UNAM

Centro de Ciencias de la Complejidad

CONABIO

CHILAM

Plataforma Proyecto 42

species

Epi-species

Epi-PUMA

PROYECTO 42

Conductome

Uso de la plataforma Proyecto 42

ETAPA A — Entrar a la plataforma

https://project42.c3.unam.mx/landing

ETAPA B — Definir el modelo

ETAPA C — Interpretar los resultados

ETAPA A

1 **https://project42.c3.unam.mx/landing**

2

3 Ubicar el mouse de lado izquierdo de la pantalla para ver el menú

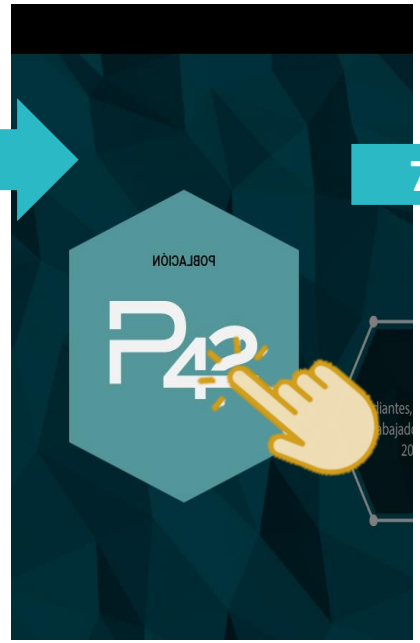
Click en POBLACIONES

4

5

Escoger la población

6



7

Click en MODELOS

ETAPA B

Definir el modelo

Los resultados aparecerán debajo de “APLICAR MODELO” hay que desplazarse hacia abajo

ETAPA B

Ejemplo

1 **POBLACIÓN** 14UNI1_CAQ

2 **SECCIONES**

Seleccionar la sección de la cual se elijaran las variables

3

Datos Personales
Antecedentes Personales
Antecedentes Familiares
Antropometría
Autoevaluación de Salud
Estilo de Vida
Información de Salud
Nutrición
Laboratorio

4

id_sexo
id_edociv
id_gestud

la variable aparece de lado derecho

Variables seleccionadas

edad
id_sexo
obesidad
salud_act
estres_act

Escoger la variable que va a ser dependiente en el modelo

5

Elige la variable dependiente

Variable dependiente las variables seleccionadas aparecen también aquí

Aparece opción de seleccionar Clase Objetivo

6

obesidad

Variable clase

Seleccionar clase
NO
SI

7 SELECCIONAR VARIABLE DEPENDIENTE

8 APLICAR MODELO

Caso de uso: Causas de Obesidad

Construcción del modelo

Queremos predecir pertenencia a una **Clase** (de variable dependiente)

Tomando en cuenta un conjunto de **características X**

Caso de uso: Causas de Obesidad

Pregunta de Investigación:

¿Cuáles son las posibles causas de la obesidad?

Hipótesis:

- Falta de ejercicio
- Consumo excesivo de alimentos, por arriba de lo recomendado
- Ambos

Clase de interés: Presencia de la obesidad

Variable dependiente: categorías de IMC (obesidad, sobrepeso, normopeso, bajo peso)

Predictores: 13 variables

DATOS PERSONALES:
Edad
Puesto
Sexo
Grado de estudios
ANTECEDENTES FAMILIARES:
Madre diabetes
Padre diabetes
Madre sobrepeso
Padre sobrepeso
ANTROPOMETRIA
Obesidad (C)
AUTOEVALUACION DE SALUD
Salud actual
Estrés actual
ESTILO DE VIDA
Ejercicio actual horas
NUTRICIÓN
Cantidad que come comparado con reco
LABORATORIO
Triglicéridos

Variable binaria

p.ej. ¿Come saludable? Sí o No

¿Hace ejercicio? Sí o No

¿Fuma? Sí o No

Variable categórica

Todas las variables se pueden convertir en **variables binarias**

p.ej. la variable ordinal ¿Cómo consideras que es tu condición física actual?

opciones de respuesta

1. Muy mala	sí	no
2. Mala	sí	no
3. Regular	sí	no
4. Buena	sí	no
5. Muy buena	sí	no

Para el clasificador Naïve Bayes cada de las opciones de respuesta se considera como una variable diferente/ clase de interes.

Variable continua – ejemplo IMC

Todas las variables se pueden convertir en **variables binarias**

p.ej. la variable continua del Índice de Masa Corporal (IMC)

cubetas

IMC	Nivel de peso
Por debajo de 18.5	Bajo peso
18.5 – 24.9	Normal
25.0 – 29.9	Sobrepeso
30.0 o más	Obesidad

¡Considerer la cantidad similar de la n (participantes) en cada cubeta!

Variable continua – ejemplo edad

Todas las variables se pueden convertir en **variables binarias**

Usando los deciles-> forma automática en la plataforma si no se definen los rangos/cubetas de otra manera

cubetas

deciles	Rangos de edad	Nx (frecuencias)
1	< 26.5	100
2	26.5 - 30.5	108
3	30.5 - 34.5	121
4	34.5 - 38.5	109
5	38.5 - 42.5	100
6	42.5 - 46.5	111
7	46.5 - 50.5	103
8	50.5 - 54.5	95
9	54.5 - 59.5	113
10	≥ 59.5	116

Selecciona los datos

POBLACIÓN
salud_medicina_1076

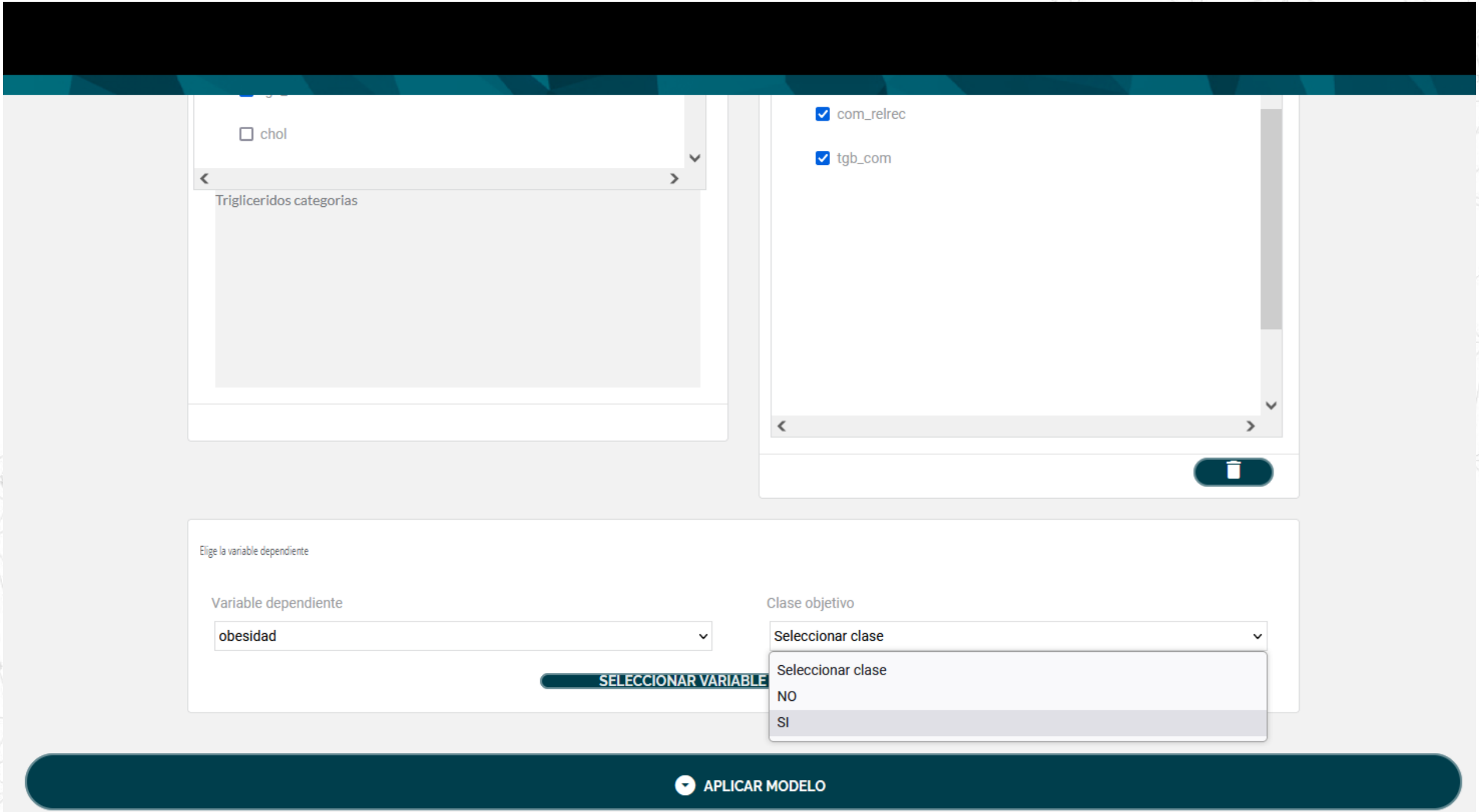
SECCIONES
Laboratorio

- [ ] tgb
- [x] tgb_com
- [ ] chol

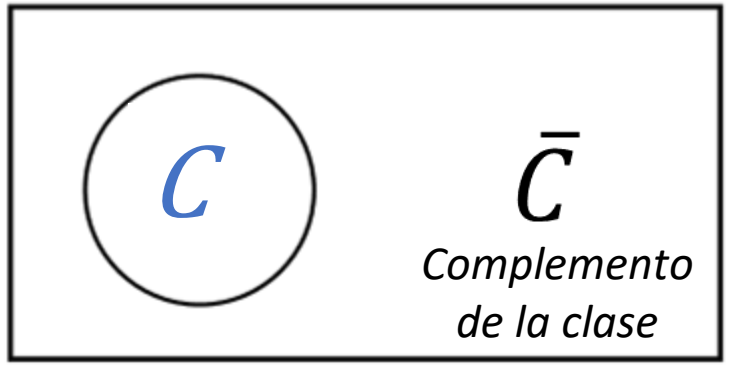
Triglicéridos categorias

Variables Seleccionadas

- [x] edad
- [x] id_puesto
- [x] id_sexo
- [x] id_gestud
- [x] mad_sobr
- [x] pad_sobr
- [x] mad_diab
- [x] pad_diab
- [x] obesidad
- [x] salud_act
- [x] estres_act
- [x] ejer_act


chol
Trigliceridos categorias
com_relrec
tgb_com
Elige la variable dependiente
Variable dependiente
obesidad
Clase objetivo
Seleccionar clase
SELECCIONAR VARIABLE
Seleccionar clase
NO
SI
APLICAR MODELO

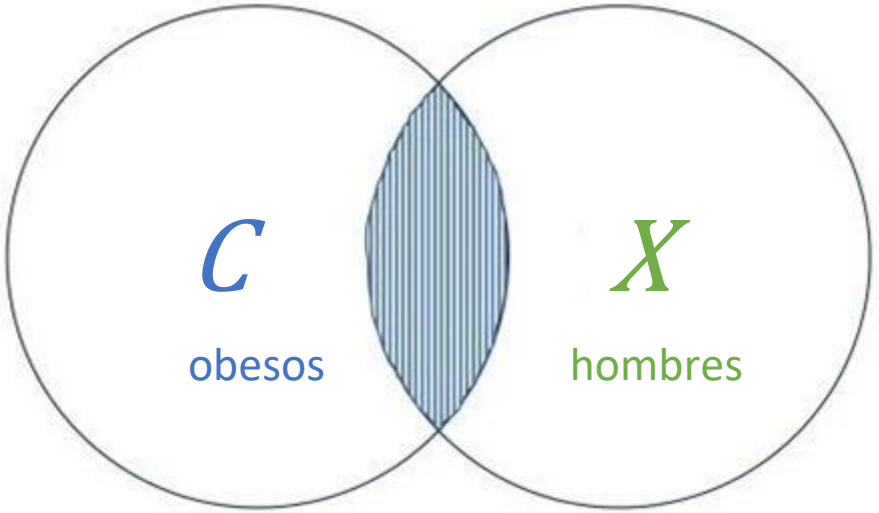
Clase vs NO Clase



Variable Dependiente:	**Clase**:	**NO Clase**:
• Categorías de IMC	• Obesos	• NO Obesos
• Consumo de cigarros	• Fumadores	• NO Fumadores
• Sexo	• Hombres	• …?
• Estatura	• Los con estatura > 1.7m	• …?

Intersección de conjuntos: C ∩ X

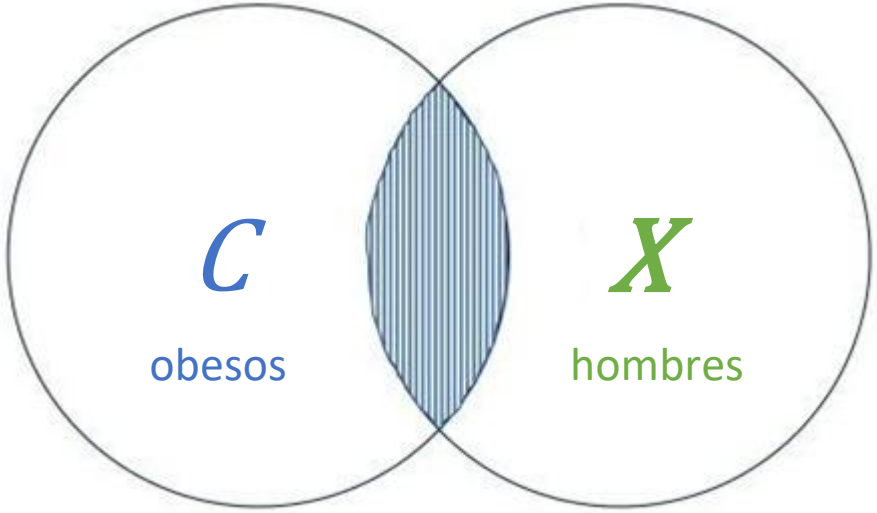
- es el suceso formado por todos los elementos que son, a la vez, de C y X.
- el suceso C ∩ X se verifica cuando ocurren simultáneamente C y X.
- se lee como “C y X".



Probabilidad condicional: P(C|X)

- es la probabilidad de que ocurra **C**, sabiendo que también sucede **X**.

$$P(C|X) = \frac{P(C \cap X)}{P(X)}$$



Probabilidades -> frecuencias

$$P(C|X) = \frac{P(C \cap X)}{P(X)}$$

en el lenguaje de “frecuencias”

$$\frac{N_{CX}}{N_X} = \frac{\frac{N_{CX}}{N}}{\frac{N_X}{N}}$$

Clase de interés: Presencia de la obesidad

Predictores X: 13 variables, p.ej. Sexo, clase: Hombre

N número de personas en toda la muestra

Nx número de personas que son hombres

Ncx número de personas con obesidad y hombres

Tabla de resultados

Los resultados están renqueados por Épsilon, desde el mayor

id	Subcategoría	Valor_Variable	Descripción	Respuesta	Epsilon	Score	nx_c	nx	pc	px_c
10	Autoevaluación de Salud	2	¿Cómo consideras tu salud actual?	MALA	5.15	0.89	29	60	0.21	0.48
10	Autoevaluación de Salud	3	¿Cómo consideras tu salud actual?	REGULAR	4.41	0.45	112	366	0.21	0.31
12	Estilo de Vida	< 0.5	¿Cuántas horas semanales dedicas al ejercicio?	0	4.29	0.41	129	436	0.21	0.3
6	Antecedentes Familiares	1	¿Tiene diabetes? Madre	SÍ	4.26	0.49	85	267	0.21	0.32
4	Datos Personales	CarTec	¿cuál es tu máximo grado de estudios?	Carrera Técnica	4.24	0.65	40	105	0.21	0.38
13	Nutrición	4	¿Cuánto piensas que comes relativo a lo que crees es lo recomendado?	MÁS DE LO RECOMENDADO	3.82	0.36	129	452	0.21	0.29
4	Datos Personales	Sec	¿cuál es tu máximo grado de estudios?	Secundaria	3.53	0.5	47	141	0.21	0.33
2	Datos Personales	Sec	¿Qué puesto ocupas?	Secretaria	3.53	0.6	26	67	0.21	0.39
14	Laboratorio	186.5 - 220.5	Trigliceridos	186.5 - 220.5	3.26	0.5	37	109	0.21	0.34
2	Datos Personales	Vig	¿Qué puesto ocupas?	Vigilante	2.85	0.49	14	34	0.21	0.41
14	Laboratorio	220.5 - 279	Trigliceridos	220.5 - 279	2.68	0.41	34	107	0.21	0.32
2	Datos Personales	Jef	¿Qué puesto ocupas?	Jefe de Área	2.41	0.37	30	96	0.21	0.31
7	Antecedentes Familiares	1	¿Tiene diabetes? (Padre)	SÍ	2.29	0.27	68	251	0.21	0.27
3	Datos Personales	42.5 - 46.5	Edad	42.5 - 46.5	2.2	0.32	33	111	0.21	0.3
2	Datos Personales	Lab	¿Qué puesto ocupas?	Laboratorista	2.06	0.33	16	48	0.21	0.33
2	Datos Personales	Int	¿Qué puesto ocupas?	Intendencia	2.03	0.29	32	110	0.21	0.29

Interpretamos hasta Épsilon = 1.96, que representa el intervalo de confianza de 95%

id	Subcategoría	Valor_Variable	Descripción	Respuesta	Epsilon	Score	nx_c	nx	pc	px_c
10	Autoevaluación de Salud	2	¿Cómo consideras tu salud actual?	MALA	5.15	0.89	29	60	0.21	0.48
10	Autoevaluación de Salud	3	¿Cómo consideras tu salud actual?	REGULAR	4.41	0.45	112	366	0.21	0.31
12	Estilo de Vida	< 0.5	¿Cuántas horas semanales dedicas al ejercicio?	0	4.29	0.41	129	436	0.21	0.3
6	Antecedentes Familiares	1	¿Tiene diabetes? Madre	SÍ	4.26	0.49	85	267	0.21	0.32

El impacto de la **característica** de **REALIZAR 0 HORAS de EJERCICIO** por semana para la **clase OBESIDAD**

- **nx** = 436 personas realizaban **0** horas de ejercicio por semana
- **nx_c** = 128 realizaban **0** horas de ejercicio por semana y tenían OBESIDAD
- **px_c** = 0.29 vs **pc** = 0.21 significa que 29% de las personas que hacen **0** horas de ejercicio eran obesos comparado con 21% en la población en lo general.
- **Epsilon** = 4.27 significa que la diferencia entre 29% y 21% es altamente estadísticamente significativa
- **Score** = 0.41 significa que su contribución al modelo es importante.

$$\text{Score} = \sum \ln \frac{P(Xi \mid C)}{P(Xi \mid \bar{C})} + \ln \frac{P(C)}{P(\bar{C})}$$

Si score de las variables *const.*

Numerador es la probabilidad de C

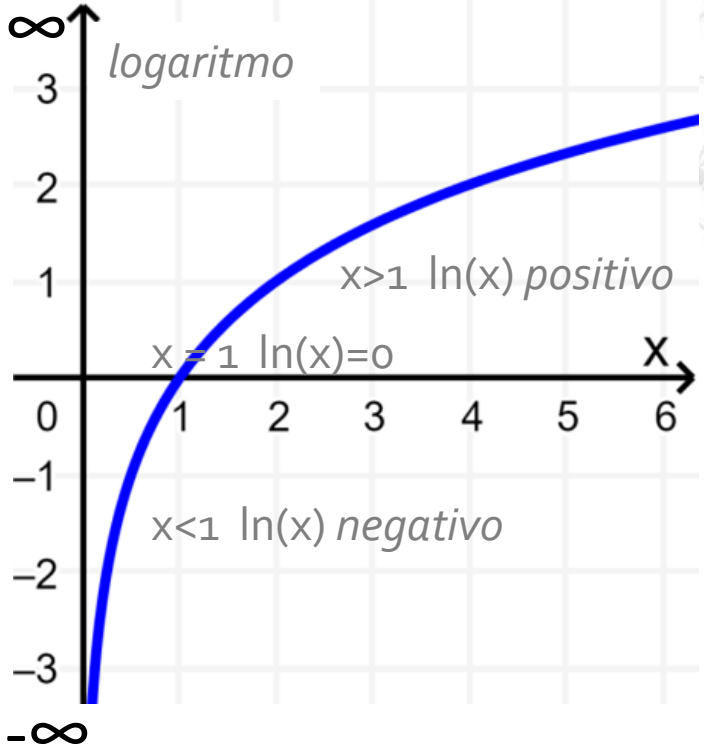
Denominador es la probabilidad de $\bar{C}$

Numerador > Denominador

el coeficiente >1 -> logaritmo/score es ***positivo***
lo que indica que la característica **pertenece** más a la **clase**.

Numerador < Denominador

el coeficiente <1 -> logaritmo/score es ***negativo***
lo que indica que la característica **pertenece** más a **NO clase**.



Score en el lenguaje de frecuencias

$$\text{Score} = \sum \ln \frac{P(Xi \mid C)}{P(Xi \mid \bar{C})} + \ln \frac{P(C)}{P(\bar{C})}$$

En el lenguaje de frecuencias

$$si = \ln \frac{\frac{Nx_iC}{Nc}}{\frac{Nx_i\bar{c}}{N\bar{c}}} = \ln \frac{\frac{Nx_iC}{Nc}}{\frac{Nx_i - Nx_iC}{N-Nc}}$$

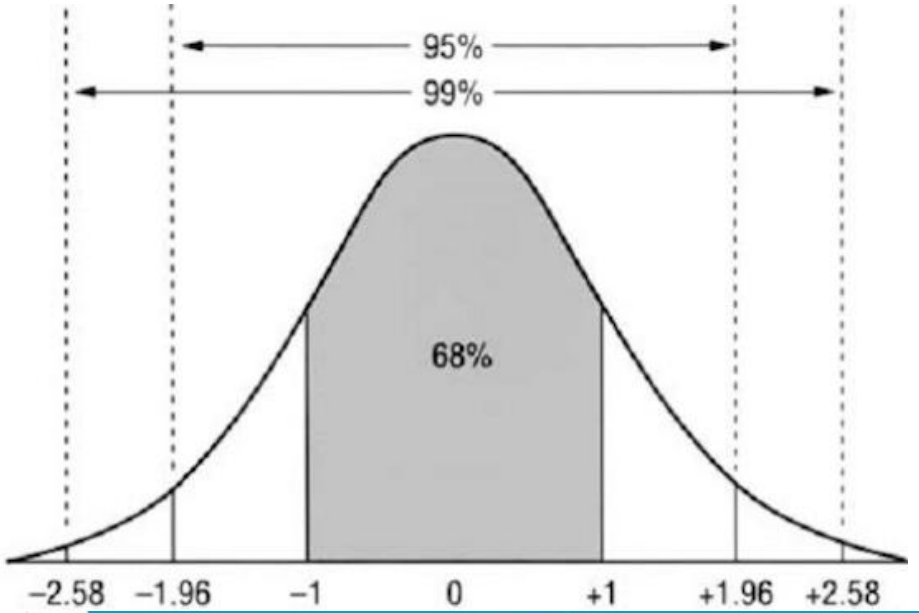
Valores de score pueden estar entre -∞ y +∞

Epsilon Ɛ – nivel de significancia

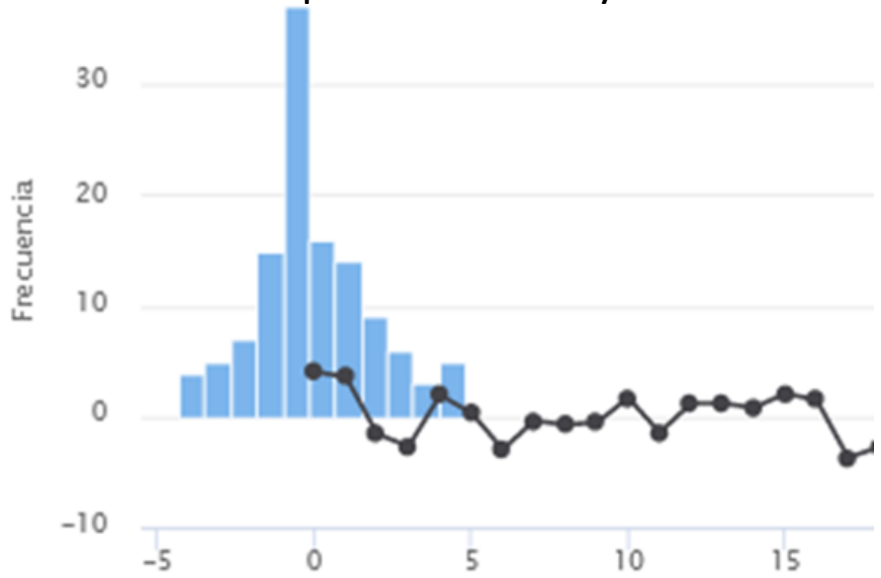
$$Ɛ(C|X) = \frac{Nx * (P(c|X) - P(C))}{\sqrt{Nx * P(C) * (1 - P(C))}}$$

- se buscan valores > 1.96 y <-1.96
- Ɛ depende de la N de la muestra

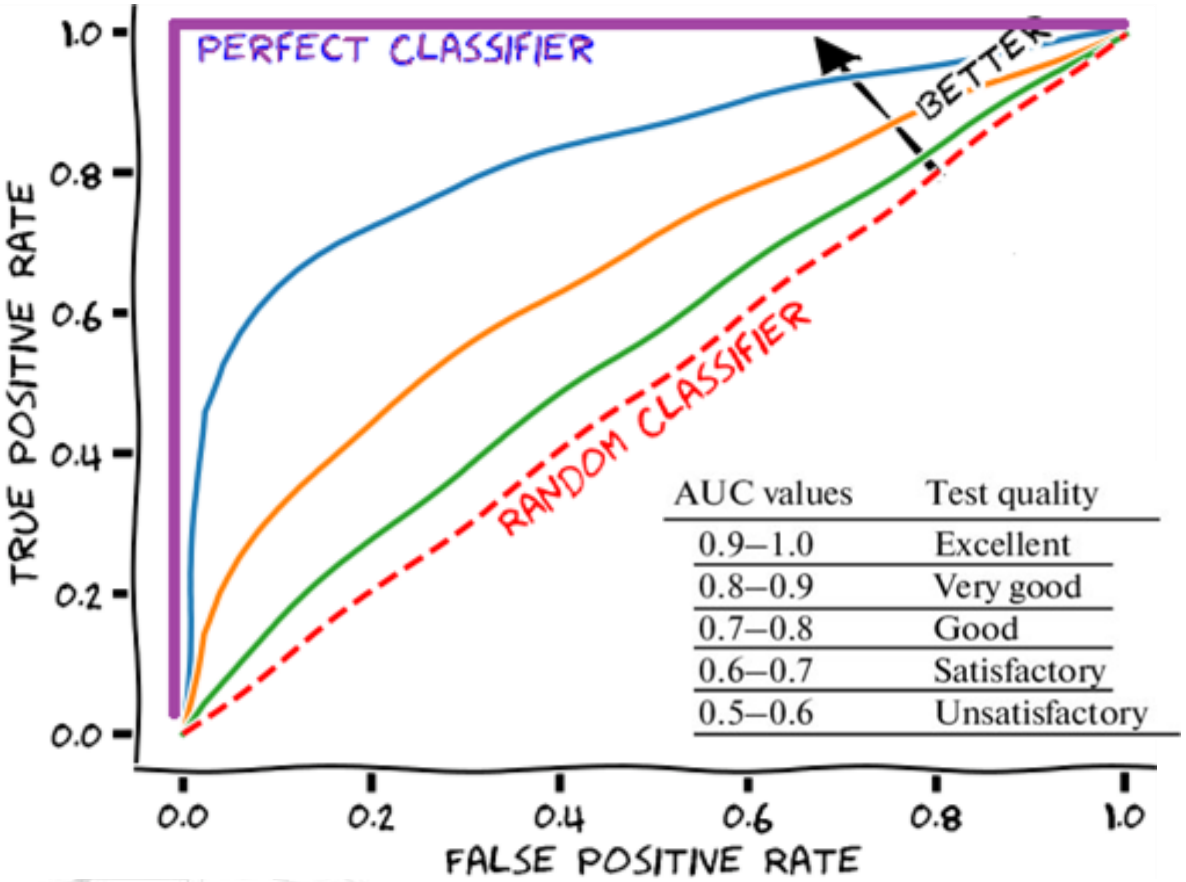
|1.96|, que representa el intervalo de confianza de 95%, 2 SD



En la plataforma Proyecto 42



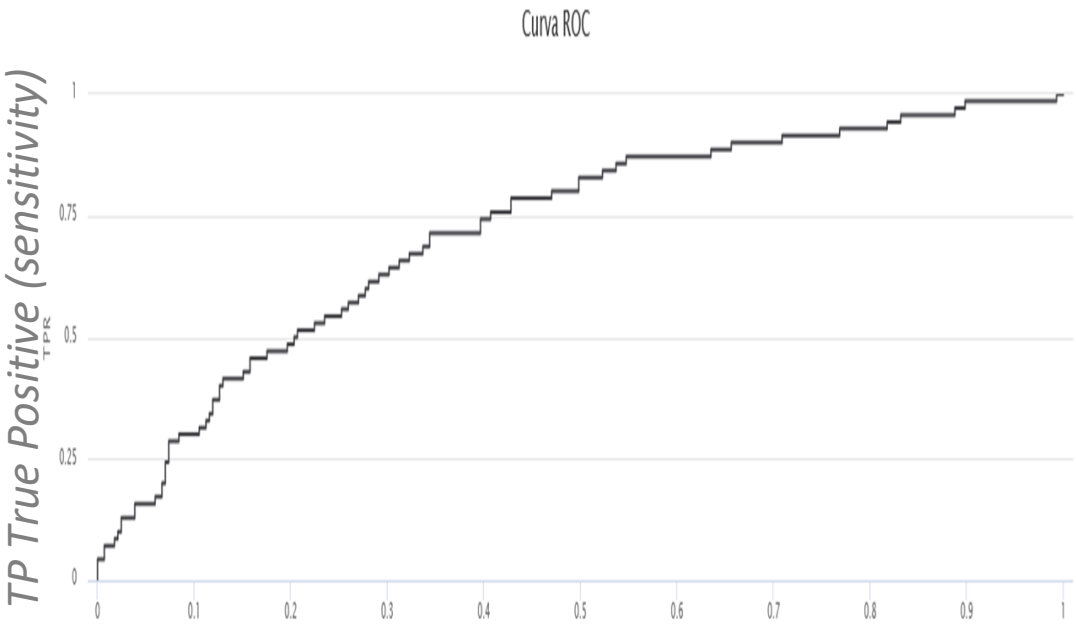
Curva de ROC para evaluar la precisión de las predicciones del modelo



AUC values	Test quality
0.9–1.0	Excellent
0.8–0.9	Very good
0.7–0.8	Good
0.6–0.7	Satisfactory
0.5–0.6	Unsatisfactory

Verdaderos Positivos

TP True Positive (sensitivity)



1- Verdaderos Negativos

1- TN True Negative (specificity)

FPR = FP / (FP + TN) = 1-TNR

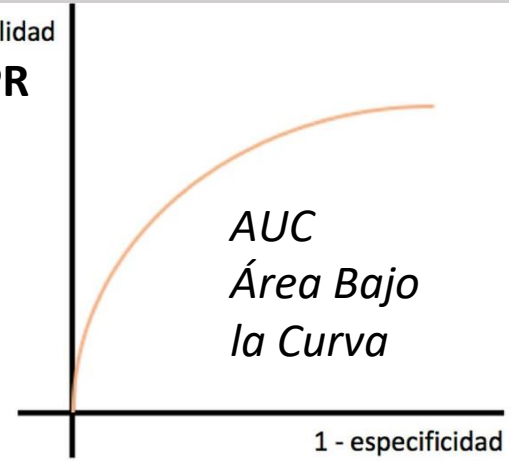
Matriz de confusión

		Predicción del Modelo	
		Positivo	Negativo
Actual	Positivo	**Verdaderos Positivos** *TP True Positive (sensitivity)*	**Falsos Negativos** *FN False Negative*
	Negativo	**Falsos Positivos** *FP False Positive*	**Verdaderos Negativos** *TN True Negative (specificity)*

La persona realmente tiene la obesidad, y el modelo la clasifica como con la obesidad

La persona realmente NO tiene la obesidad, y el modelo la clasifica como SIN la obesidad

Curva ROC
Receiver Operating Characteristic



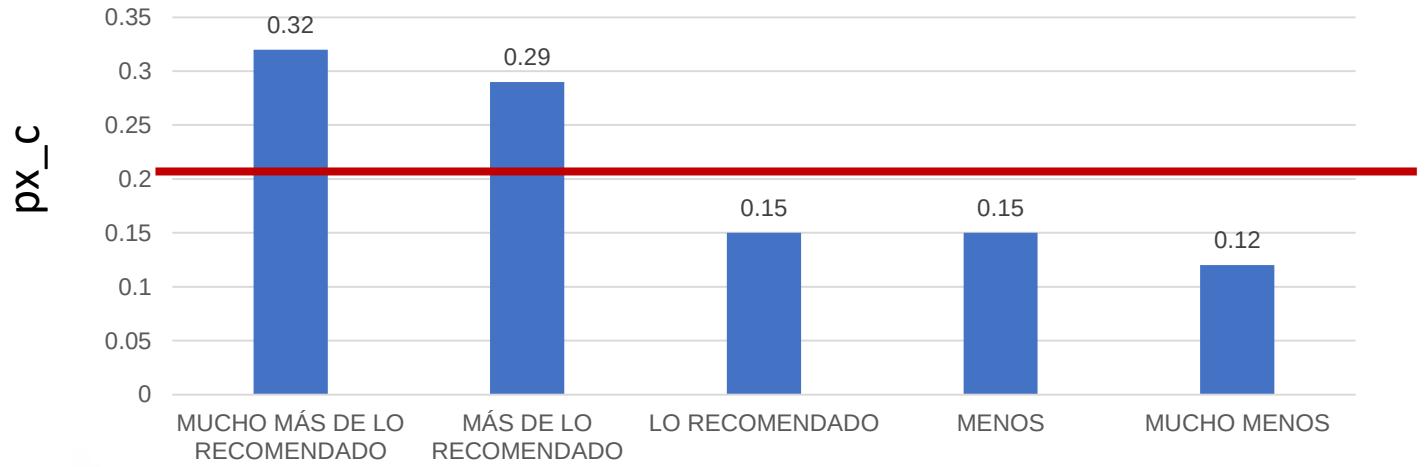
¿Qué me dice Épsilon negativo?

id	Subcategoría	Valor_Variable	Descripción	Respuesta	Epsilon	Score	nx_c	nx	pc	px_c
14	Laboratorio	< 68.5	Trigliceridos	< 68.5	-4.57	-1.79	3	104	0.21	0.03
10	Autoevaluación de Salud	5	¿Cómo consideras tu salud actual?	MUY BUENA	-4.3	-1.71	3	95	0.21	0.03
10	Autoevaluación de Salud	4	¿Cómo consideras tu salud actual?	BUENA	-3.76	-0.46	78	536	0.21	0.15
3	Datos Personales	< 26.5	Edad	< 26.5	-3.72	-1.28	6	100	0.21	0.06
2	Datos Personales	ED	¿Qué puesto ocupas?	Estudiante de Doctorado	-3.58	-1.39	4	81	0.21	0.05
13	Nutrición	3	¿Cuánto piensas que comes relativo a lo que crees es lo recomendado?	LO RECOMENDADO	-3.08	-0.46	54	369	0.21	0.15
4	Datos Personales	Mast	¿cuál es tu máximo grado de estudios?	Maestría	-3	-0.67	22	182	0.21	0.12
2	Datos Personales	Acade	¿Qué puesto ocupas?	Académico	-2.81	-0.54	32	234	0.21	0.14
3	Datos Personales	26.5 - 30.5	Edad	26.5 - 30.5	-2.8	-0.84	11	108	0.21	0.1
4	Datos Personales	Doc	¿cuál es tu máximo grado de estudios?	Doctorado	-2.72	-0.69	17	143	0.21	0.12
6	Antecedentes Familiares	0	¿Tiene diabetes? Madre	NO	-2.5	-0.24	142	807	0.21	0.18
2	Datos Personales	E	¿Qué puesto ocupas?	Estudiante	-2.38	-1.02	4	52	0.21	0.08
12	Estilo de Vida	≥ 9.5	¿Cuántas horas semanales dedicas al ejercicio?	≥ 9.5	-2.33	-1.01	4	51	0.21	0.08
2	Datos Personales	EM	¿Qué puesto ocupas?	Estudiante de Maestría	-2.05	-0.74	8	71	0.21	0.11
12	Estilo de Vida	3.5 - 4.5	¿Cuántas horas semanales dedicas al ejercicio?	3.5 - 4.5	-1.99	-0.69	9	76	0.21	0.12
14	Laboratorio	68.5 - 85.5	Trigliceridos	68.5 - 85.5	-1.89	-0.56	14	103	0.21	0.14

Renqueo por “Descripción X”

Para realizar este renqueo primero hay que descargar la tabla desde la plataforma Proyecto 42

id	Subcategoría	Valor_Variable	Descripción	Respuesta	Epsilon	Score	nx_c	nx	pc	px_c
13	Nutrición	5	¿Cuánto piensas que comes relativo a lo que crees es lo recomendado?	MUCHO MÁS DE LO RECOMENDADO	1.67	0.24	12	37	0.21	0.32
13	Nutrición	4	¿Cuánto piensas que comes relativo a lo que crees es lo recomendado?	MÁS DE LO RECOMENDADO	3.82	0.36	129	452	0.21	0.29
13	Nutrición	3	¿Cuánto piensas que comes relativo a lo que crees es lo recomendado?	LO RECOMENDADO	-3.08	-0.46	54	369	0.21	0.15
13	Nutrición	2	¿Cuánto piensas que comes relativo a lo que crees es lo recomendado?	MENOS	-1.87	-0.49	19	131	0.21	0.15
13	Nutrición	1	¿Cuánto piensas que comes relativo a lo que crees es lo recomendado?	MUCHO MENOS	-0.6	-0.68	1	8	0.21	0.12



pc = 0.21 para la clase de prsencia de obesidad

Entonces ¿cuál es la causa de la obesidad?

- ¿Percibir que tengo mala salud o salud regular?
- ¿No hacer ejercicio?
- ¿Tener madre con diabetes?
- ¿Tener Carrera Técnica o Secundaria como máximo grado de estudios?
- ¿Comer más de lo recomendado?
- ¿Ocupar puesto de secretaria o vigilante?
- ¿Tener triglicéridos altos?
- etc

¿Qué hay atrás de estas características?

UNAM

Centro de Ciencias de la Complejidad

CONABIO

CHILAM

Backchart

species

Epi-species

Epi-PUMA

PROYECTO 42

Conductome

ETAPA C

Interpretar los resultados

Los resultados aparecerán debajo de "APLICAR MODELO", hay que desplazarse hacia abajo.

n frecuencias

p probabilidades

Tabla Resultados

id	Subcategoría	Valor_Variable	Descripción (Pregunta)	Respuesta (Característica X)	Épsilon	Score	nx_c	nx	pc	px_c
5	Datos Personales	T	¿cuál es tu máximo grado de estudios?	CARRERA TÉCNICA	4.05	0.62	39	105	0.21	0.37
6	Nutrición	4.0	¿Cuánto piensas que comes relativo a lo que crees es lo recomendado?	MÁS DE LO RECOMENDADO	3.83	0.36	128	451	0.21	0.28

nx el **número** de personas con la característica X
nx_c el **número** de personas con la característica X y dentro de la clase C

pc la **probabilidad** que alguien está en la clase C independientemente de sus características
px_c la **probabilidad** posterior P(C|X) que es la probabilidad que una persona con característica X está en la clase C (nx_c/nx)

Epsilon una medida de significancia estadística, se buscan valores > |1.96|
Score la contribución de la característica al modelo predictivo

Para descargar la tabla de resultados hay que desplacerse hacia el fin de la tabla

DESCARGAR TABLA