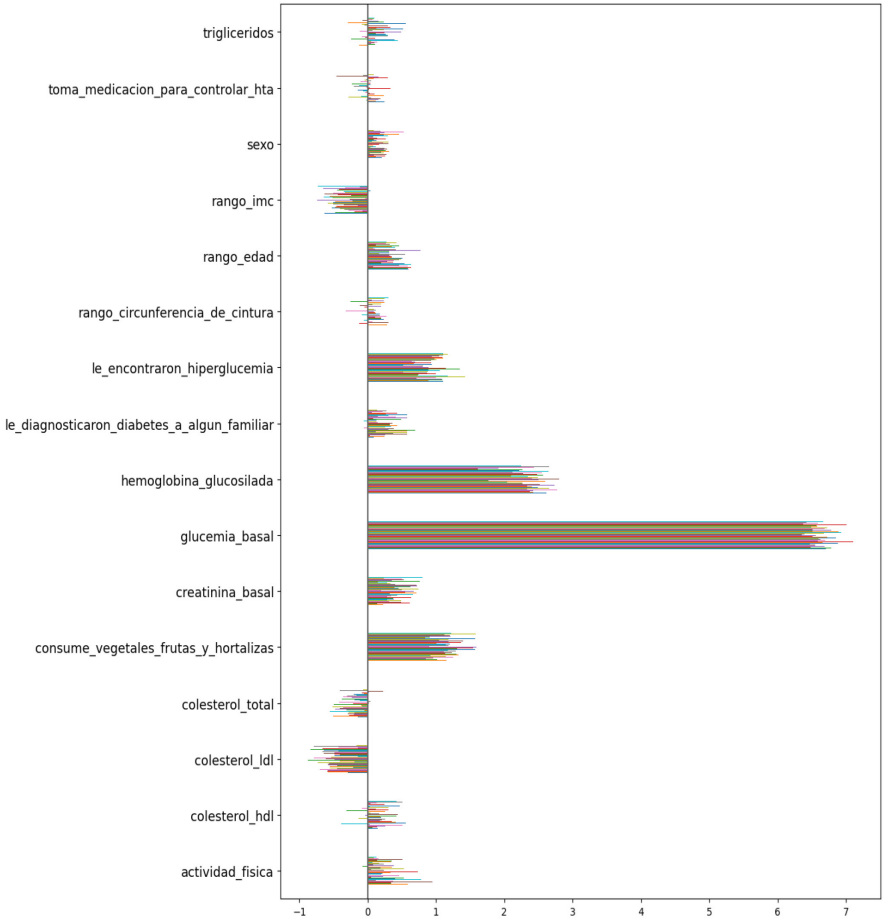
Modelos aplicados en evaluación (clase positiva = “Con riesgo”)

modelo	F-SCORE	precision	RECALL	accuracy
ANN	54.27 ± 3.71	58.61 ± 3.06	51.05 ± 6.43	57.55 ± 2.07
RF	43.61 ± 4.06	63.77 ± 4.13	33.42 ± 4.84	57.33 ± 1.93
LOR	47.94 ± 3.81	60.88 ± 4.27	39.92 ± 5.25	57.16 ± 2.32
DT	52.67 ± 5.01	55.30 ± 3.07	51.16 ± 9.03	54.98 ± 2.47
kNN	53.77 ± 3.03	54.45 ± 2.53	53.32 ± 4.79	54.55 ± 2.45

selección de características

	Dataset (D)	N	in	LOR	DT	k-NN	RF	ANN
D_0	DCL-bin	503	16	85.56%	93.43%	71.62%	94.64%	91.14%
D_1	D_0 - hem___glucosilada	503	15	86.64%	93.27%	70.01%	94.75%	91.60%
D_2	D_1 - creatinina_basal	503	14	86.58%	93.48%	70.61%	94.87%	91.66%
D_3	D_2 - colesterol_ldl	503	13	86.77%	93.34%	71.15%	94.91%	91.52%
D_4	D_3 - trigliceridos	503	12	87.15%	93.01%	74.05%	94.86%	91.68%
D_5	D_4 - colesterol_hdl	503	11	87.06%	92.82%	75.63%	94.95%	92.08%
D_6	D_5 - colesterol_total	503	10	87.63%	93.72%	79.70%	94.85%	92.70%
D_7	D_6 - glucemia_basal	503	9	59.21%	57.21%	56.50%	59.68%	58.61%

Accuracy de los modelos experimentados sobre el conjunto de evaluación utilizando CV



Pesos asociados a cada atributo por el modelo LOR en evaluación para DCL-bin.

DCL · 16 variables de entrada - 503 ejemplos

Modelos aplicados en evaluación

modelo	accuracy	MSE	RMSE	MAE	R2
RF	93.71 ± 1.6	0.07 ± 0.02	0.27 ± 0.04	0.12 ± 0.01	0.78 ± 0.07
DT	91.14 ± 2	0.11 ± 0.03	0.32 ± 0.05	0.11 ± 0.02	0.68 ± 0.09
ANN	88.79 ± 2.48	0.11 ± 0.02	0.33 ± 0.03	0.24 ± 0.02	0.68 ± 0.06
LR	86.42 ± 2.65	0.15 ± 0.02	0.38 ± 0.03	0.30 ± 0.01	0.56 ± 0.07
kNN	67.85 ± 3.41	0.24 ± 0.02	0.49 ± 0.02	0.40 ± 0.02	0.27 ± 0.06

Discusión clasificadores

Resultados comparativo de los modelos de clasificación aplicados en evaluación

DATASET	variables	ejemplos	modelo	F-SCORE	accuracy
DCL-bin	16	503	RF	94.77 ± 1.60	94.56 ± 1.60
			DT	93.84 ± 1.50	93.46 ± 1.58
			ANN	91.87 ± 1.90	91.22 ± 2.06
DCG-bin	10	1233	RF	92.68 ± 1.3	93.23 ± 1.12
			DT	91.39 ± 1.73	91.87 ± 1.67
			ANN	90.65 ± 1.76	91.05 ± 1.67
DG-bin	9	1233	ANN	54.27 ± 3.71	57.55 ± 2.07
			RF	43.61 ± 4.06	57.33 ± 1.93
			LOR	47.94 ± 3.81	57.16 ± 2.32

Los modelos RF, DT y ANN demostraron un gran poder de clasificación

Aunque no es (del todo) correcta la comparación, prácticamente no hay diferencias en las métricas de los modelos para las segmentaciones

Baja significativa en el rendimiento de los modelos

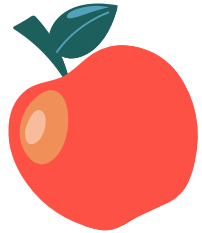
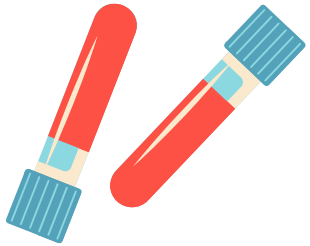
Discusión general

Resultados comparativo de los modelos aplicados en evaluación para las segmentaciones planteadas

DATASET	CLASIFICADOR /REGRESOR	variables	ejemplos	modelo	accuracy	otras métricas
DCL	Regresor	16	503	RF	93.71 ± 1.60	MAE: 0.12± 0.01 MSE: 0.07± 0.02
				DT	91.14 ± 2	MAE: 0.12± 0.01 MSE: 0.07± 0.02
				ANN	88.79 ± 2.48	MAE: 0.12± 0.01 MSE: 0.07± 0.02
DCL-bin	Clasificador	16	503	RF	94.56 ± 1.60	F-score: 92.68± 1.3
DCG-bin	Clasificador	10	1233	RF	93.23 ± 1.12	F-score: 92.68± 1.3
DG-bin	Clasificador	9	1233	ANN	57.55 ± 2.07	F-score: 54.27± 3.71

Los modelos RF, DT y ANN demostraron un gran poder de regresión

Es posible visualizar un patrón de comportamiento independientemente del problema abordado


05

conclusiones y trabajos futuros





conclusiones

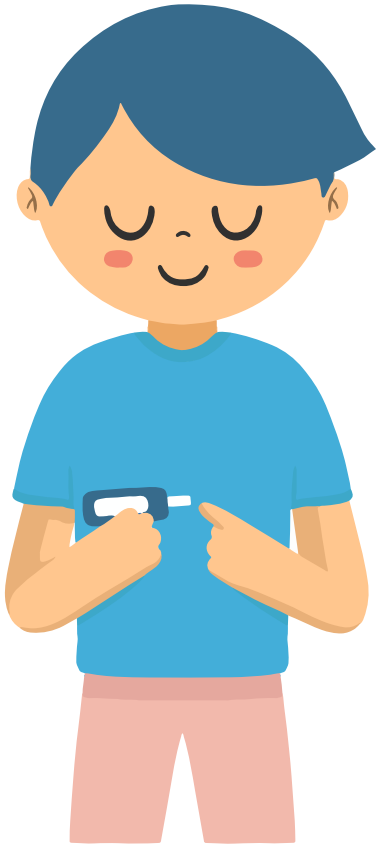
- La mayoría de las variables carecen de relevancia para el problema de clasificación La `glucemia_basal` es el atributo más influyente seguida por la `hemoglobina_glucosilada`. En la ausencia de toda información de laboratorio, `le_encontraron_hiperglucemia` es la más influyente.
- Los modelos RF, DT y ANN demostraron un gran poder predictivo tanto en clasificación como regresión (DCL-bin, DCG-bin y DCL). Los modelos LOR y LR se ubicaron por debajo de ellos, seguido por kNN, con el rendimiento más bajo.
- Aunque no es (del todo) correcta la comparación, prácticamente no hay diferencias en las métricas conseguidas por los modelos para las segmentaciones DCL-bin/DCG-bin/DCL, lo que tendría incidencia importante en la práctica.
- La segmentación que solo considera la información clínica genera un impacto crítico en el rendimiento de los modelos.
- Del proceso de selección de características, se destaca la importancia crítica de la `glucemia_basal` en la evaluación y detección del riesgo de diabetes.
- La binarización de la clase favoreció al balanceo y simplificó el problema.

Debido a limitaciones propias de la base de datos, no es posible afirmar que los resultados sean concluyentes, aunque sí resultan promisorios.

Considerando la vacancia de herramientas de este tipo, este trabajo representa el primer paso hacia modelos más sofisticados.

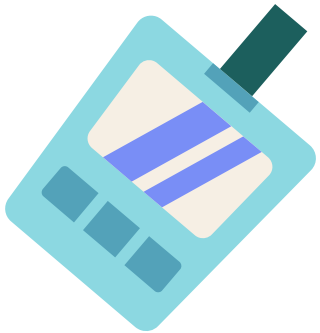
trabajos futuros

- Evaluar modelos de clasificación o regresión adicionales, como *Naïve-Bayes* o *Support Vector Machine*.
- Realizar una búsqueda de hiperparámetros a fin de obtener la mejor versión de los modelos experimentados
- Analizar los valores nulos presentes e implementar diversas técnicas para abordar su completitud.
- Incrementar el tamaño del conjunto de datos, a partir de la recolección de nuevos registros o utilizando técnicas de *data-augmentation*.
- Desarrollar una aplicación web para obtener predicciones en tiempo real.



¿preguntas?

¡Muchas gracias!



Gonzalo Tittarelli

CENEXA
CONICET
U N L P
CEA: CICPBA

Facultad de Informática
Universidad Nacional de La Plata

III-LIDI
Instituto de Investigación en Informática - LIDI