

CLASIFICADOR DE RIESGO CARDÍACO

PROFESOR

Sebastian Ferraro

TUTOR

Juan Manuel Romero

GRUPO

Ignacio Silva - Daniela Flores - Dimas Torres
Camila Baron - Erica Müller

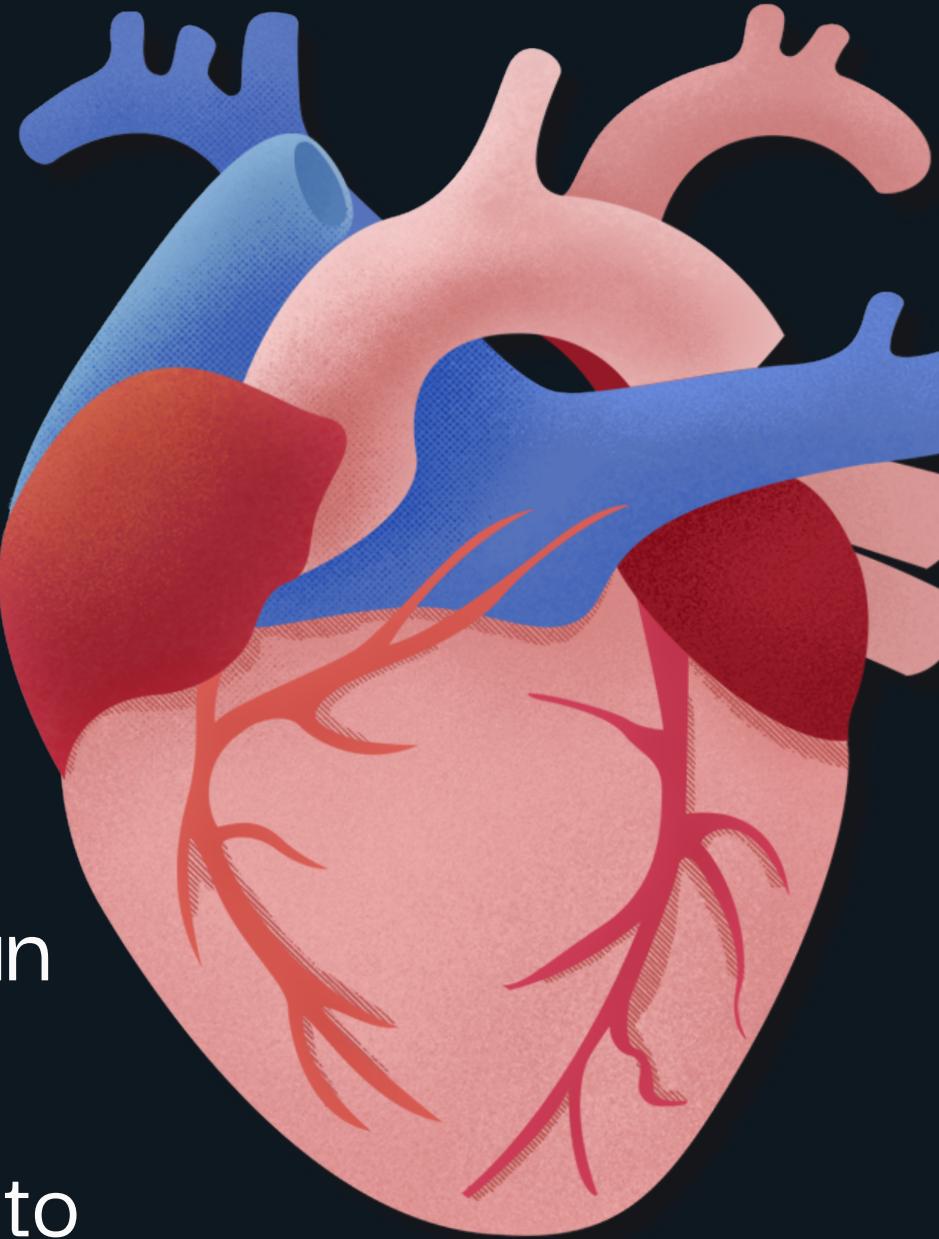
OBJETIVO

El objetivo de este análisis es identificar el mejor modelo de clasificación que nos permita detectar a través de las variables presentadas las posibilidades de sufrir una enfermedad cardiaca.

DATASET

Origen: Sistema de Vigilancia de Factores de Riesgo del Comportamiento (BRFSS) EE.UU.

Motivo de su elección: Creemos que es un tema de gran importancia. Las tasas de enfermedades cardiovasculares continúan aumentando debido a distintos factores y nos resulta importante e interesante estudiar la relevancia de todos los parámetros que puedan incidir en el aumento de probabilidades de desarrollar una de estas enfermedades.



Variables CATEGORICAS

HeartDisease, Smoking, AlcoholDrinking, Stroke, DiffWalking, Sex, Race,
Diabetic, PhysicalActivity, GenHealth, Asthma, KidneyDisease, SkinCancer

Variables CONTINUAS

BMI, PhysicalHealtH, MentalHealth, AgeCategory, SleepTime

EDA

| Se observaron las variables que existen y la calidad de los datos. Confirmamos los tipos de variables y que no tenemos campos nulos. Determinamos la Variable Target: Heart Disease. El trabajo se inició con un data set de 319795 registros y tuvimos que reducirlo a 4000 para poder trabajar.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 4000 entries, 13706 to 118707
Data columns (total 18 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   HeartDisease    4000 non-null   object  
 1   BMI              4000 non-null   float64 
 2   Smoking          4000 non-null   object  
 3   AlcoholDrinking 4000 non-null   object  
 4   Stroke           4000 non-null   object  
 5   PhysicalHealth   4000 non-null   int64  
 6   MentalHealth     4000 non-null   int64  
 7   DiffWalking      4000 non-null   object  
 8   Sex               4000 non-null   object  
 9   AgeCategory      4000 non-null   object  
 10  Race              4000 non-null   object  
 11  Diabetic         4000 non-null   object  
 12  PhysicalActivity 4000 non-null   object  
 13  GenHealth        4000 non-null   object  
 14  SleepTime        4000 non-null   int64  
 15  Asthma            4000 non-null   object  
 16  KidneyDisease    4000 non-null   object  
 17  SkinCancer        4000 non-null   object  
dtypes: float64(1), int64(3), object(14)
memory usage: 593.8+ KB
```

Análisis Univariado

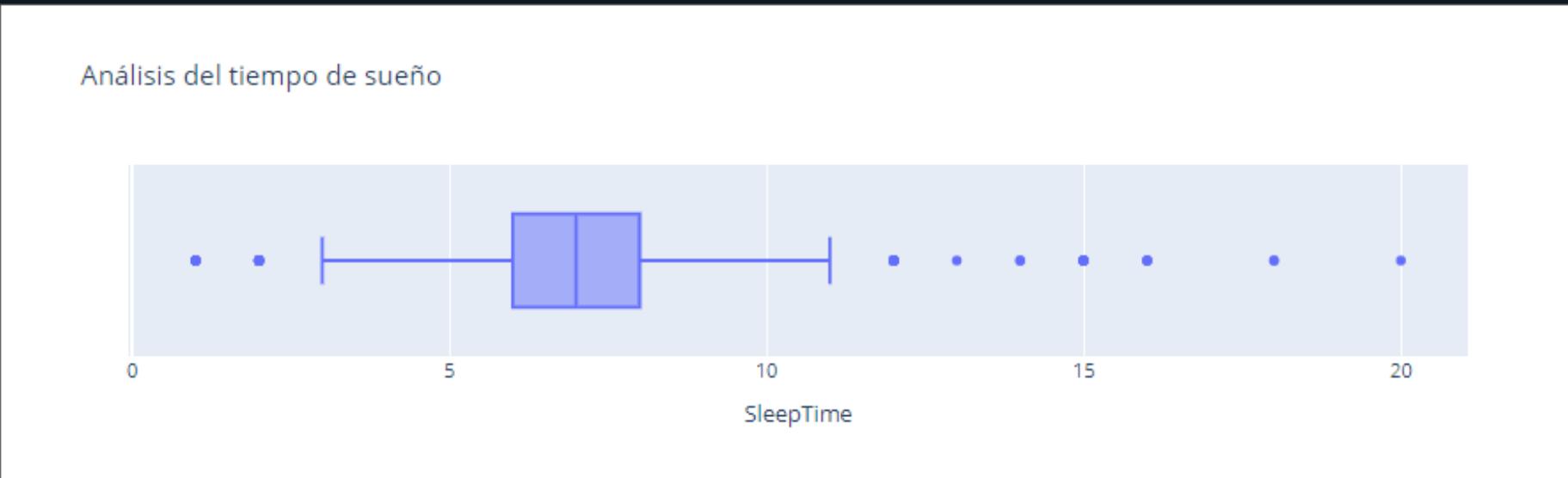
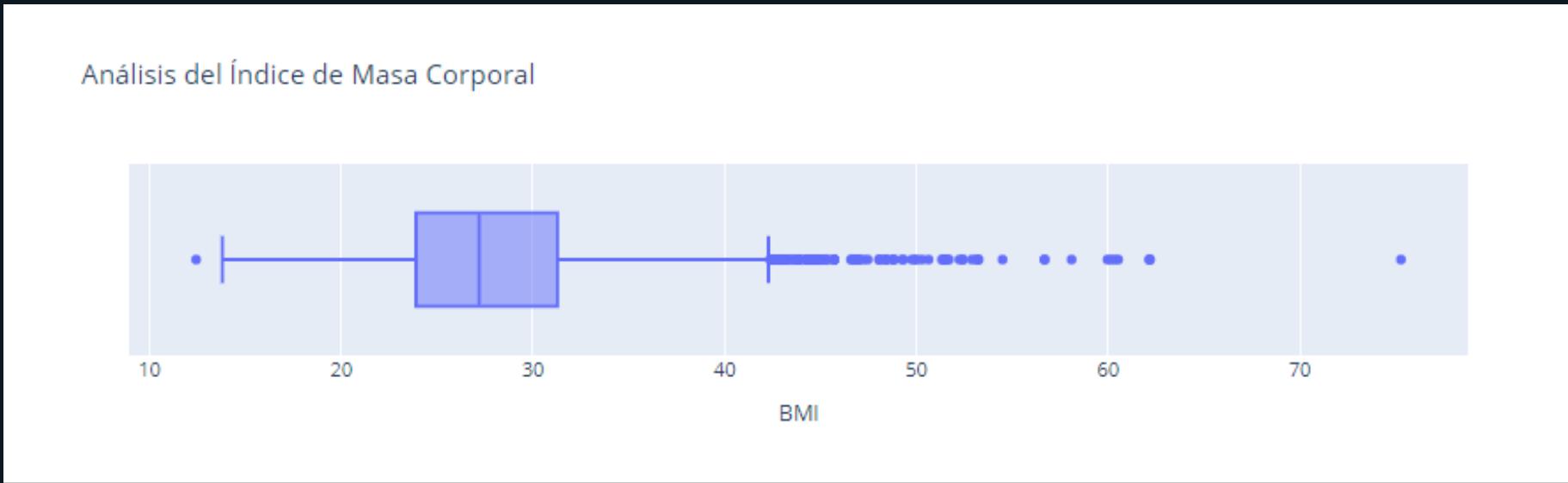
Variables CONTINUAS

	BMI	PhysicalHealth	MentalHealth	SleepTime
count	4000.00	4000.00	4000.00	4000.00
mean	28.14	3.36	3.96	7.07
std	6.20	7.96	7.97	1.43
min	12.44	0.00	0.00	1.00
25%	23.91	0.00	0.00	6.00
50%	27.20	0.00	0.00	7.00
75%	31.28	2.00	3.00	8.00
max	75.28	30.00	30.00	20.00

| La tabla nos muestra la distribución de las variables continuas. Podemos ver que existen valores extremos por lo que iniciamos un análisis mas profundo.

Análisis Univariado

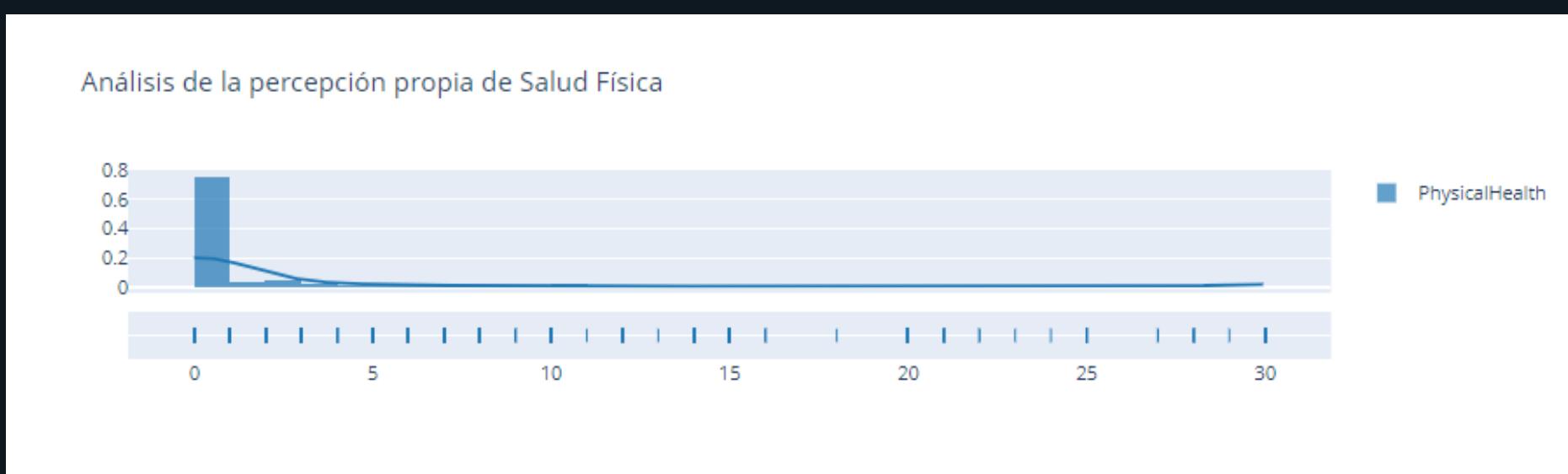
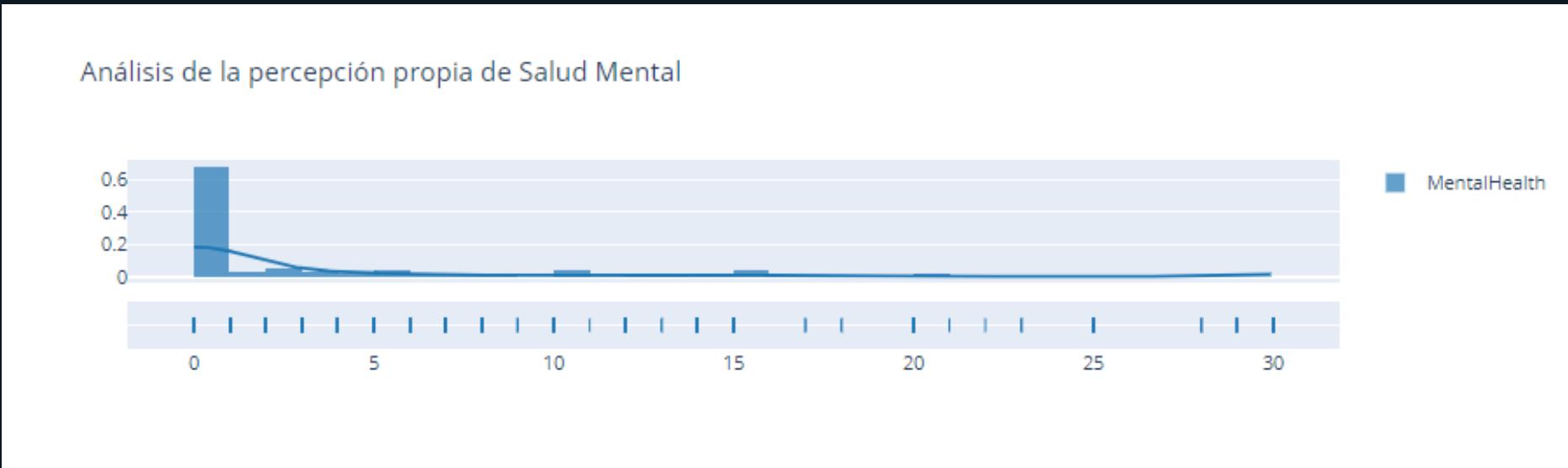
Variables CONTINUAS



| En ambos graficos confirmamos que existen valores extremos, pero que por fuera de estos, existe una variable bastante normal.

Análisis Univariado

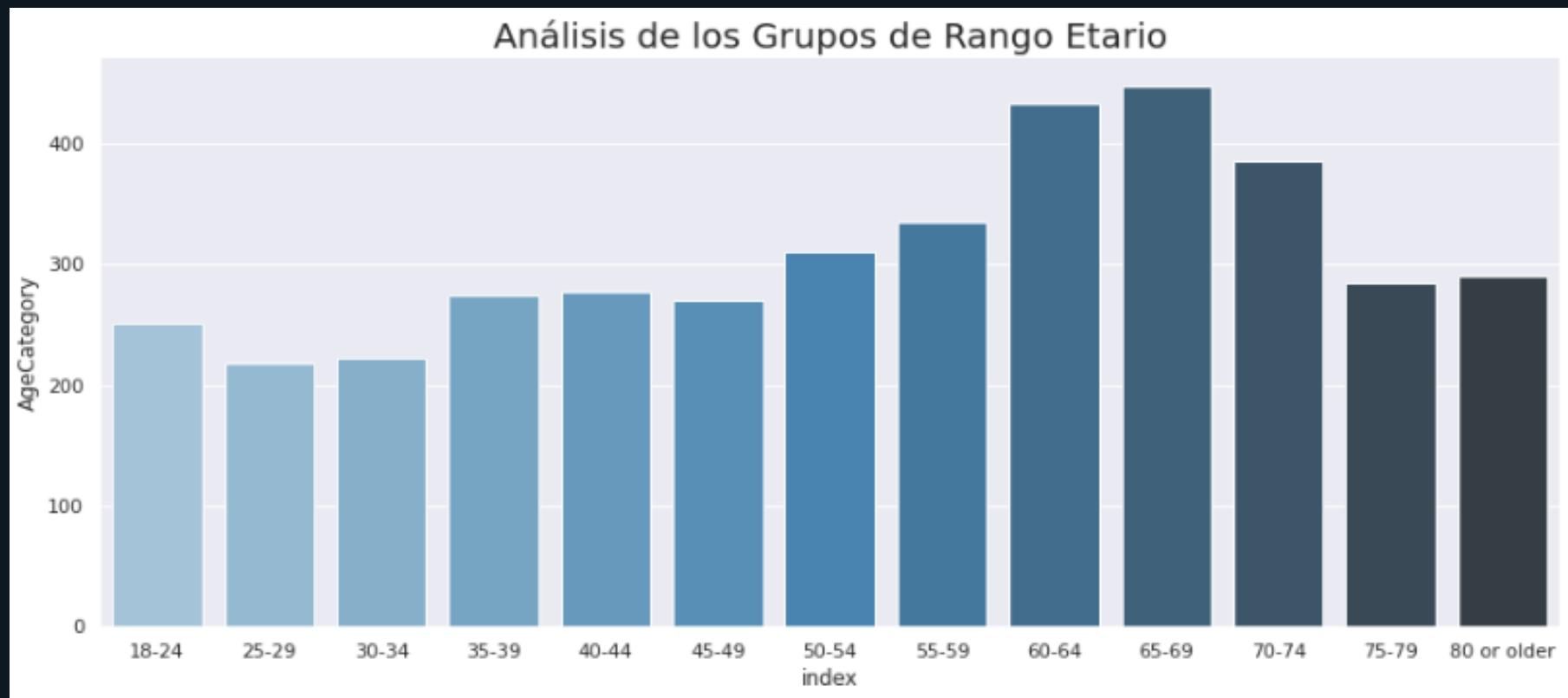
Variables CONTINUAS



En este caso vemos que existe una relación entre Salud Mental y Física, al menos en la percepción de las personas que brindaron sus datos.

Análisis Univariado

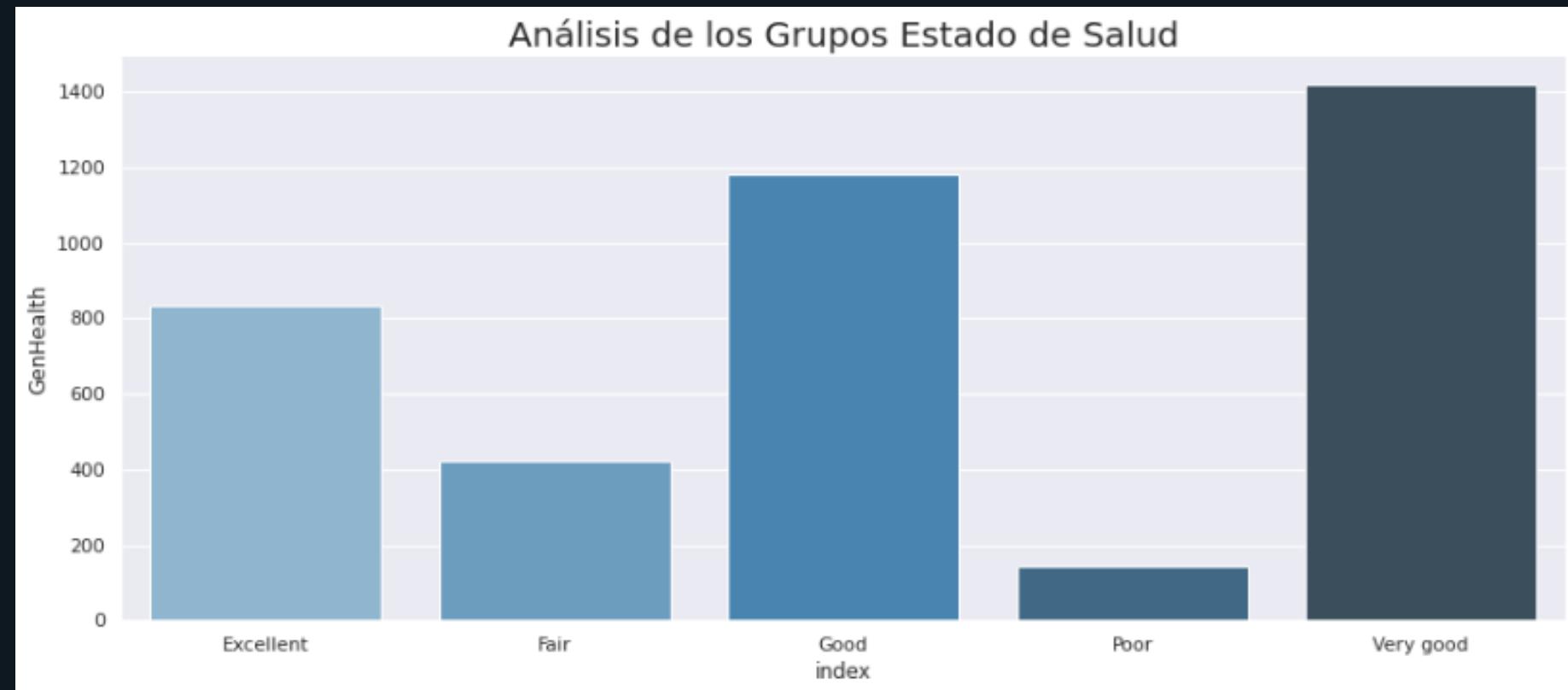
Variables CONTINUAS



| La distribución de edades es bastante pareja aunque se ve una preferencia por informantes cercanos a los 50.

Análisis Univariado

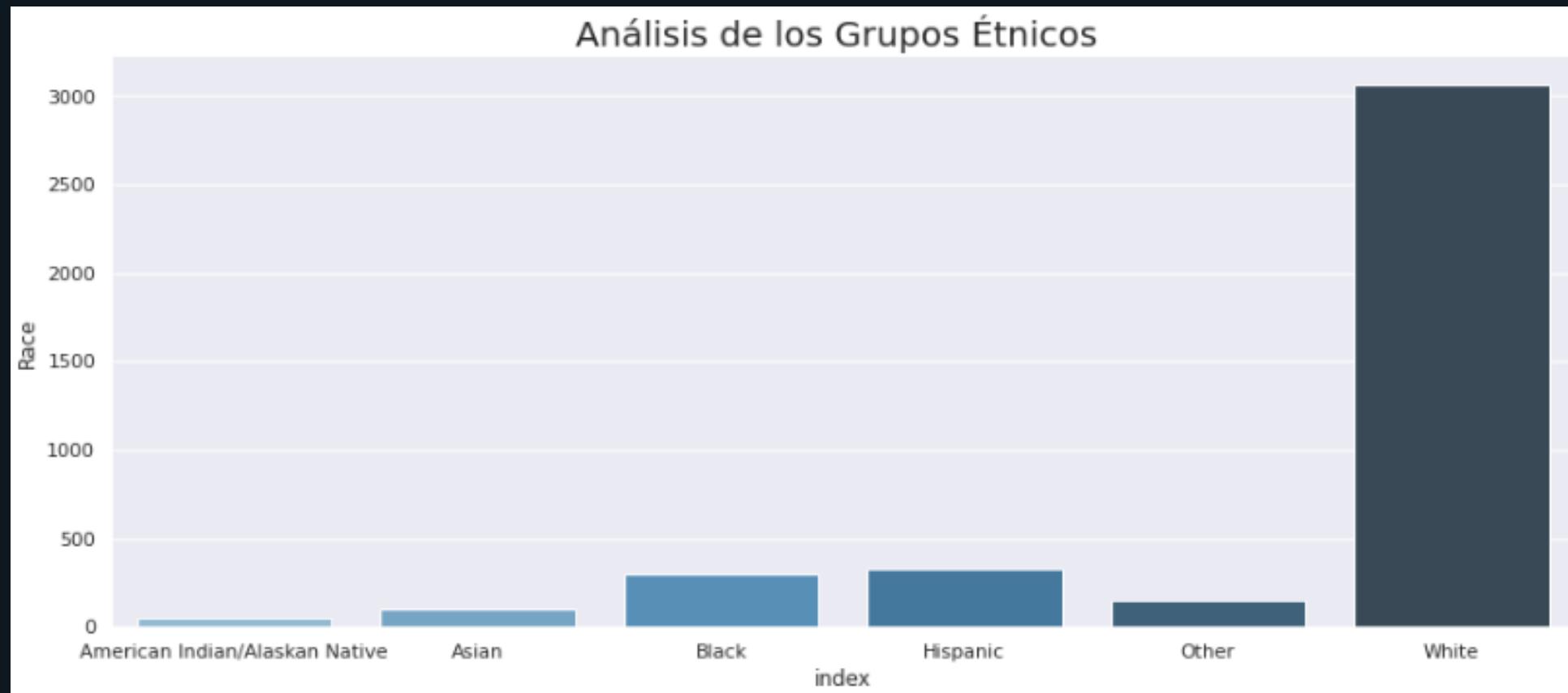
Variables CUALITATIVAS



Vemos una notable diferencia entre las personas que creen tener un estado de salud deficiente frente a las demás opciones.

Análisis Univariado

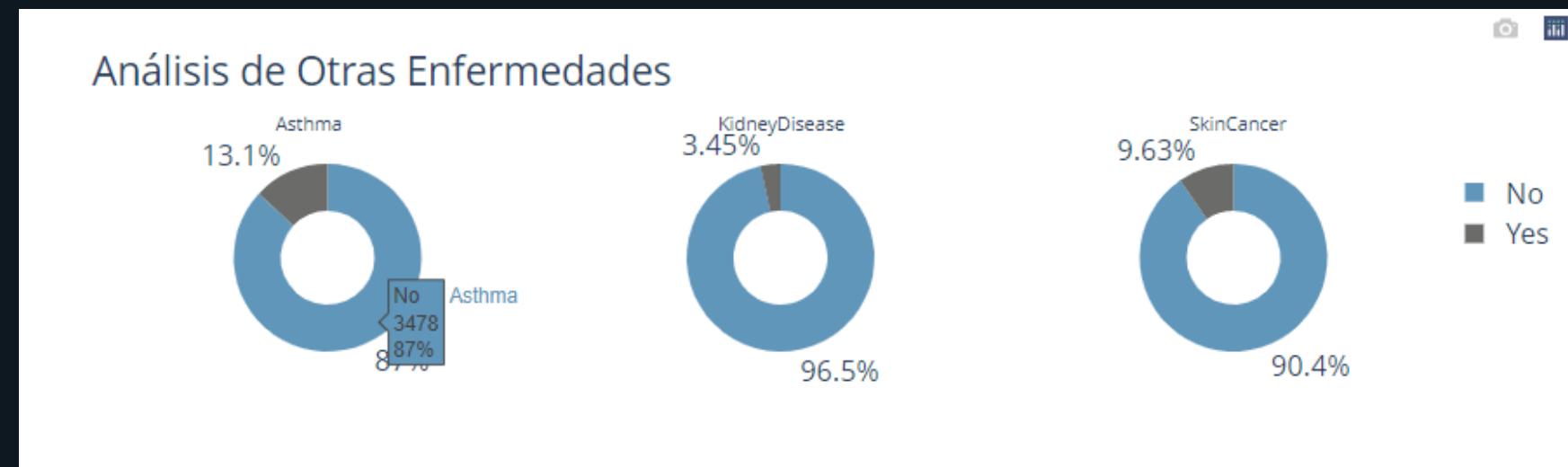
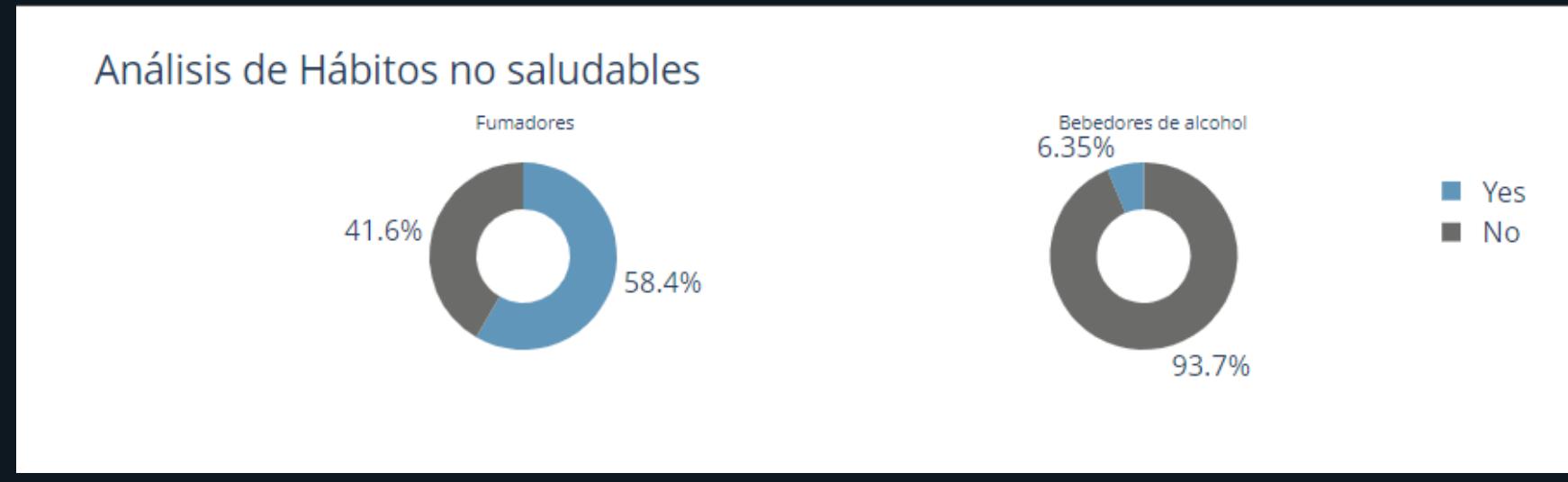
Variables CUALITATIVAS



| Existe un gran predominio de personas de Etnia Blanca lo que es un dato entendible teniendo en cuenta el origen del dataset.

Análisis Univariado

Variables CUALITATIVAS



| Los gráficos de Hábitos no saludables nos muestran que tenemos un predominio de fumadores sobre bebedores. Por los de Otras Enfermedades, hay un predominio marcado de personas que no tienen antecedentes.

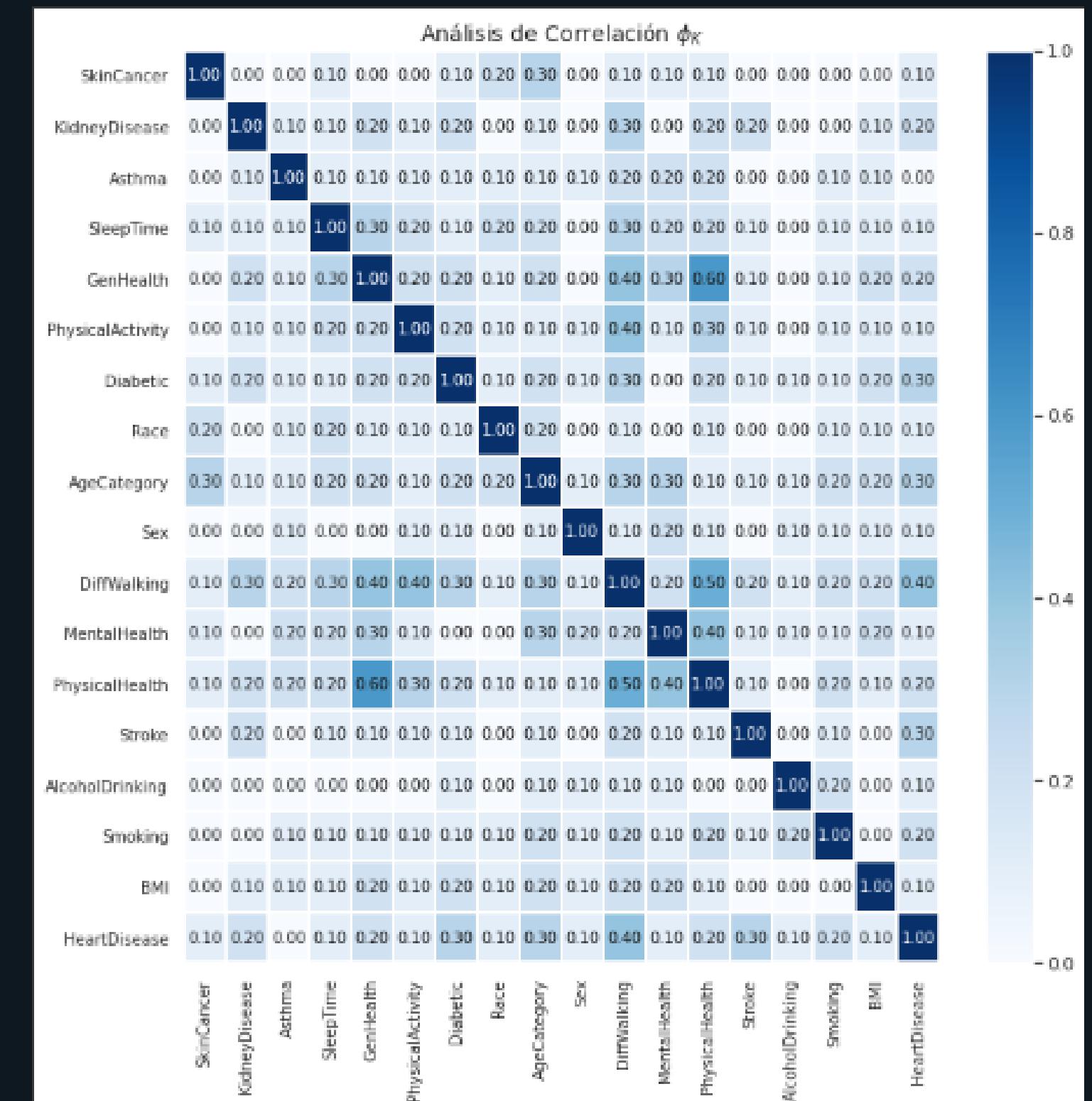
Análisis Bivariado

χ^2 de Pearson (chi-cuadrado)

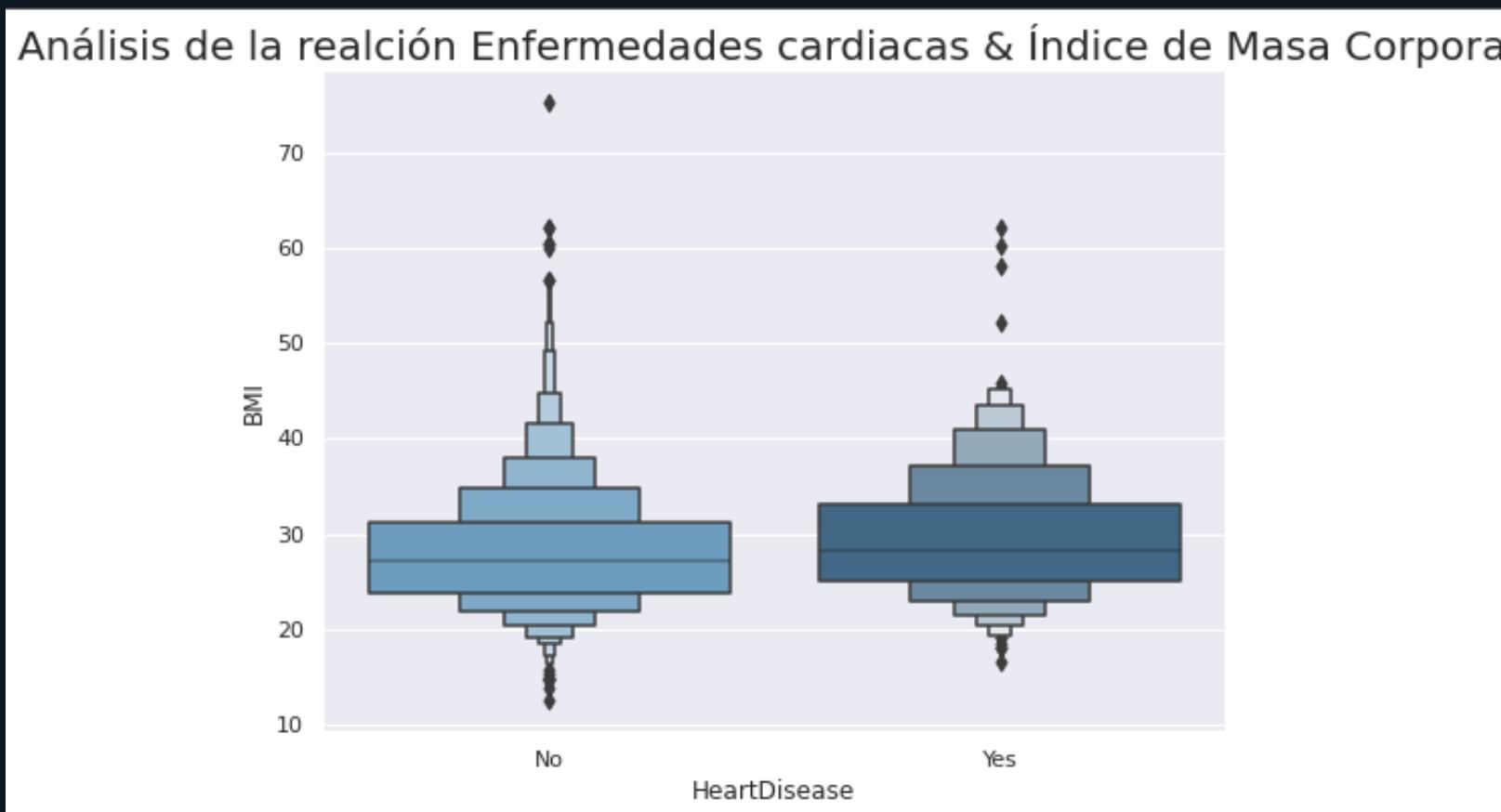
El gráfico de correlación nos muestra una relación entre salud física y salud general, un poco menor entre salud física y salud mental.

Por otro lado, vemos vinculada la salud física a la tendencia a caminar.

No observamos ninguna relación importante entre las variables y la variable target.

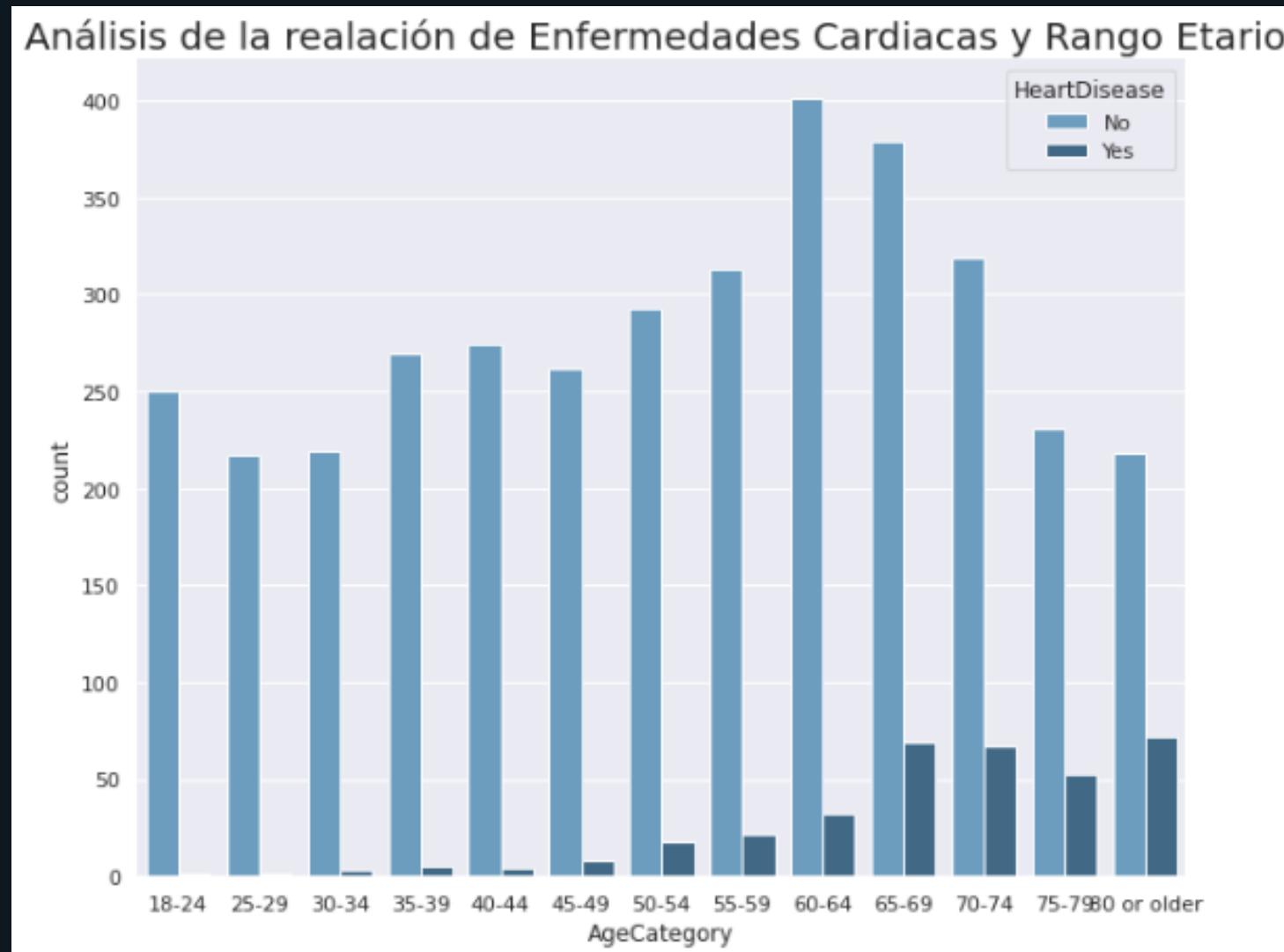


Análisis Bivariado



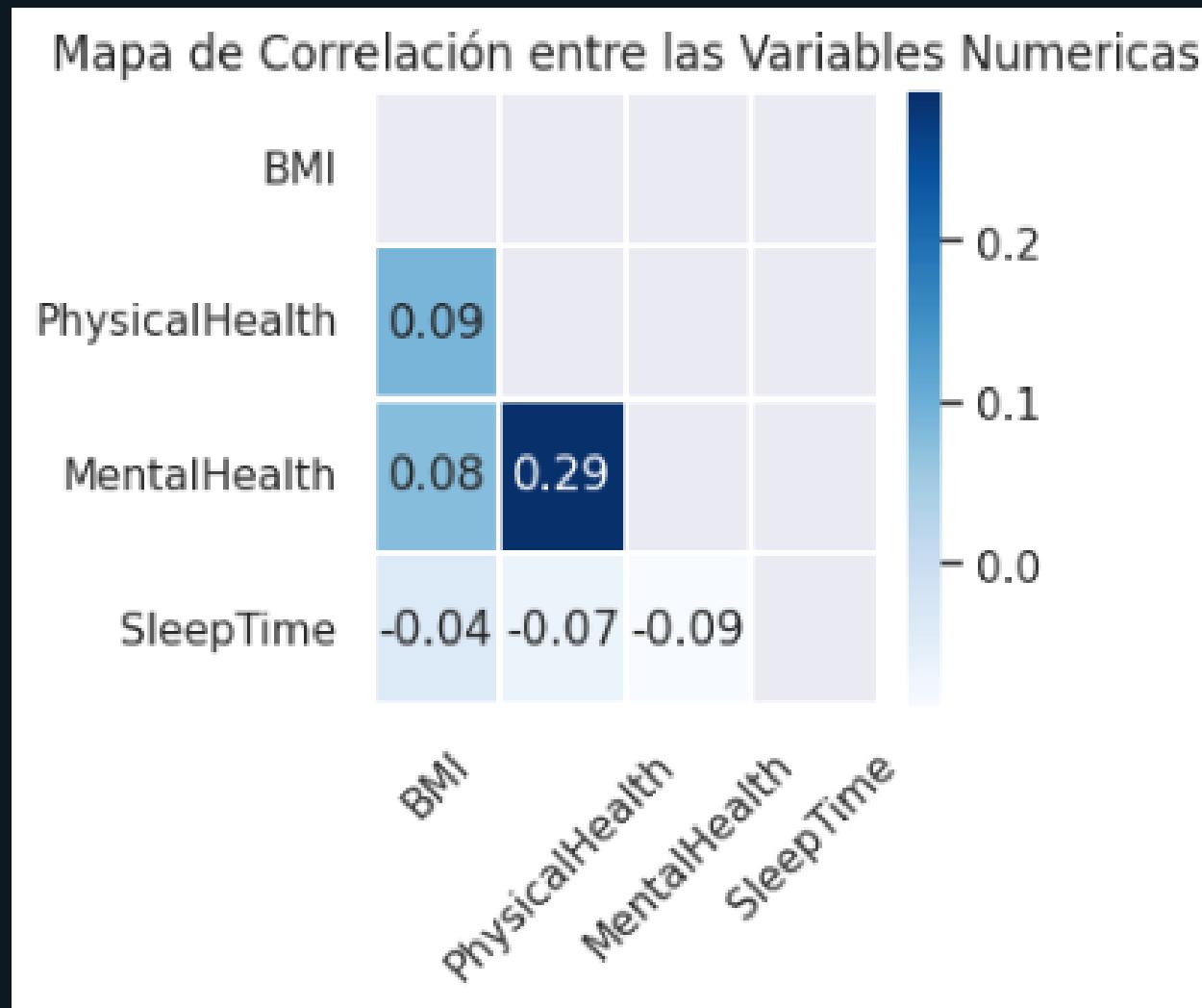
| La distribucion de indice de masa corporal no parece diferir entre las personas entre el grupo con enfermedades cardiacas y el que no tiene.

Análisis Bivariado



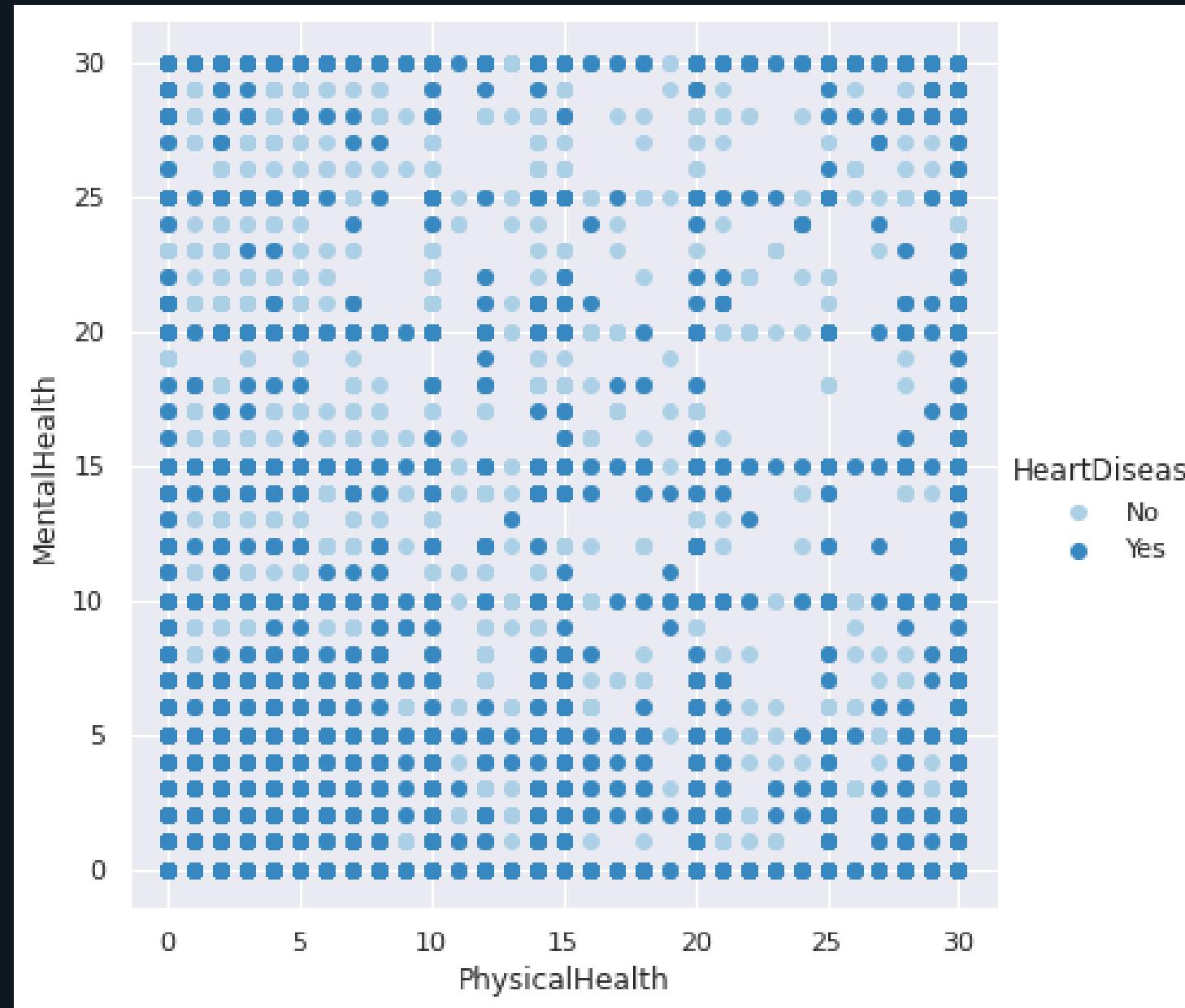
| Las categorías de edad mas avanzada tiene mayor volumen de personas con enfermedades cardiacas.

Análisis Multivariado



Solo vemos una leve relación entre salud mental y salud física.

Análisis Multivariado

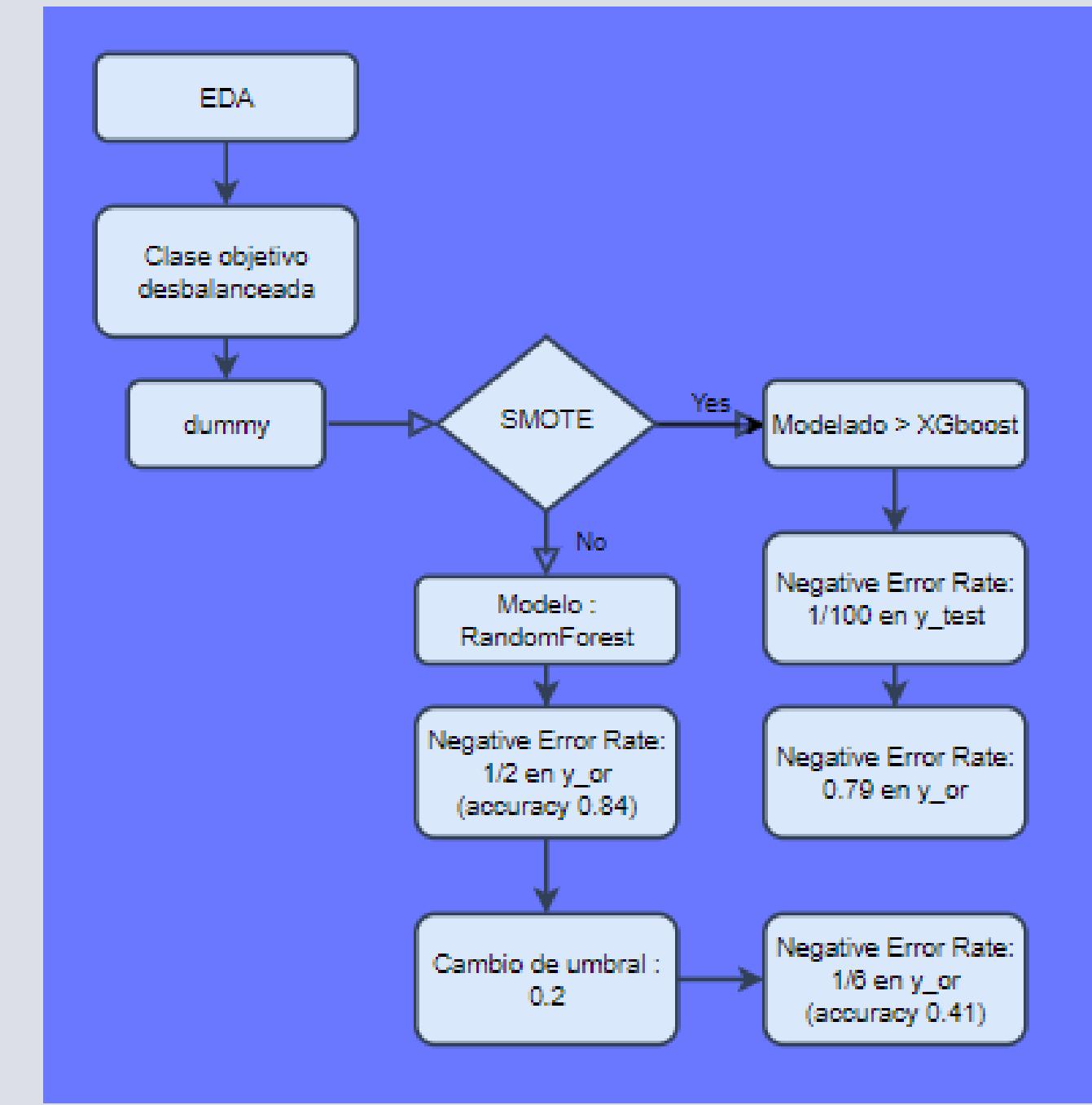


| La imagen nos muestra que las personas con menor puntaje en estado de salud es mas frecuente a tener enfermedades cardiacas.

Modelos

Nuestro dataset original tiene un desbalanceo de clases importante, por lo que se aplicó la Técnica SMOTE para resolverlo.

Al procesar el nuevo dataset balanceado percibimos que el modelo podría estar sobre entrenado y decidimos volver a trabajarla sin hacer oversample.



Resultados de los Modelos

Regresión Logistica

fit_time	0.466319
score_time	0.021385
test_f1	0.192174
test_accuracy	0.895000
test_roc_auc	0.824501
test_recall	0.129167
test_precision	0.520238

Bosque Aleatorio

fit_time	0.155558
score_time	0.033010
test_f1	0.000000
test_accuracy	0.902500
test_roc_auc	0.822515
test_recall	0.000000
test_precision	0.000000

XGBoost

fit_time	0.167880
score_time	0.008938
test_f1	0.000000
test_accuracy	0.901250
test_roc_auc	0.789194
test_recall	0.000000
test_precision	0.000000

Adaboost

fit_time	0.946045
score_time	0.192434
test_f1	0.047059
test_accuracy	0.898750
test_roc_auc	0.803623
test_recall	0.026667
test_precision	0.200000

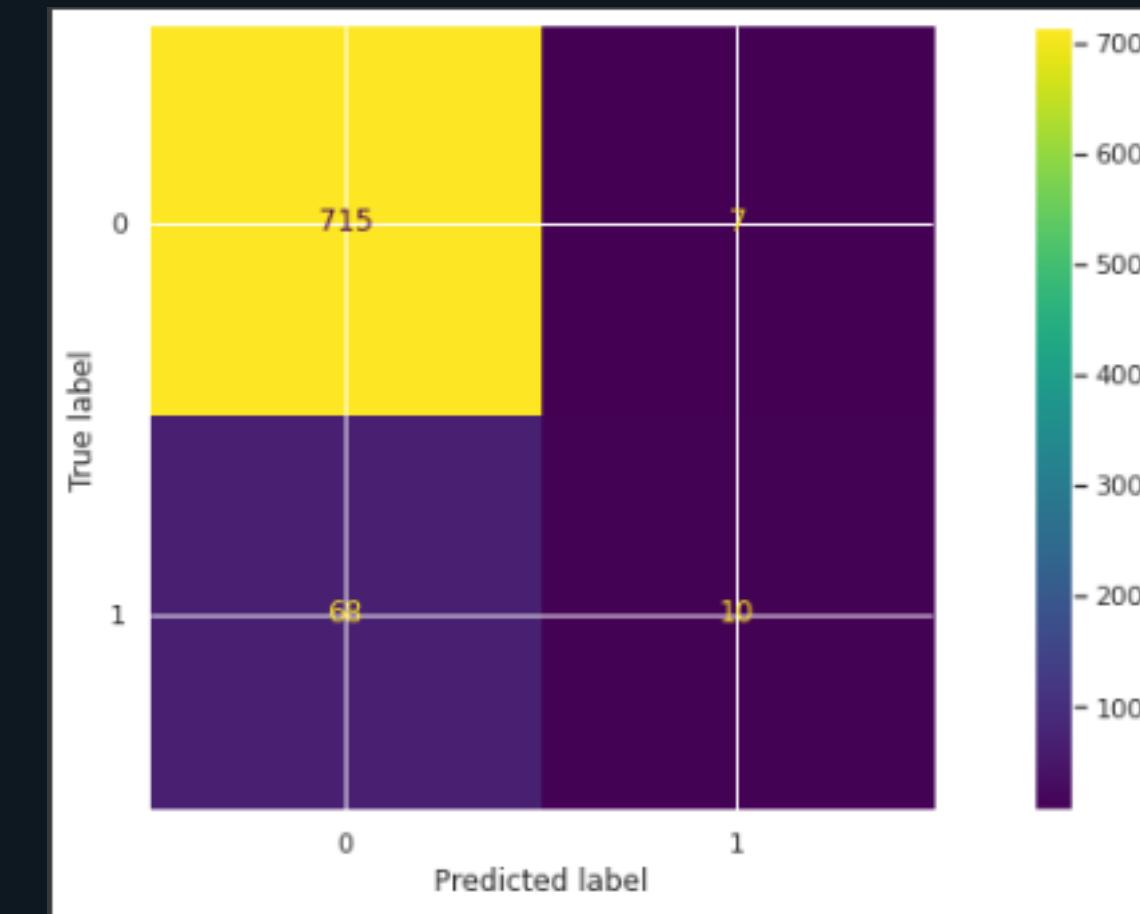
LGBMC

fit_time	0.208055
score_time	0.022358
test_f1	0.262303
test_accuracy	0.888750
test_roc_auc	0.755098
test_recall	0.205000
test_precision	0.387302

Mejor Modelo

Regresión Logistica

fit_time	0.466319
score_time	0.021385
test_f1	0.192174
test_accuracy	0.895000
test_roc_auc	0.824501
test_recall	0.129167
test_precision	0.520238

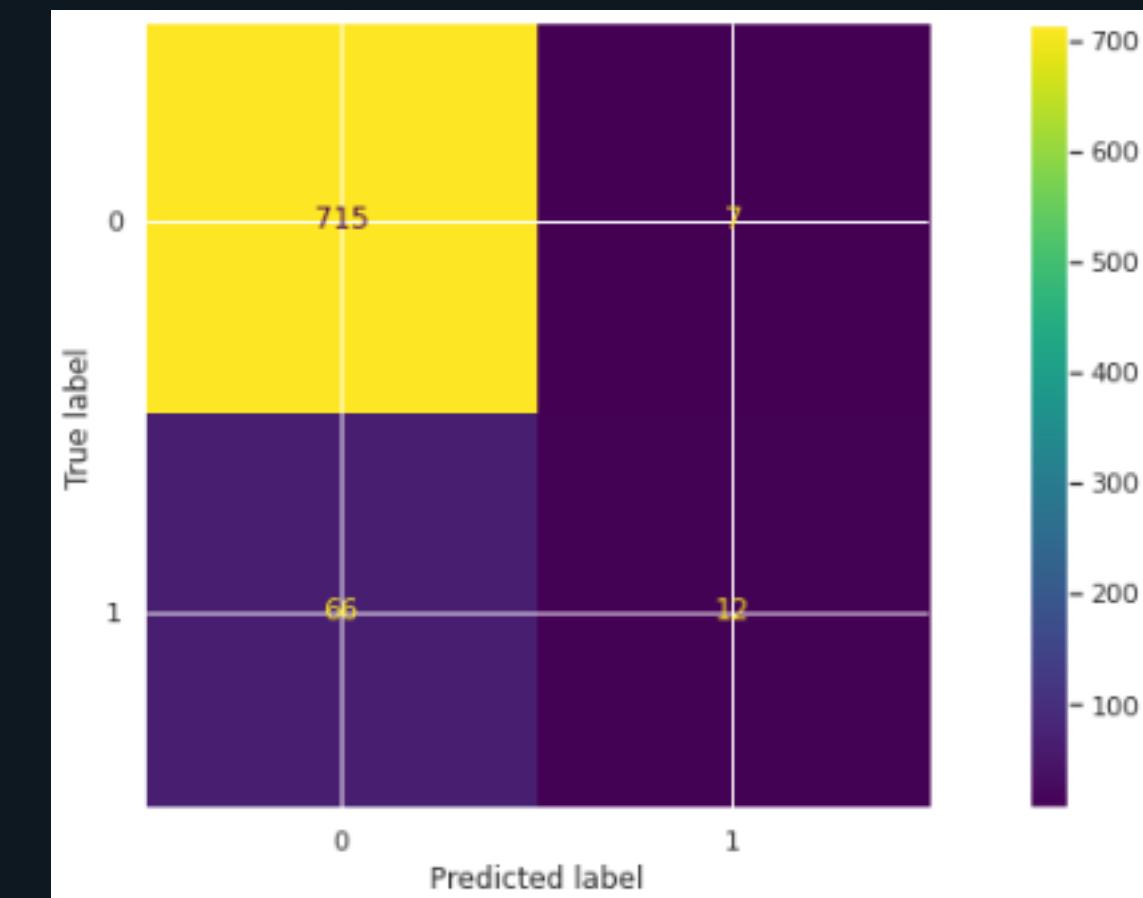


Mejor Modelo

Optimización del Modelo
Con SMOTE

fit_time	0.246966
score_time	0.010910
test_f1	0.285902
test_accuracy	0.898750
test_roc_auc	0.822495
test_recall	0.204167
test_precision	0.523810

miss rate :0.8461538461538461
selectivity :0.009695290858725761

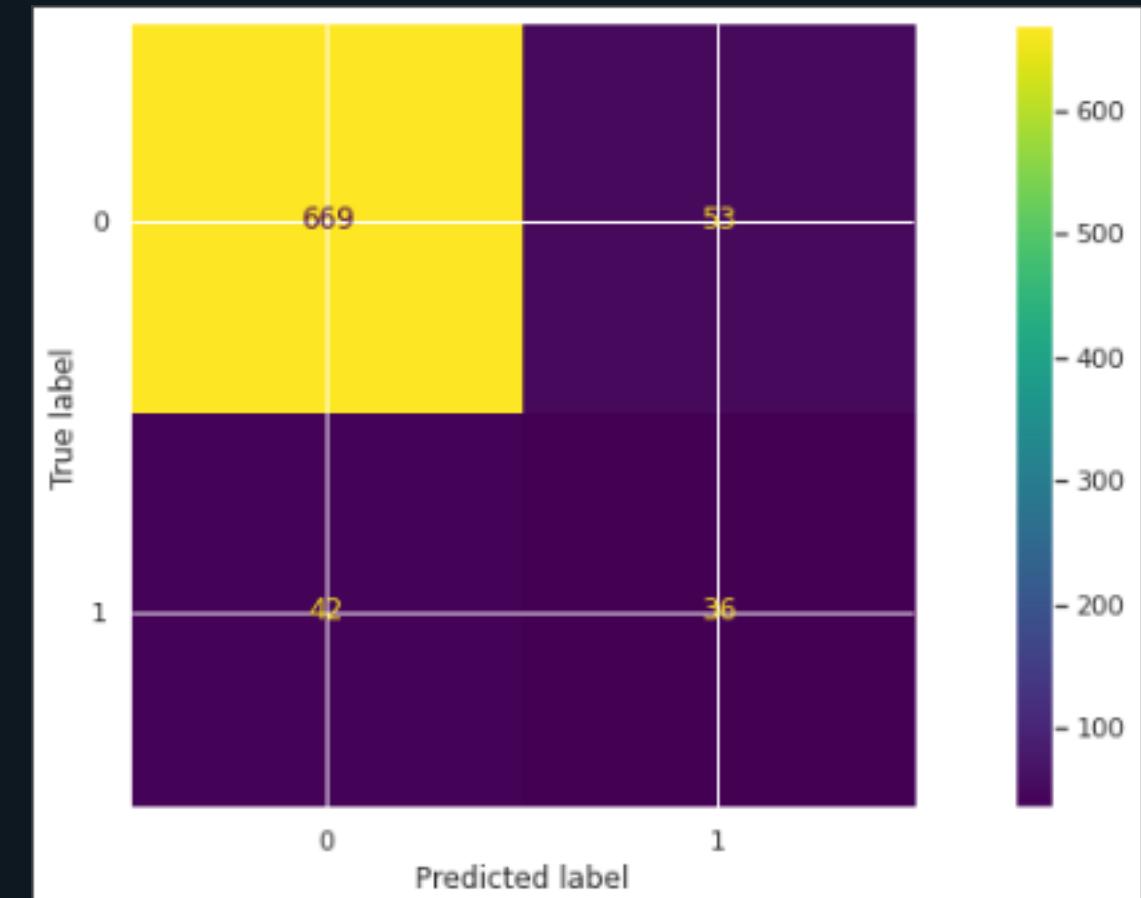


Mejor Modelo

Optimización del Modelo Sin SMOTE

```
resultados de la muestra desbalanceada /nfit_time          0.560881
score_time        0.021015
test_f1           0.428044
test_accuracy     0.863750
test_roc_auc      0.818782
test_recall       0.528333
test_precision    0.366732
dtype: float64
resultados de la muestra balanceada /nfit_time          0.108854
score_time        0.010916
test_f1           0.177747
test_accuracy     0.895000
test_roc_auc      0.820804
test_recall       0.116667
test_precision    0.507143
dtype: float64
```

```
miss rate :0.8461538461538461
selectivity :0.009695290858725761
```



Mejor Modelo

Optimización del Modelo Sin SMOTE

```
miss rate :0.11538461538461539  
selectivity :0.3490304709141274
```

Cambiando el umbral de nuestro modelo podemos alcanzar una menor tasa de error a cambio de sacrificar parte de nuestra precisión. Sólo 1 de cada 8 personas de alto riesgo serán clasificadas equivocadamente, un error tres veces menor que en el caso de pacientes de bajo riesgo.