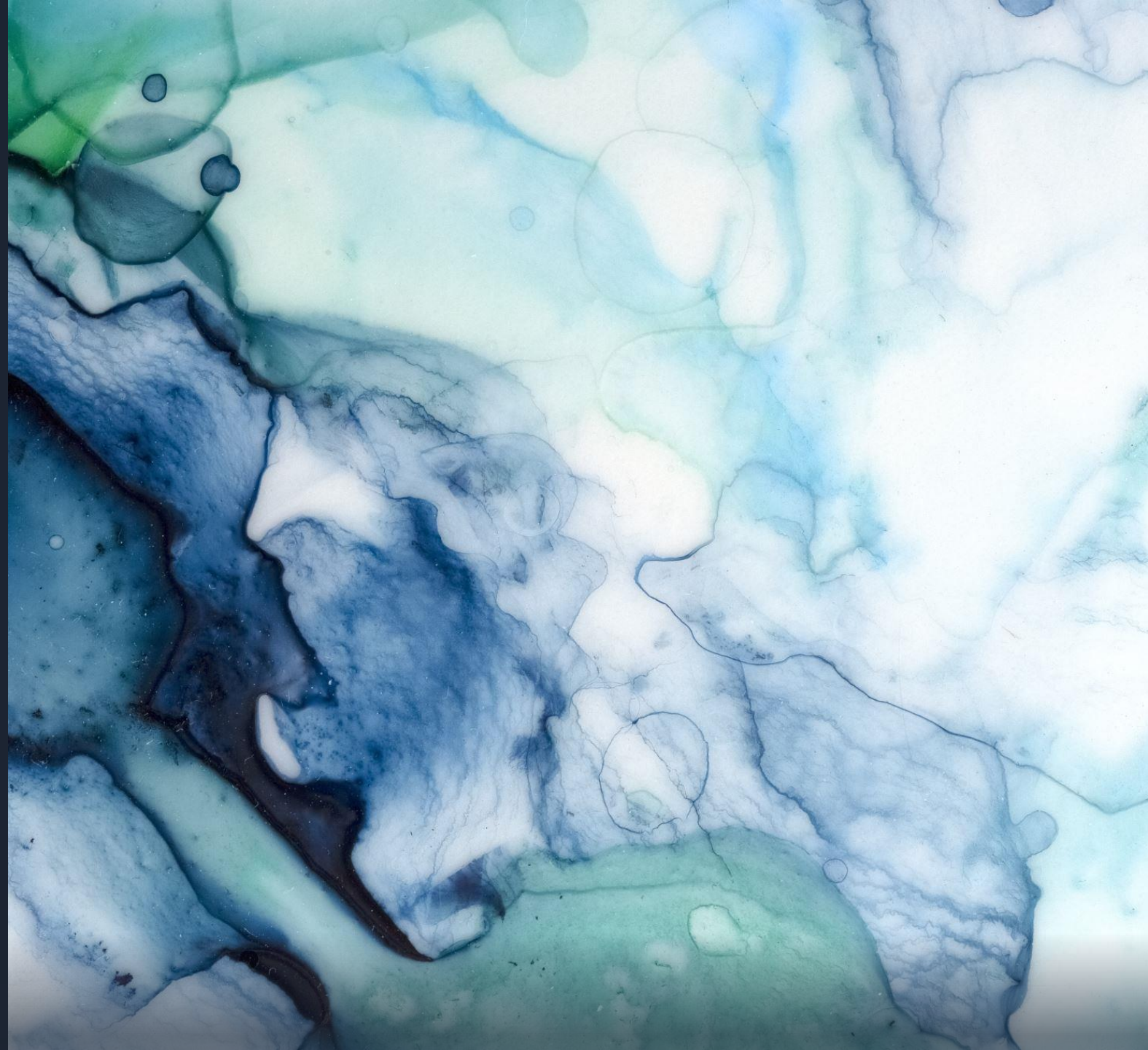
# Proyecto final

Gabriel Lizana

Modelo de estimación de enfermedades cardíacas



# Índice

# Objetivos

- Con la utilización de una base de datos pública, se intenta modelar un algoritmo de *machine learning* que pueda predecir enfermedades cardíacas.

- Los datos provienen de la *Behavioral Risk Factor Surveillance System* (BRFSS) que mediante una encuesta telefónica releva datos de salud de la población de EEUU.

# Machine Learning como herramienta de diagnóstico

- El machine learning es una rama de la inteligencia artificial que permite crear sistemas capaces de aprender de los datos y mejorar su rendimiento sin necesidad de programación explícita. Esta tecnología ofrece grandes oportunidades para el diagnóstico temprano de condiciones de salud, ya que puede analizar grandes cantidades de información y detectar patrones ocultos que podrían indicar un riesgo potencial.
- Uno de los aspectos más interesantes del machine learning es que puede utilizar los pequeños hábitos que tienen las personas en su día a día como fuente de datos. Por ejemplo, el consumo de alimentos, el nivel de actividad física, el sueño, el estrés, el humor, etc. Estos hábitos pueden reflejar el estado de salud de una persona y servir como indicadores de posibles enfermedades cardíacas, diabetes, depresión, etc.
- El proyecto tiene como objetivo desarrollar un modelo de machine learning que pueda predecir la probabilidad de padecer una enfermedad cardíaca a partir de los hábitos diarios de las personas. Para ello, utilizaremos un conjunto de datos que contiene información sobre variables clínicas, demográficas y conductuales de más de 10.000 individuos. Nuestro modelo podrá proporcionar una estimación del riesgo de enfermedad cardíaca para cada persona y sugerir acciones preventivas para mejorar su salud.

# Los datos

- A partir de la encuesta telefónica BRFSS que tiene como objetivo relevar datos que cualquier individuo es capaz de responder, se realiza una supervisión de la condición de salud de la población. En el proyecto se implementa la base de datos de la encuesta del 2021. Entre otras, se relevaron las siguientes variables a través de una o varias preguntas relacionadas

- La base de datos se descargó desde kraggle como parte de un proyecto público y ya estaba facilitada la limpieza de datos.

## Variables del Dataset

1) *Variables subjetivas*
General_Health: ¿En qué condición de salud se encuentra?
Checkup: Última vez que se realizó un control médico

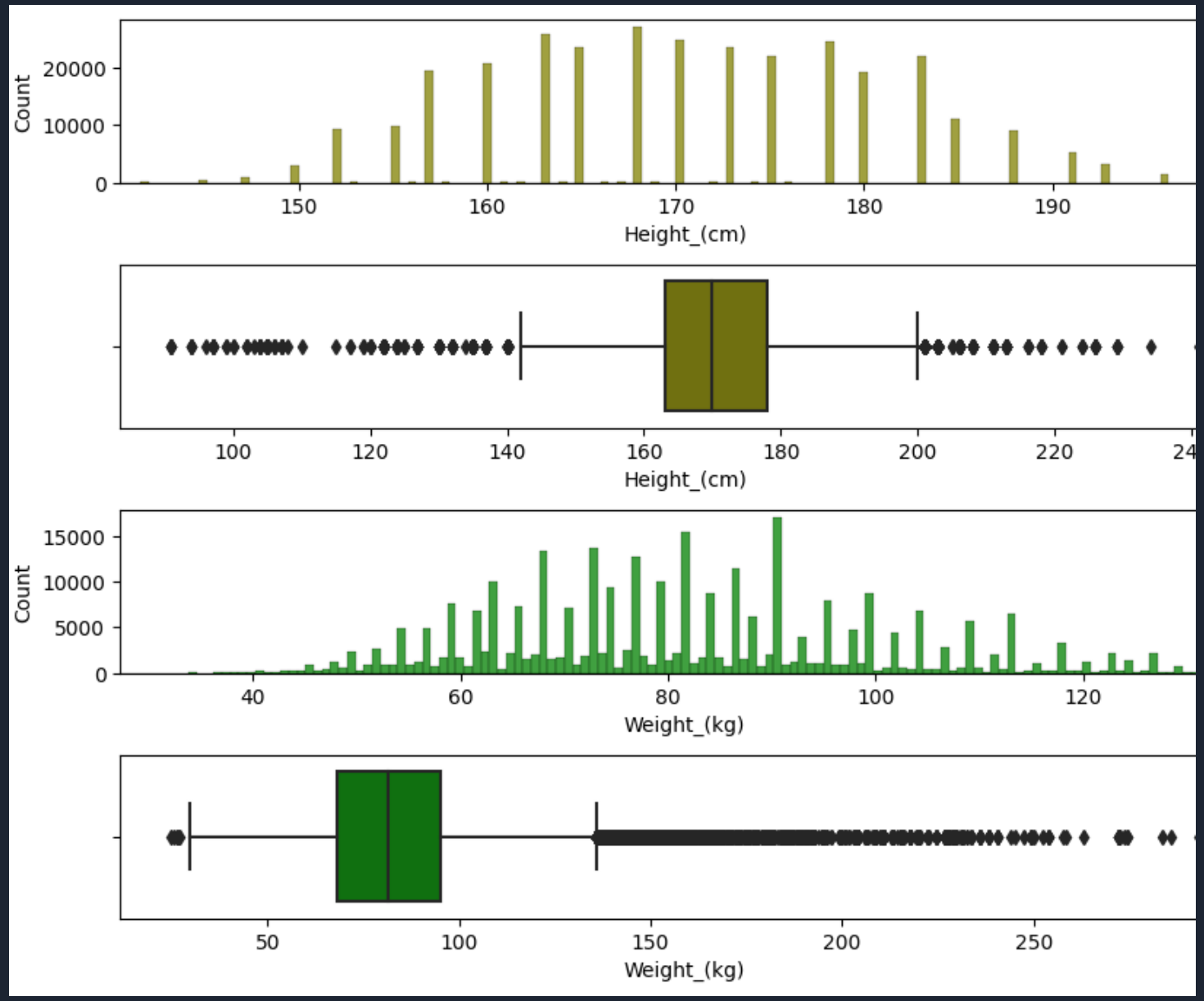
2) *Variables de diagnóstico*
Heart_Disease: Diagnosticada/o con problemas cardíacos - *Variable objetivo* -
Skin_Cancer: Diagnosticada/o con Cáncer de piel.
Other_Cancer: Diagnosticada/o con cualquier otro cáncer.
Depression: Diagnosticada/o con algún grado de depresión.
Diabetes: Diagnosticada/o con diabetes, contempla casos agudos y temporales como hiperglucemia por embarazo; pero también considera casos límites como no tener el diagnóstico, pero sí tener la etapa previa a diabetes.
Arthritis: Diagnosticada/o con artritis.

3) *Variables biológicas:*
Sex: Sexo del encuestado.
Age_Category: Grupo etario. Se dividen en intervalos de 5 años cada uno, empezando desde los 18 hasta 80. El primer intervalo se extiende hasta los 24, como así también de los 80 en adelante. El resultante es 18-24;25-29;...;75-79;80+
Height_(cm): Altura en Centímetros
Weight_(kg): Peso en Kilogramos
BMI: (Body Mass Index) - Índice de masa corporal. Coeficiente descriptivo de la distribución del peso entre los tejidos del cuerpo. Estimado a partir de la altura y del peso, pero también mensurable con herramientas médicas.

4) *Variables de hábitos*
Smoking_History: Es fumador activo o con al menos 100 cigarrillos consumidos
Alcohol_Consumption: Cantidad días que tomó alcohol de los últimos 30 días
Fruit_Consumption: Cantidad de comidas (desayuno, almuerzo, merienda, cena) que incluían fruta en los últimos 30 días
Green_Vegetables_Consumption: Cantidad comidas con Vegetales en los últimos 30 días
FriedPotato_Consumption: Cantidad de comidas con fritura en los últimos 30 días

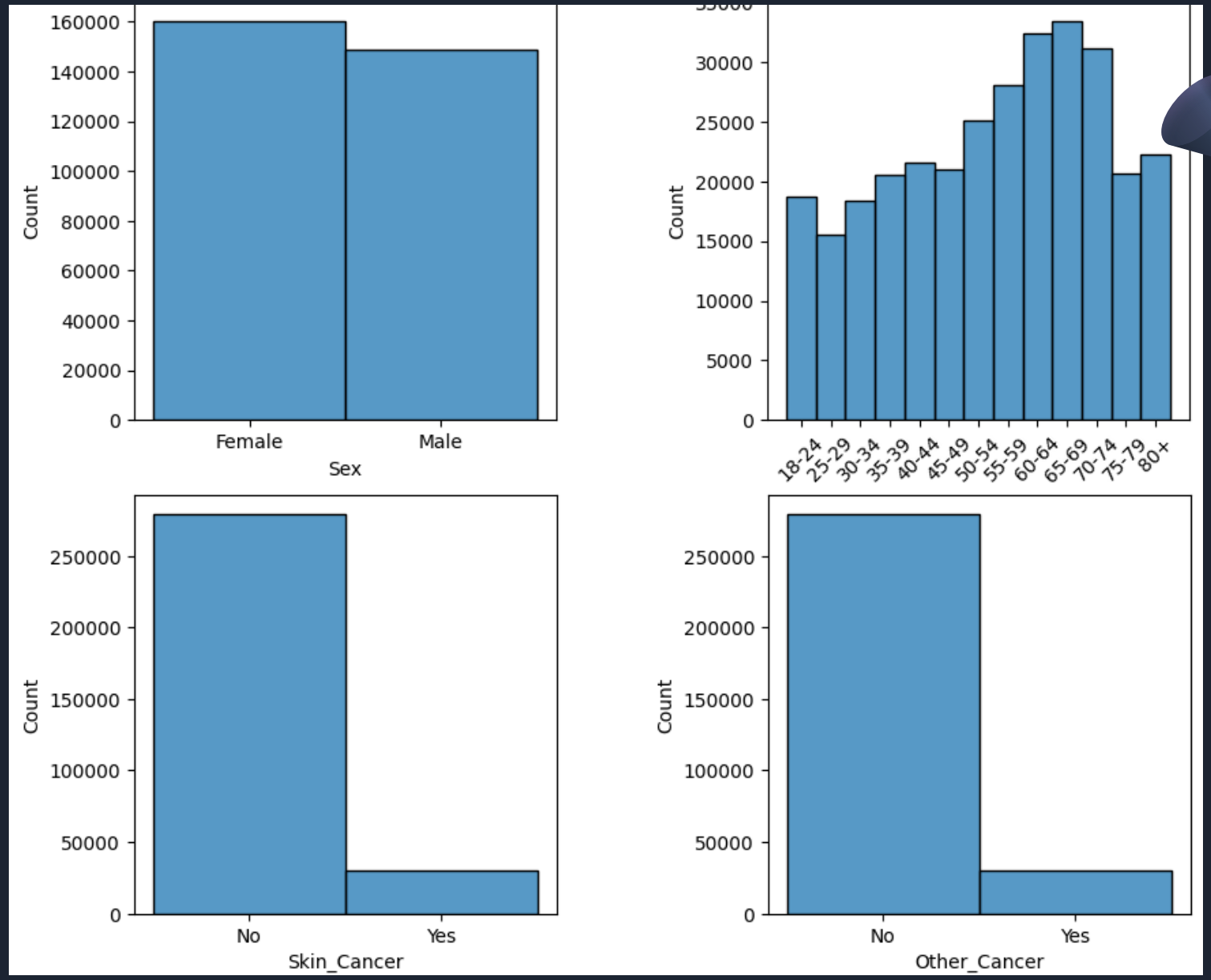
# Un poco de contexto...

- A partir de los datos obtenidos, se observa una distribución bastante centrada en el caso de altura (*Height*) y una distribución más asimétrica en peso (*Weight*)



# Un poco de contexto...

- Las poblaciones observadas de la encuesta son principalmente personas sanas, de una gran variedad de rango etario y aproximadamente iguales según género.
- Podemos dar por sentado que se trata de una muestra significativa en términos de aplicabilidad a una población normal para el diagnóstico a partir de variables secundarias



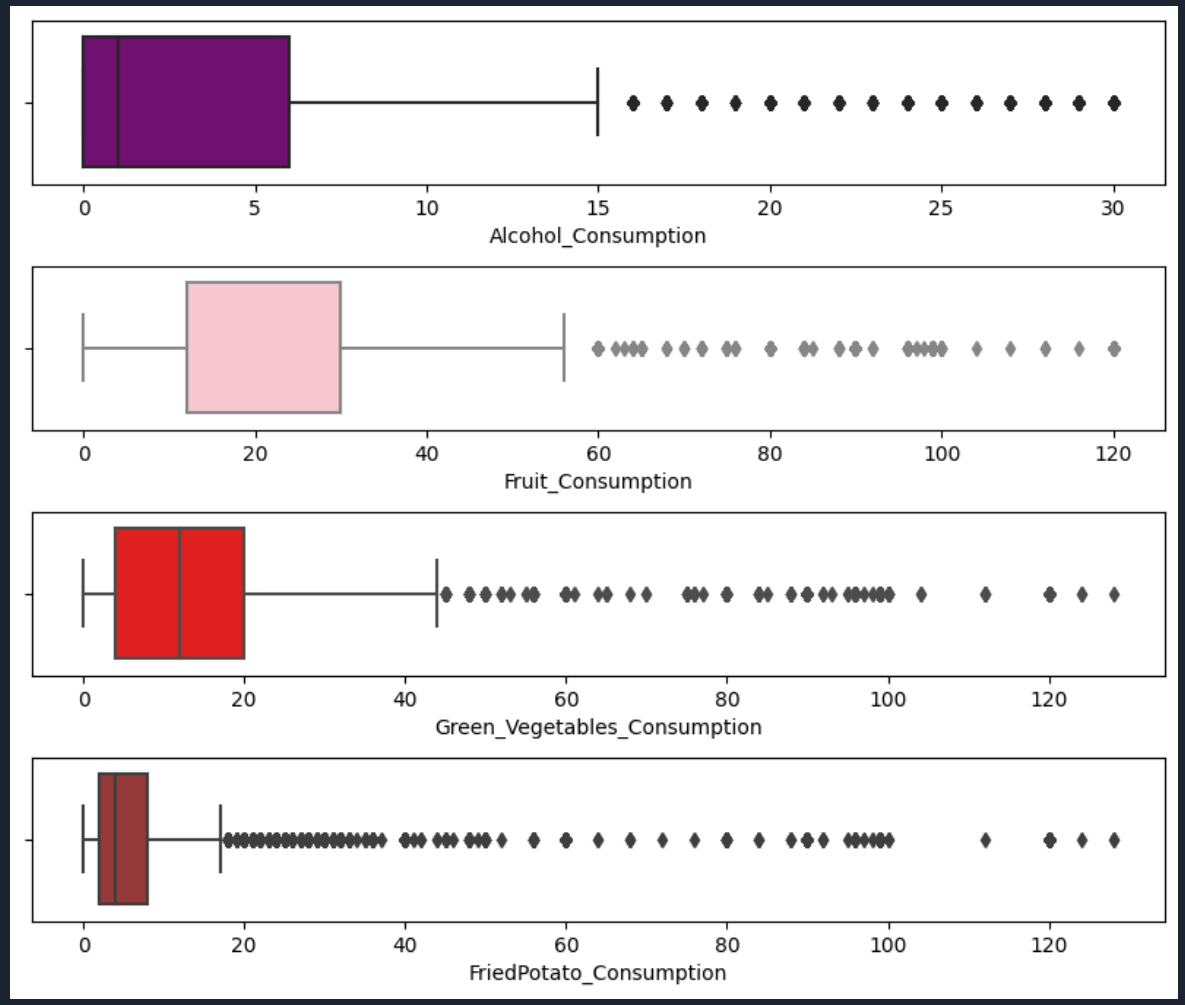
# La información importante

Tenemos ahora entonces los
Datos sobre los hábitos saludables y no tan saludables de la misma población.

Vale la pena destacar que el consumo de alcohol en exceso es una excepción, así como también las frituras. Aun así, se incluyen para el estudio.

En el caso de frutas y verduras, se reconoce una mayor participación en agregar estos alimentos a la dieta.

Sin embargo, esta información se vuelve poco relevante al comparar la inferencia de la edad y el consumo de tabaco.



# El principio del modelo

Luego de calcular la correlación que existe entre el diagnóstico de aquellos que padecen enfermedades cardíacas, con los otros aspectos que forman parte de la encuesta, se llega al siguiente caso: Ninguna de las variables por sí solas puede predecir una enfermedad cardíaca con precisión.

Sin embargo, es posible que la combinación de variables, sobre todo aquellas que muestran una mayor correlación lineal, puedan dar una predicción más certera.

Con esta información se procede a modelar mediante machine learning un algoritmo de clasificación para predecir posteriormente nuevos casos de enfermedades cardíacas.

| Correlación lineal con enfermedades cardíacas: | |
|---|---|
| Skin_Cancer | 0.090848 |
| Other_Cancer | 0.092387 |
| Depression | 0.032526 |
| Diabetes | 0.177285 |
| Arthritis | 0.153913 |
| Height_(cm) | 0.015780 |
| Weight_(kg) | 0.045875 |
| BMI | 0.042666 |
| Age_Category | 0.229011 |
| Smoking_History | 0.107797 |
| Alcohol_Consumption | -0.036569 |
| FriedPotato_Consumption | -0.009227 |

# Construcción del modelo

Para resolver el problema de clasificación y predicción se utilizó **Gradient Boosting Clasifier**

*Algunas de las ventajas de Gradient Boosting son:*

- Tiene un alto rendimiento y precisión en problemas de regresión y clasificación.
- Puede utilizar métodos de regularización para evitar el sobreajuste.

*Algunas de las desventajas son:*

- Es más lento y consume más memoria que otros algoritmos de ensamble como Random Forest3.

# Los resultados

- Una vez normalizadas y codificadas las variables, se implementó el modelo de GBC sobre las variables previamente mencionadas.
- Los resultados del modelo son bastante prometedores, sobre todo si se considera que no se realizó ningún ajuste de hiperparámetros fino al respecto
- Los resultados reflejan que hubo 79 casos de falsos positivos, lo cual es posible que con la asesoría de un médico y estudios pertinentes se despeje la posibilidad de un problema cardíaco, sin embargo 6156 casos pasaron desapercibidos todavía. Lo que refleja que el modelo necesita más información sobre los individuos para poder predecir con mayor exactitud

```
#Ahora comparemos con GBM

gbrt = GradientBoostingClassifier(random_state = 42)
gbrt.fit(X_train, y_train)
print("Accuracy de Gradient Boosting:\n",gbrt.score(X_test, y_test))

y_pred_gbc = gbrt.predict(X_test)
cm = confusion_matrix(y_test, y_pred_gbc)
print("Matriz de confusión:\n",cm)
```

```
Accuracy de Gradient Boosting:
 0.919250395006087
Matriz de confusión:
 [[70906    79]
 [ 6156    73]]
```

# Conclusiones

- Es posible a partir de múltiples variables distintas de los hábitos cotidianos, peso y edad de un individuo, a obtener una predicción del riesgo que corre a tener enfermedades cardíacas. El modelo requiere más información tal vez para ser científicamente válido (precisión mayor al 95%) o incluso más para ser comercialmente viable.
- Los distintos modelos de clasificación (DecisionTreeClassifier, GaussianNB, GradientBoostingClassifier) pueden aplicarse para la predicción de riesgo de enfermedades, en distinta medida cada uno, según el costo computacional, pueden conseguir mayor precisión y volverse una herramienta útil para el cuidado de la salud del ser humano.

# Agradecimientos


- A todo el equipo de CoderHouse, que hizo posible el desarrollo del curso de Data Scientist.
- A los profesores a cargo de la cursada, Orlando Ramos Pérez y Nestor Jose Escudero Mora. Que con empeño llevaron a cabo la tarea de explicar los conceptos a pesar de las dificultades encontradas.
- A todos los compañeros de cursada que brindaron apoyo y ayuda con sus consultas y buena voluntad.

# Fuentes

- Dataset: https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset
- Datos del censo: https://www.cdc.gov/brfss/annual_data/annual_2021.html
- Preguntas de la encuesta: https://www.cdc.gov/brfss/questionnaires/pdf-ques/2021-BRFSS-Questionnaire-1-19-2022-508.pdf