ITBA
Instituto Tecnológico de Buenos Aires

82.05 Análisis Predictivo

Trabajo Práctico N° 1 - Definición del problema y análisis exploratorio

Integrantes: Manuel Dominguez, Danna Mindiuk

Caso de negocio: desarrollo de estrategias y campañas de marketing para promover el dejar de fumar

Objetivos

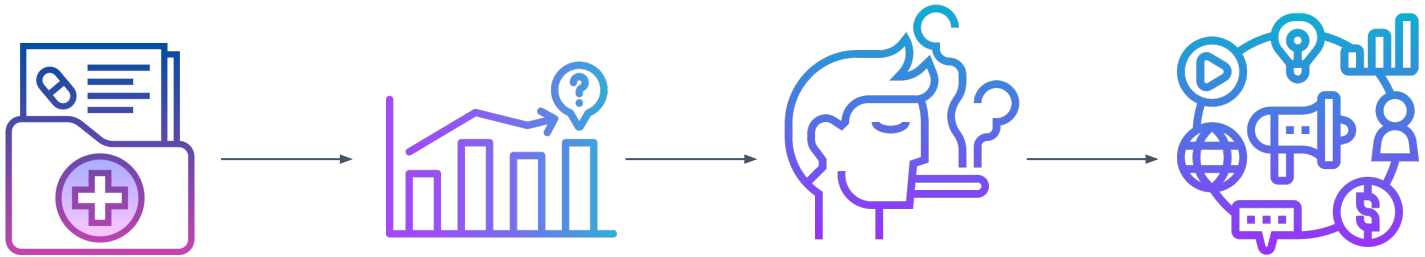
- Predecir con precisión si una persona fuma o no en base a diversos indicadores de salud
- Segmentación de la audiencia fumadora en subgrupos según las distintas características de salud
- Establecer un target concreto para implementar estrategias de marketing
- Desarrollar estrategias de marketing personalizadas para promover la cesación del tabaquismo

Variable respuesta

Buscamos predecir si una persona es o no fumadora, por lo que la variable respuesta a predecir es ***smoking***

¿Cómo funcionaría el modelo en producción?

Se parte de una base de datos con información sobre la salud de cierta cantidad de personas y se busca predecir cuáles de ellas son fumadoras. Una vez obtenidas estas conclusiones, se desarrollará una estrategia de marketing personalizada y con estas personas como target, con el fin de incentivarlas a dejar de fumar.



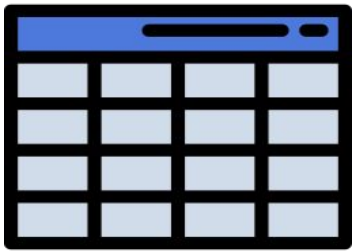
Dataset: descripción de los datos

Descripción del dataset utilizado

Se trabajó con el dataset ***ML Olympiad - Smoking Detection in Patients***, el cual contiene información sobre ciertas métricas de salud de una gran cantidad de personas, junto a una variable que indica si éstas fuman o no.

A partir de los datos, se creó un dataframe con las siguientes características:

- 159.256 filas
- 24 columnas
- Contiene 6 variables categóricas y 18 numéricas
- No contiene valores nulos
- Las variables toman valores de tipo int o float



Descripción de las variables iniciales

`id`	id que identifica a cada registro
`age`	edad de la persona
`weight (kg)`	peso de la persona en kilogramos
`waist (cm)`	tamaño de la cintura de la persona en centímetros
`height (cm)`	altura de la persona en centímetros
`eyesight (left)`	nivel de visión del ojo izquierdo de la persona
`eyesight (right)`	nivel de visión del ojo derecho de la persona
`hearing (left)`	nivel auditivo del oído izquierdo de la persona, toma los valores 1 o 2
`hearing (right)`	nivel auditivo del oído derecho de la persona, toma los valores 1 o 2
`systolic`	presión sistólica de la persona en milímetros de mercurio (mm Hg)
`relaxation`	presión diastólica de la persona en milímetros de mercurio (mm Hg)

`fasting blood sugar`	glucemia en ayunas, medido en miligramos de glucosa por decilitros de sangre (mg/dL)
`Cholesterol`	colesterol de la persona, medido en miligramos de colesterol por decilitro de sangre (mg/dL)
`triglyceride`	cantidad de triglicéridos en sangre, medido en en miligramos de triglicéridos por decilitro de sangre (mg/dL)
`HDL`	colesterol “bueno” de la persona, medido en miligramos de colesterol por decilitro de sangre (mg/dl)
`LDL`	colesterol “malo” de la persona, medido en miligramos de colesterol por decilitro de sangre (mg/dl)
`hemoglobin`	cantidad de hemoglobina en sangre, medido en gramos de hemoglobina por decilitro de sangre (gr/dL)
`Urine protein`	cantidad de proteína en la orina, medio en miligramos de proteína por decilitro de orinal (mg/dL)
`serum creatinine`	cantidad de creatinina en sangre, medido en miligramos de creatinina por decilitro de sangre (mg/dL)

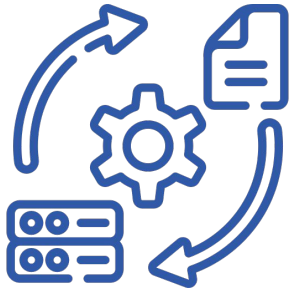
AST	cantidad de enzimas AST en sangre, medido en unidades de la enzima por litro de sangre (U/L)
ALT	cantidad de enzimas ALT en sangre, medido en unidades de la enzima por litro de sangre (U/L)
Gtp	cantidad de enzimas GGT en sangre, medido en unidades de la enzima por litro de sangre (U/L)
dental caries	variable binaria que indica si la persona tiene caries o no, tomando los valores de 1 o 0, respectivamente
smoking	variable binaria que indica si la persona fuma o no, tomando los valores de 1 o 0, respectivamente

Transformaciones realizadas

Variables *age, weight (kg)* y *height (cm)*

- Estas variables toman valores en rangos de 5 unidades, pero encontramos valores por fuera de estos rangos.
- Decidimos aproximar estos valores al rango correspondiente según su cercanía al mismo, ya que se trataba de pocos registros y los valores eran muy cercanos a los intervalos establecidos.
- Para cada una de estas variables, se modificó la siguiente cantidad de registros:

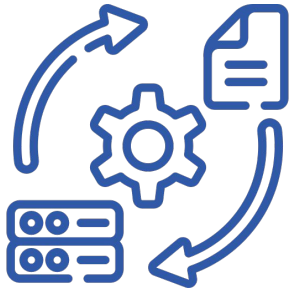
age	*weight (kg)*	*height (cm)*
4	7	2



Creación de nuevas variables

A partir de las variables del dataset, creamos las siguientes nuevas variables para profundizar el análisis:

- BMI
- BMI_Cat
- Hipertensión
- Cholesterol_Cat
- LDL_Cat
- HDL_Cat
- triglyceride_Cat
- Anemic

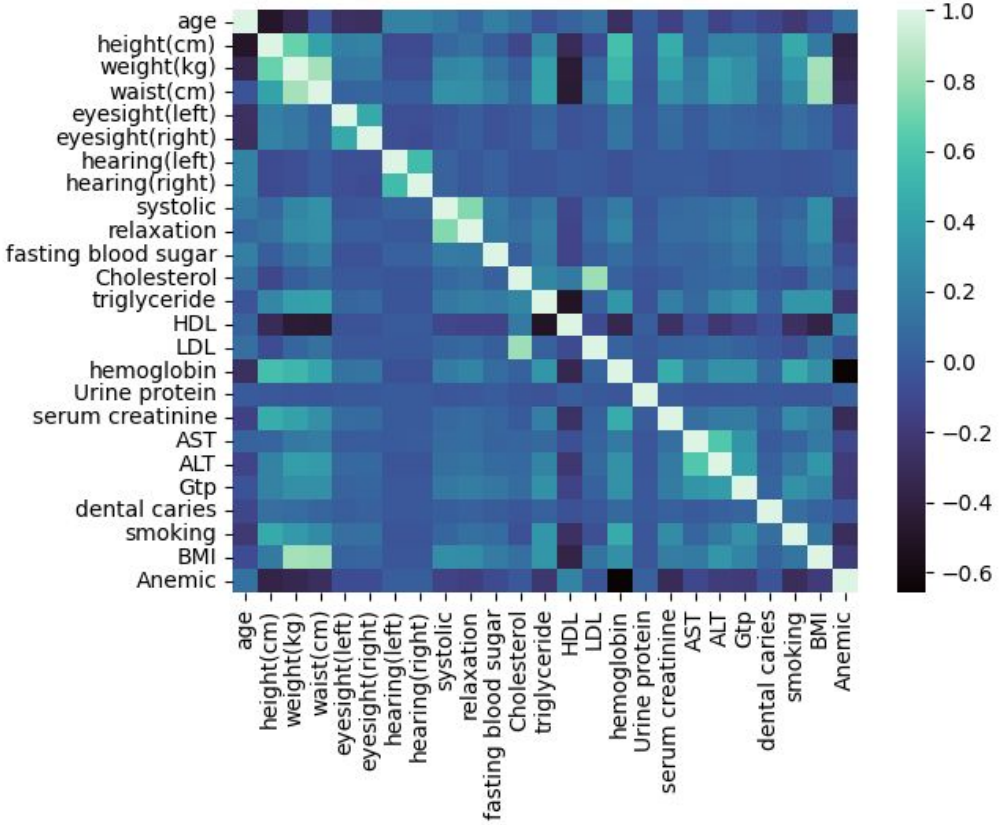


Descripción de las variables creadas

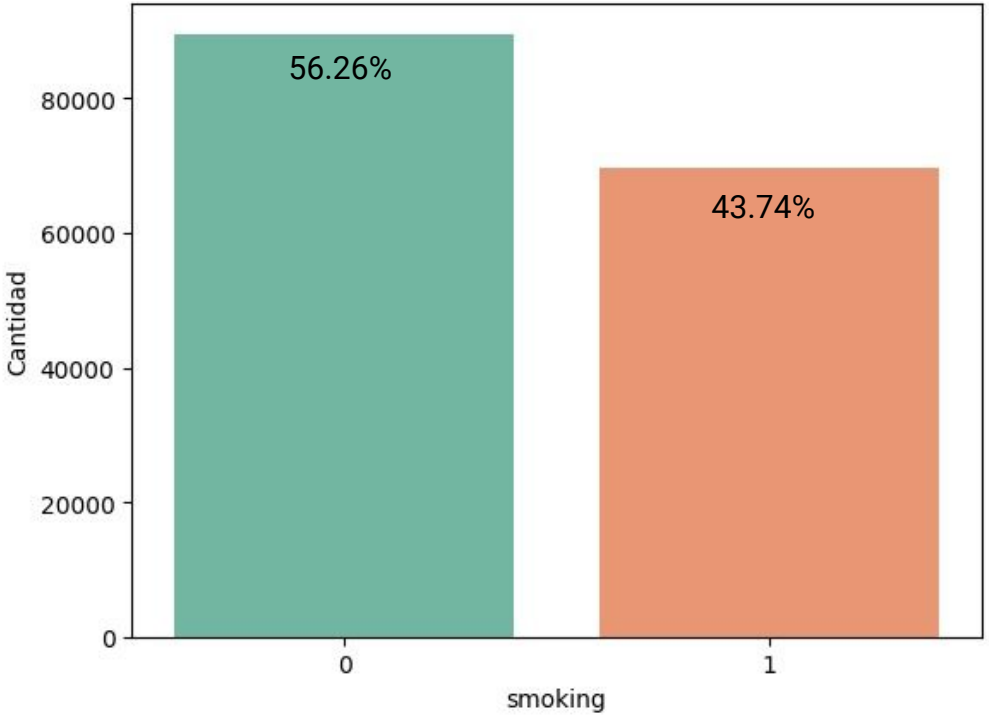
`BMI`	índice de masa corporal, calculado a partir de ***weight (kg)*** y ***height (cm)***
`BMI_Cat`	categoría del BMI, pudiendo tomar los valores *Underweight, Normal, Overweight, Obese G1, Obese G2* y *Obese G3*
`Hipertension`	categoría creada a partir de ***systolic***, pudiendo tomar los valores *Normal, Elevated, Ita G1, Alta G2 y Emergencia*
`Cholesterol_Cat`	categoría creada a partir de ***Cholesterol***, pudiendo tomar los valores *Óptimo, Límite* y *Alto*
`LDL_Cat`	categoría creada a partir de ***LDL***, pudiendo tomar los valores *Óptimo, Casi Óptimo, Límite, Alto* y *Muy Alto*
`HDL_Cat`	categoría creada a partir de ***HDL***, pudiendo tomar los valores *Óptimo, Normal y Alto*
`triglyceride_Cat`	categoría creada a partir de ***triglyceride***, pudiendo tomar los valores *Óptimo, Normal y Alto*
`Anemic`	categoría creada a partir de ***hemoglobin***, pudiendo tomar los valores *0* y *1*, según sea anémica la persona o no

Análisis Exploratorio

Correlación entre las variables



Cantidad de fumadores y no fumadores



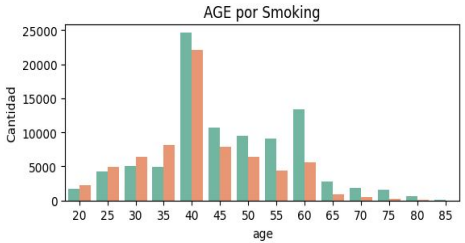
Cantidad de personas que fuman

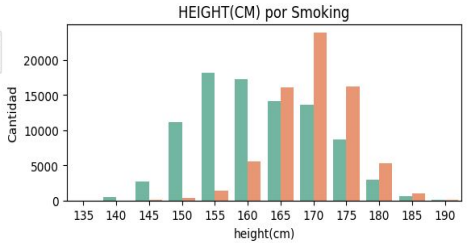
69653

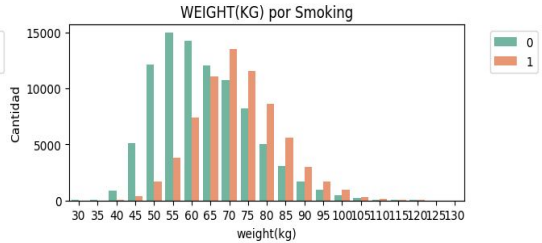
Cantidad de personas que no fuman

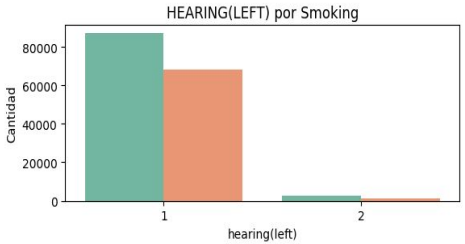
89603

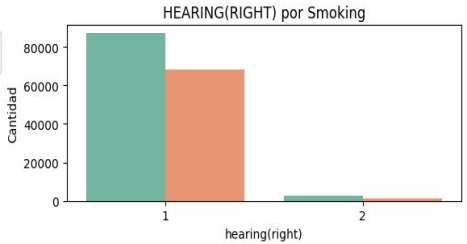
Distribuciones de las variables discretas

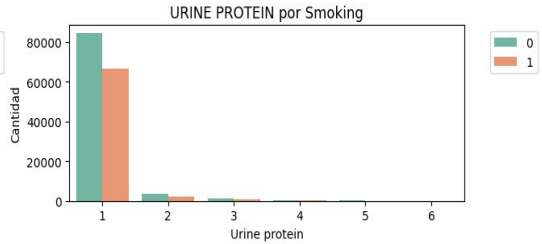


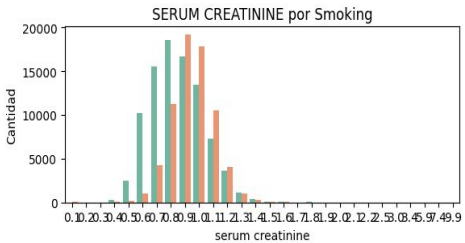


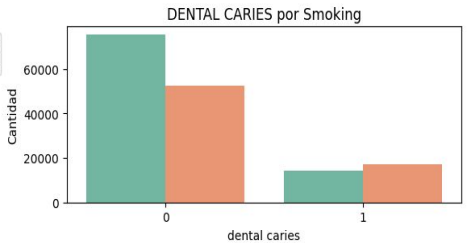


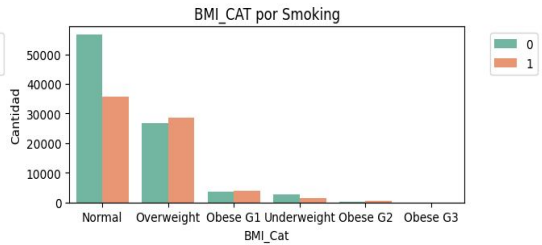




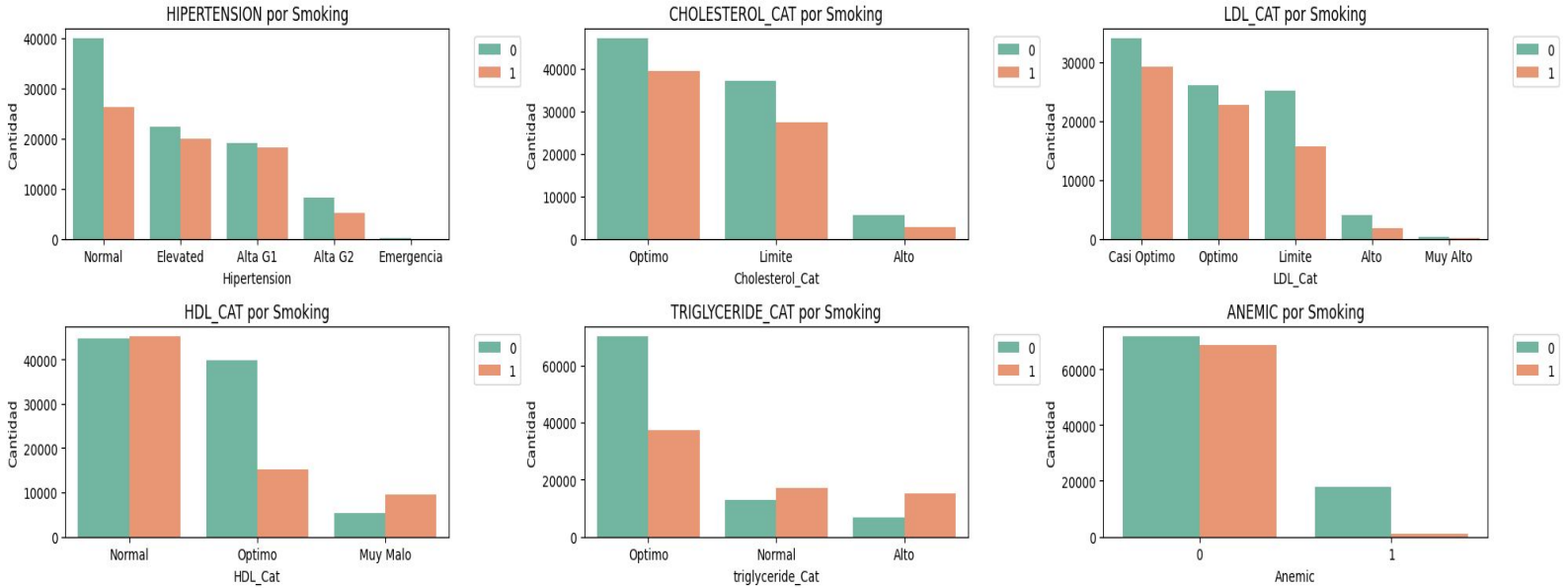




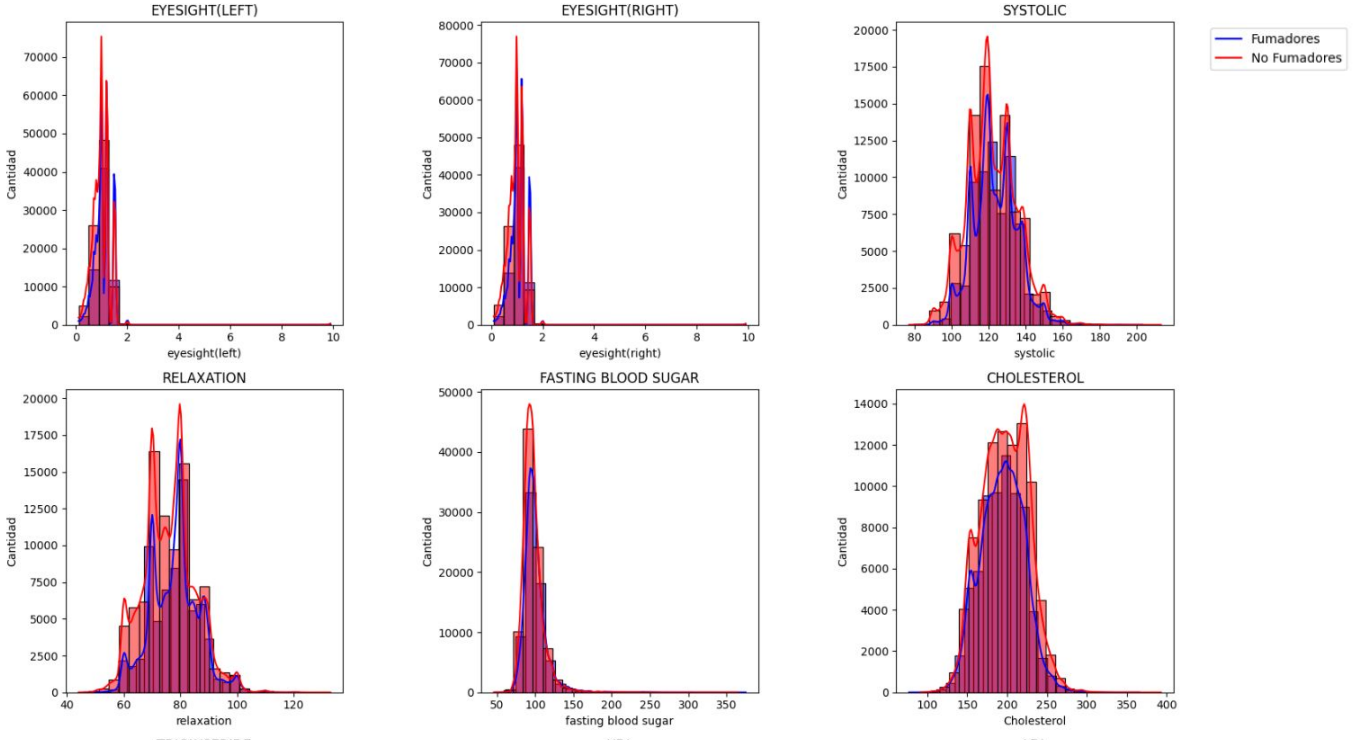




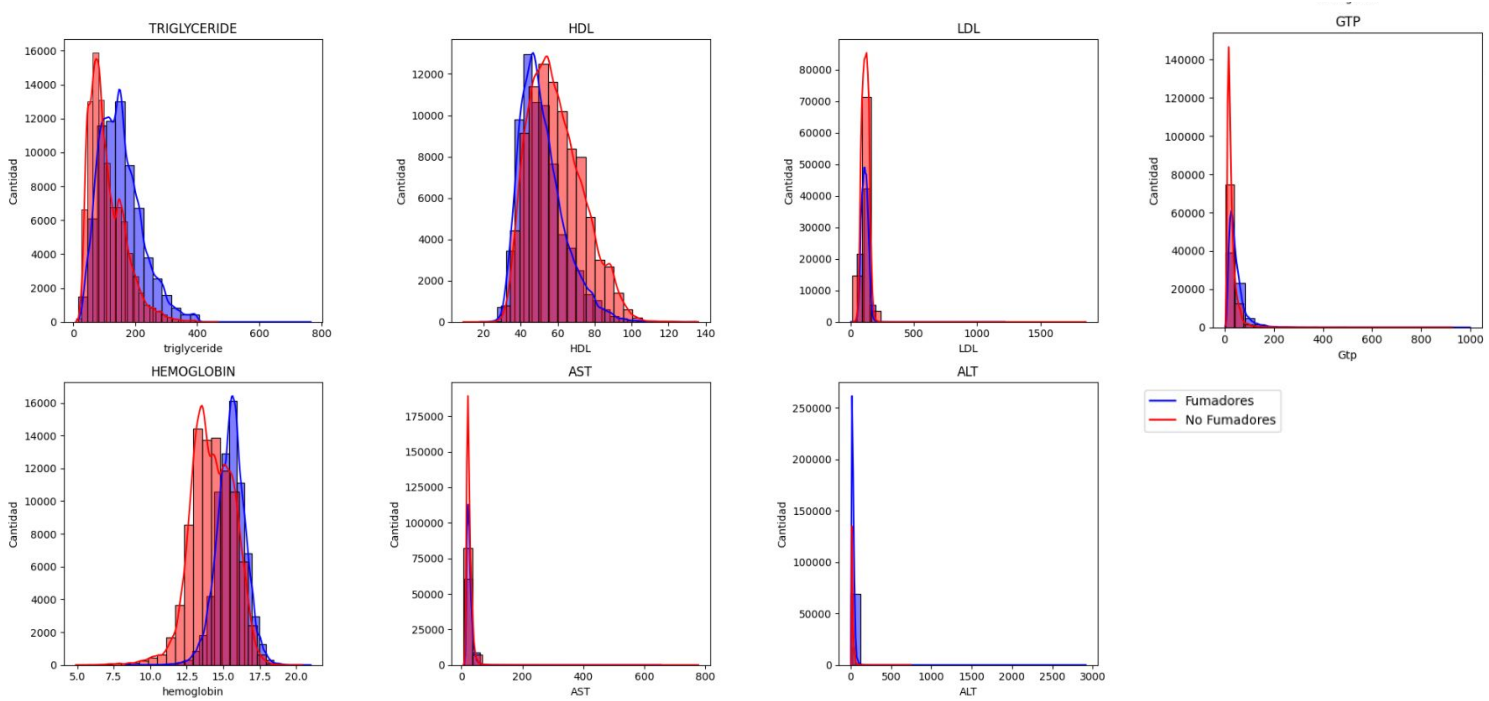
Distribuciones de las variables discretas



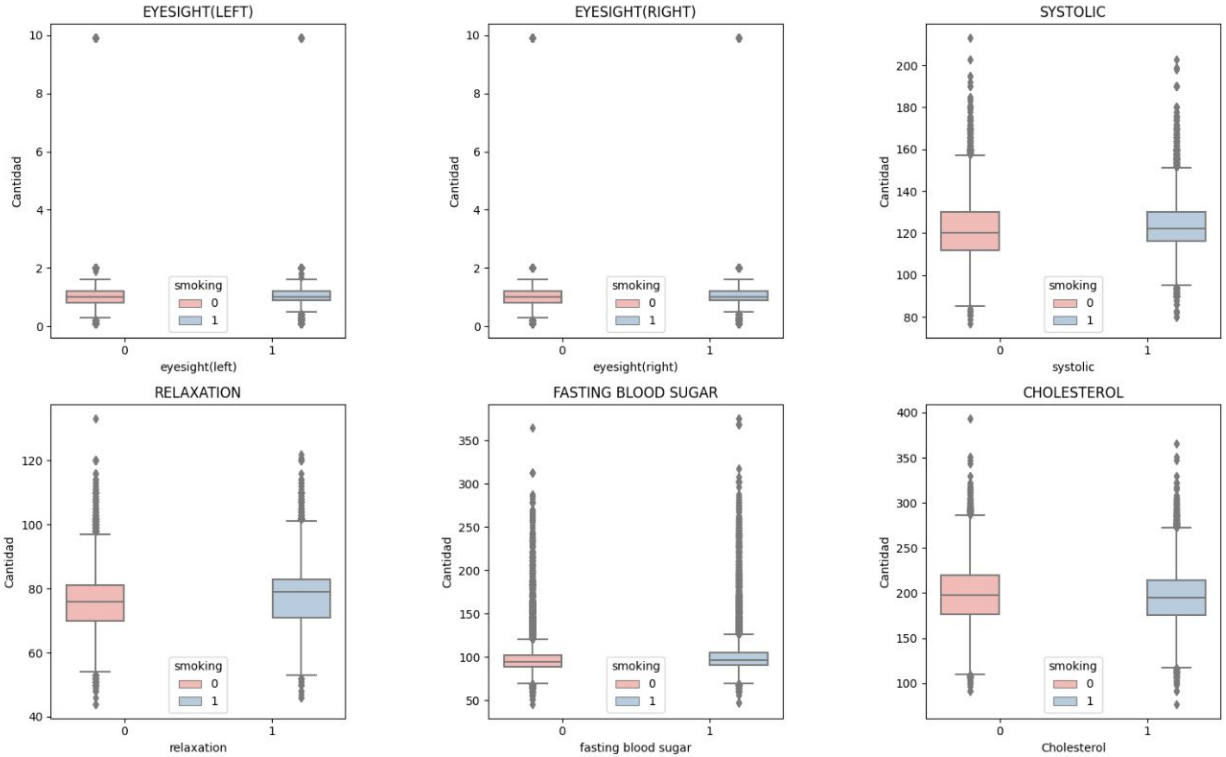
Distribuciones de las variables continuas



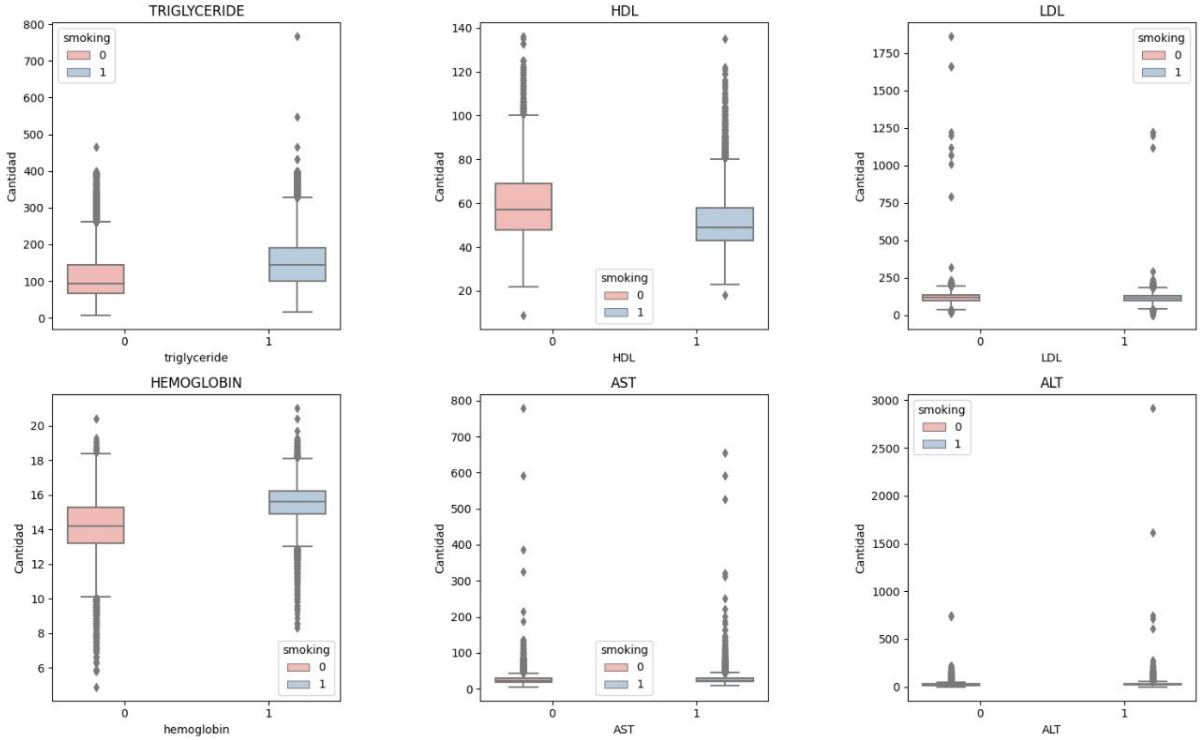
Distribuciones de las variables continuas

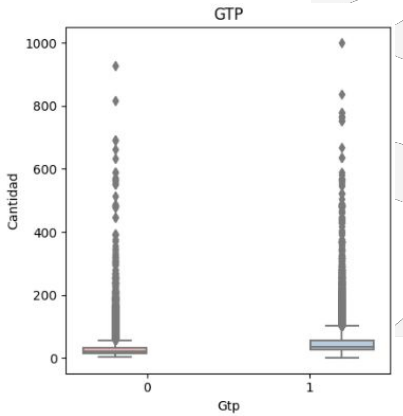


Detección y análisis de los outliers



Detección y análisis de los outliers



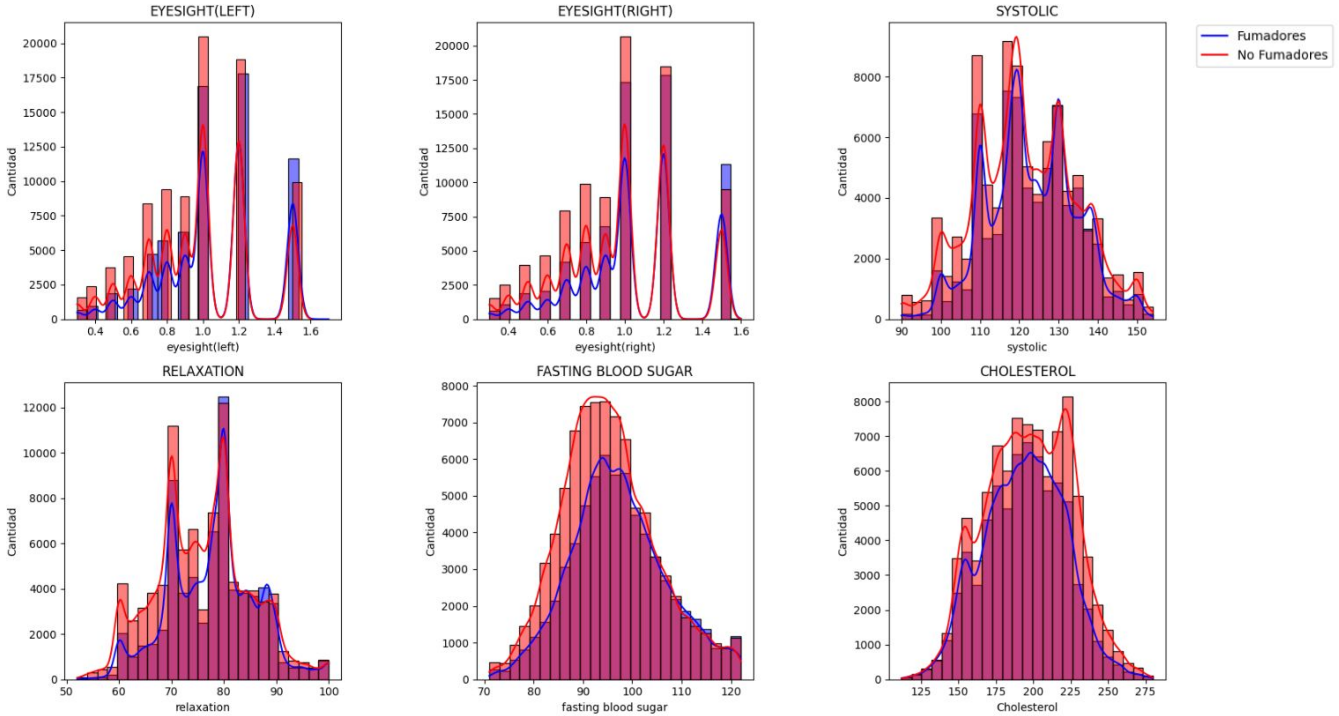


Detección y análisis de los outliers

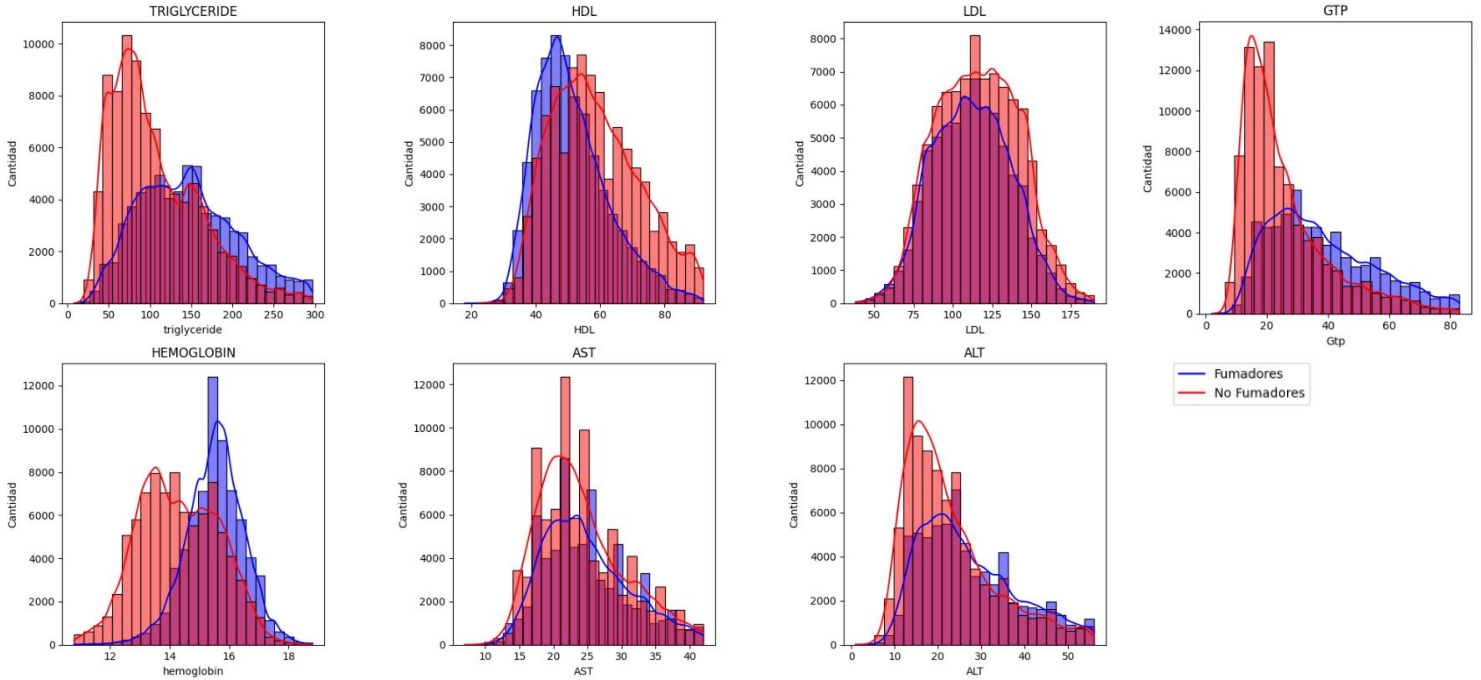
Porcentaje de outliers por variable continua

eyesight (left)	0.01%
eyesight (right)	0.02%
systolic	0.01%
relaxation	0.01%
fasting blood sugar	0.05%
triglyceride	0.02%
HDL	0.01%

Distribución de variables continuas sin outliers



Distribución de variables continuas sin outliers



Métricas por variable según fumadores y no fumadores - Media

	No Fumadores	Fumadores
Age	46,45	41,53
Systolic	121,84	12,33
Relaxation	76	77,99
Hemoglobin	14,22	15,52
Urine protein	1,08	1,06
Mediana AST	23	24
BMI	24,03	24,99
Cholesterol	197,09	194,12
Serum creatinine	0,84	0,94
Mediana Gtp	21	37
Weight(kg)	63,24	72,16

Test de Chi - Cuadrado para las variables categóricas

Variable	Estadístico Chi Cuadrado
BMI_Cat	2791.44
Hipertension	1206.87
Cholesterol_Cat	526.27
LDL_Cat	1207.60
HDL_Cat	9821.38
triglyceride_Cat	11611.62
Anemic	13245.09

Para un nivel de significación del 1%, se rechaza la hipótesis nula para todas las variables y se puede afirmar que existe dependencia entre éstas y la variable ***smoking***

Conclusiones

Conclusiones del análisis exploratorio

- A partir de los test de Chi Cuadrado pudimos confirmar la existencia de dependencia entre las variables categóricas creadas y la variable respuesta
- Analizando las medias de las variables más relevantes por grupo, con el objetivo de describir a la persona media fumadora y no fumadora, encontramos que presentan similitudes y están en los mismos grupos, aunque la persona fumadora tiende a acercarse más a grupos menos saludables.
- Las personas fumadoras tienden a tener menos HDL (colesterol bueno) y más LDL (colesterol malo). Además tienden a presentar valores en AST y Gtp que se asocian con un mayor riesgo de presentar daño en el hígado.
- Consideramos crucial para la continuación del análisis la segmentación de los fumadores en subgrupos según los distintos niveles de cada variable. Esto permitirá comprender mejor los perfiles y comportamientos de los fumadores, identificando patrones específicos de riesgo. Esta segmentación nos facilitará desarrollar estrategias de marketing más precisas y efectivas, adaptadas a las necesidades de cada grupo.

ITBA
Instituto Tecnológico de Buenos Aires

¡MUCHAS GRACIAS!

MÁS INFORMACIÓN > **www.itba.edu.ar**