**Laboratorio 1 - Clasificación**

[**1.** **Modelos Escogidos:** 1](#_Toc113216984)

[**2.** **Explicación de Tratamiento de Datos:** 1](#_Toc113216985)

[**3.** **Explicación de la tabla:** 3](#_Toc113216986)

[**Dashboards en Tableau:** 4](#_Toc113216987)

[**Datos Sin Depurar:** 4](#_Toc113216988)

[**Dashboard Unificado:** 5](#_Toc113216989)

[**Datos Depurados:** 6](#_Toc113216990)

[**Dashboard Unificado:** 7](#_Toc113216991)

[**4.** **Modelos de clasificación:** 9](#_Toc113216992)

[**KNN:** Hecho por Martín Ubaque 9](#_Toc113216993)

[**Árboles de decisión:** Hecho por Diego Alejandro Tovar 10](#_Toc113216994)

[**Regression:** 11](#_Toc113216995)

[**Bayesiano Ingenuo Gaussiano:** 15](#_Toc113216996)

[**Comparación:** 16](#_Toc113216997)

[**Resultados:** 16](#_Toc113216998)

[**Problemas:** 16](#_Toc113216999)

[**Referencias:** 17](#_Toc113217000)

# 

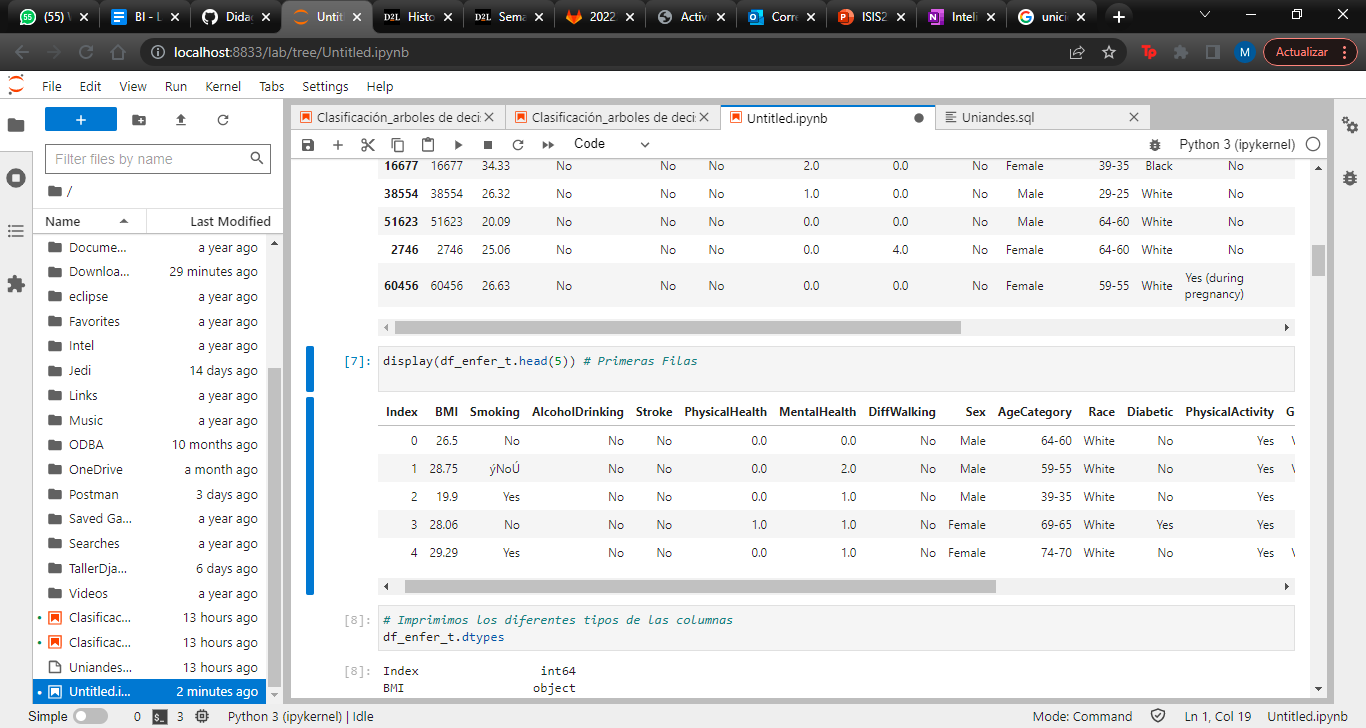
# **Modelos Escogidos:**

* 1. KNN
  2. Árboles de decisión
  3. Regression
  4. Naive Gaussian Bayessian

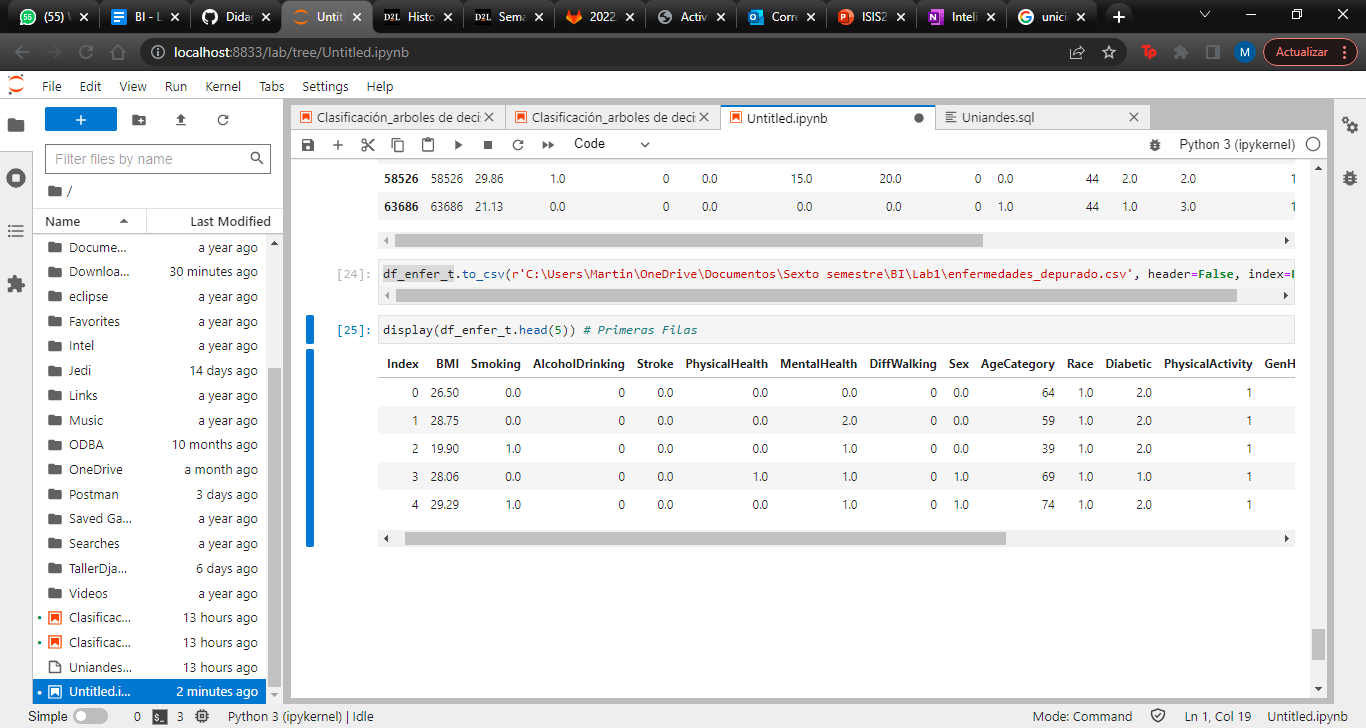
# **Explicación de Tratamiento de Datos:**

Originalmente, nos habían dado 67,500 datos, sin embargo, algunos valores en las distintas columnas no pusieron ninguna de las posibles opciones como respuesta. Cuando esto pasó, y si no contenía alguna de las posibles respuestas dentro de lo que se escribió, entonces todos los datos de esa persona se borraban. Por esta razón, al final había 40607 datos. Esto significa que al final se borraron 26892 datos, ya que una o más casillas tenían en ellas datos erróneos. El proceso utilizado para modificar los datos con el fin de conseguir los modelos más detallados siguió el protocolo ASUM-DM. Una vez se recibieron los datos, el primer paso fue intentar realizar un perfilamiento de los datos. Tras la carga y la exploración de dichos datos, se encontró una gran falla en la calidad de estos. Debido a esta falta, dicho perfilamiento se vio afectado, por lo que se realizó una purificación de los datos. Los datos no tenían duplicados, su principal fallo fue la validez, ya que había muchos valores ingresados incorrectamente, los cuales fueron corregidos para que fueran consistentes con los demás datos o, en el caso de que no pudieran ser identificados para ser corregidos, fueron eliminados. Además de esto, había una irreparable falta de completitud de los datos por dos razones. La primera era el hecho de que, según la empresa, la encuesta recopiló 100,000 datos, pero solo nos brindaron 67,500. La segunda fue el hecho de que la falta de calidad nos obligó a eliminar la cantidad mencionada anteriormente. Una vez se logró verificar la validez de los datos, fueron modificados de tal forma que pudieran ser utilizados por los algoritmos. Para esto, convertimos todos los datos en datos numéricos de forma que los mismos datos puedan ser utilizados por cualquier algoritmo. La forma en la que se modificaron dichos datos está explicada en la tabla del punto 3. A continuación se ve un ejemplo de los datos antes y después de ser procesados.

Antes:



Después:

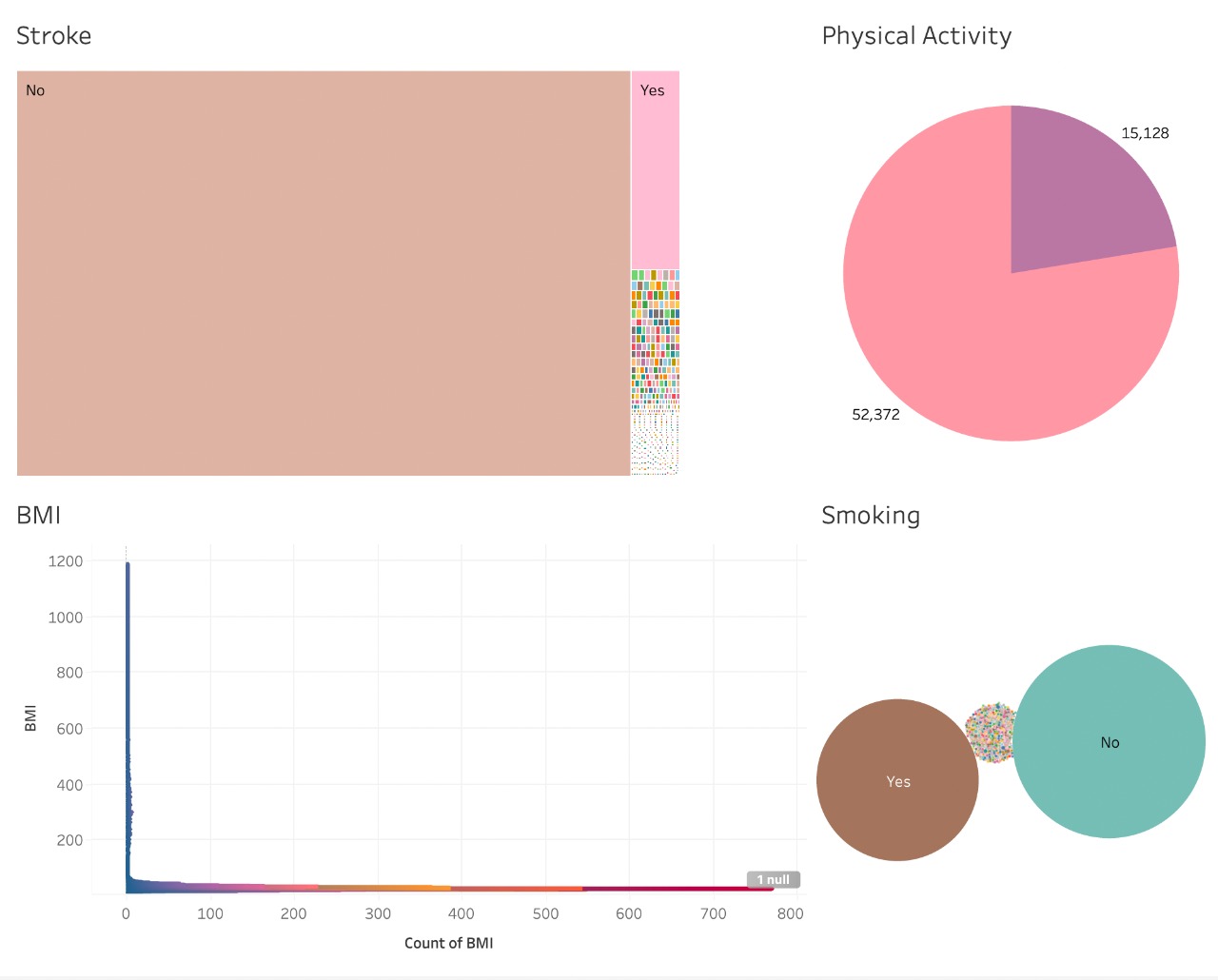
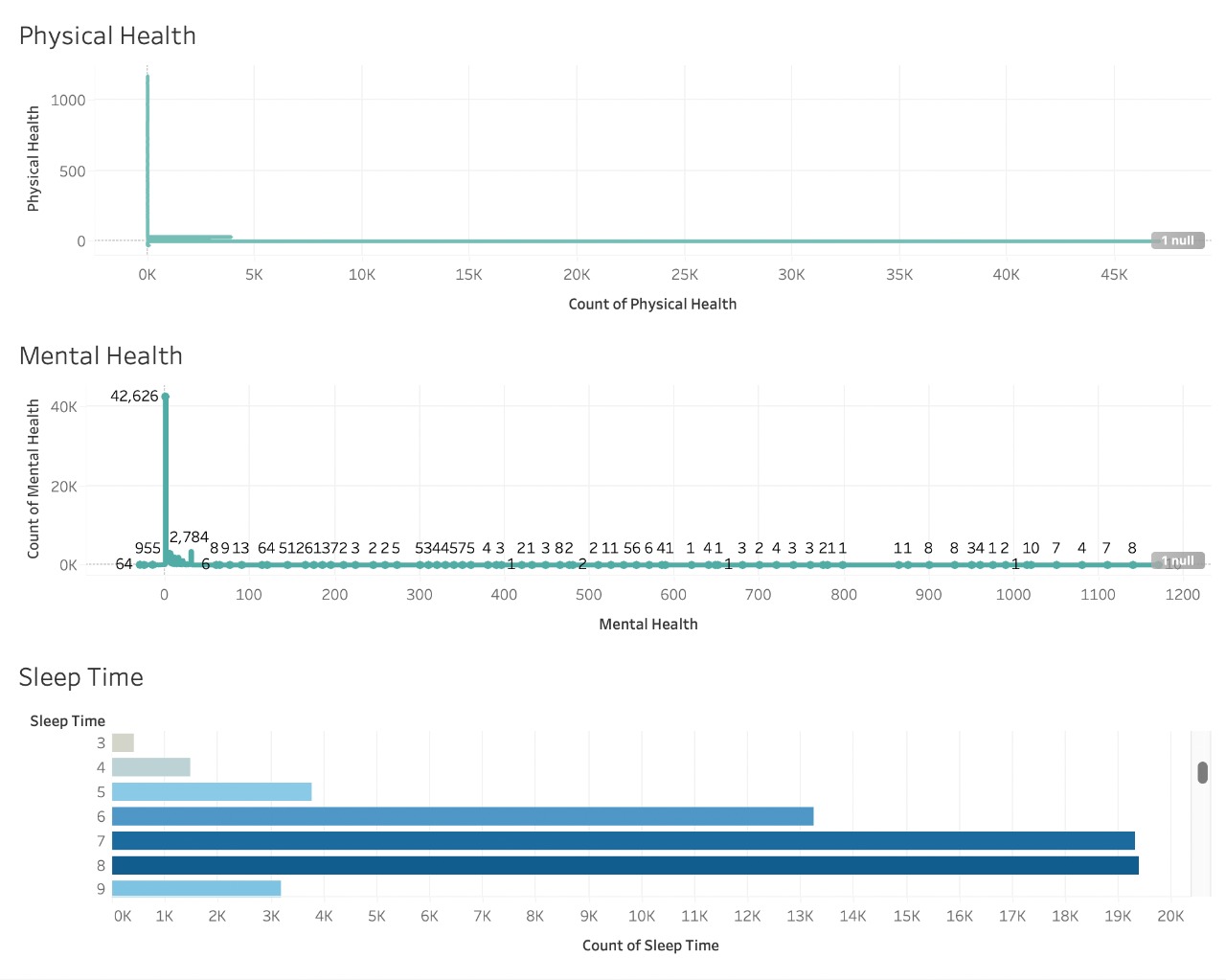


# **Explicación de la tabla:**

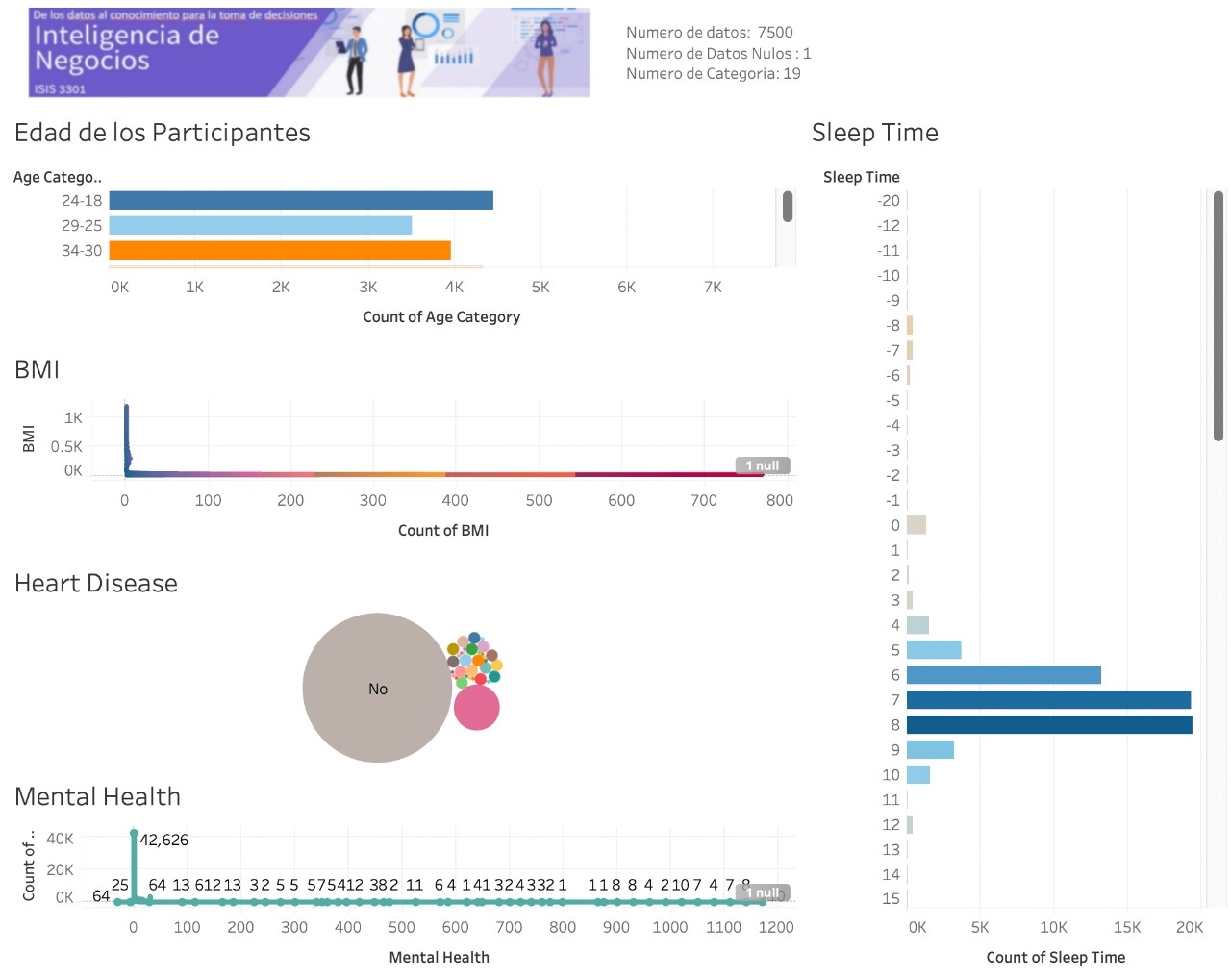
|  |  |  |
| --- | --- | --- |
| Column name | Initial values | Modified values |
| HeartDisease | Yes/No | 1/0 |
| BMI | 1 < x < 99 | 1 < x < 99 |
| Smoking | Yes/No | 1/0 |
| AlcoholDrinking | Yes/No | 1/0 |
| Stroke | Yes/No | 1/0 |
| PhysicalHealth | 0 <= x <= 30 | 0 <= x <= 30 |
| MentalHealth | 0 <= x <= 30 | 0 <= x <= 30 |
| DiffWalking | Yes/No | 1/0 |
| Sex | F/M | 1/0 |
| AgeCategory | 50 <= x | 50 <= x <= 80 |
| Diabetic | Yes/No | 1/0 |
| PhysicalActivity | Yes/No | 1/0 |
| GenHealth | Poor/ Fair/ Good/Very Good/ Excellent | 1/2/3/4/5 |
| SleepTime | 0<=x<=24 | 0<=x<=24 |
| Asthma | Yes/No | 1/0 |
| KidneyDisease | Yes/No | 1/0 |
| SkinCancer | Yes/No | 1/0 |

# **Dashboards en Tableau:**

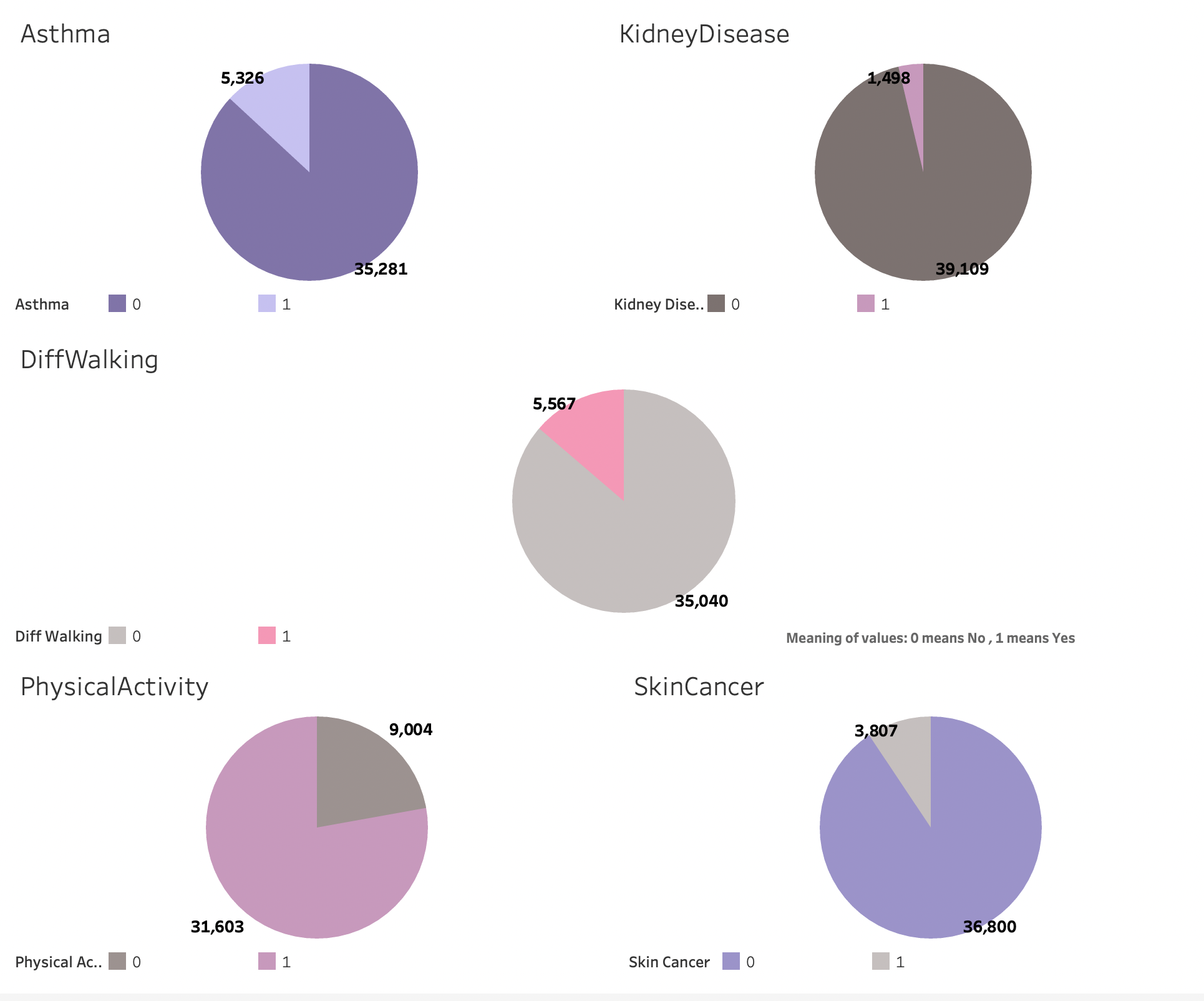
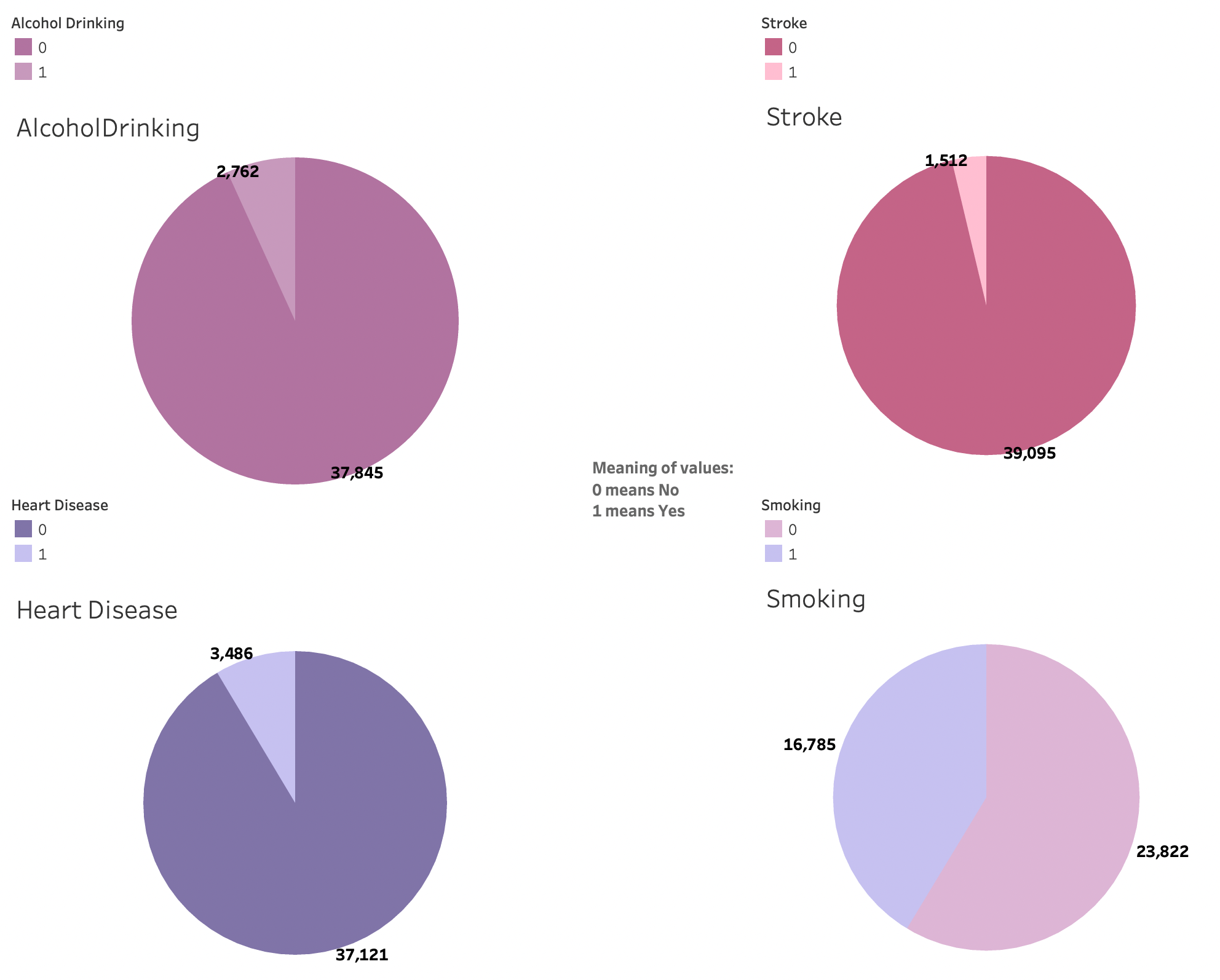
## **Datos Sin Depurar:**

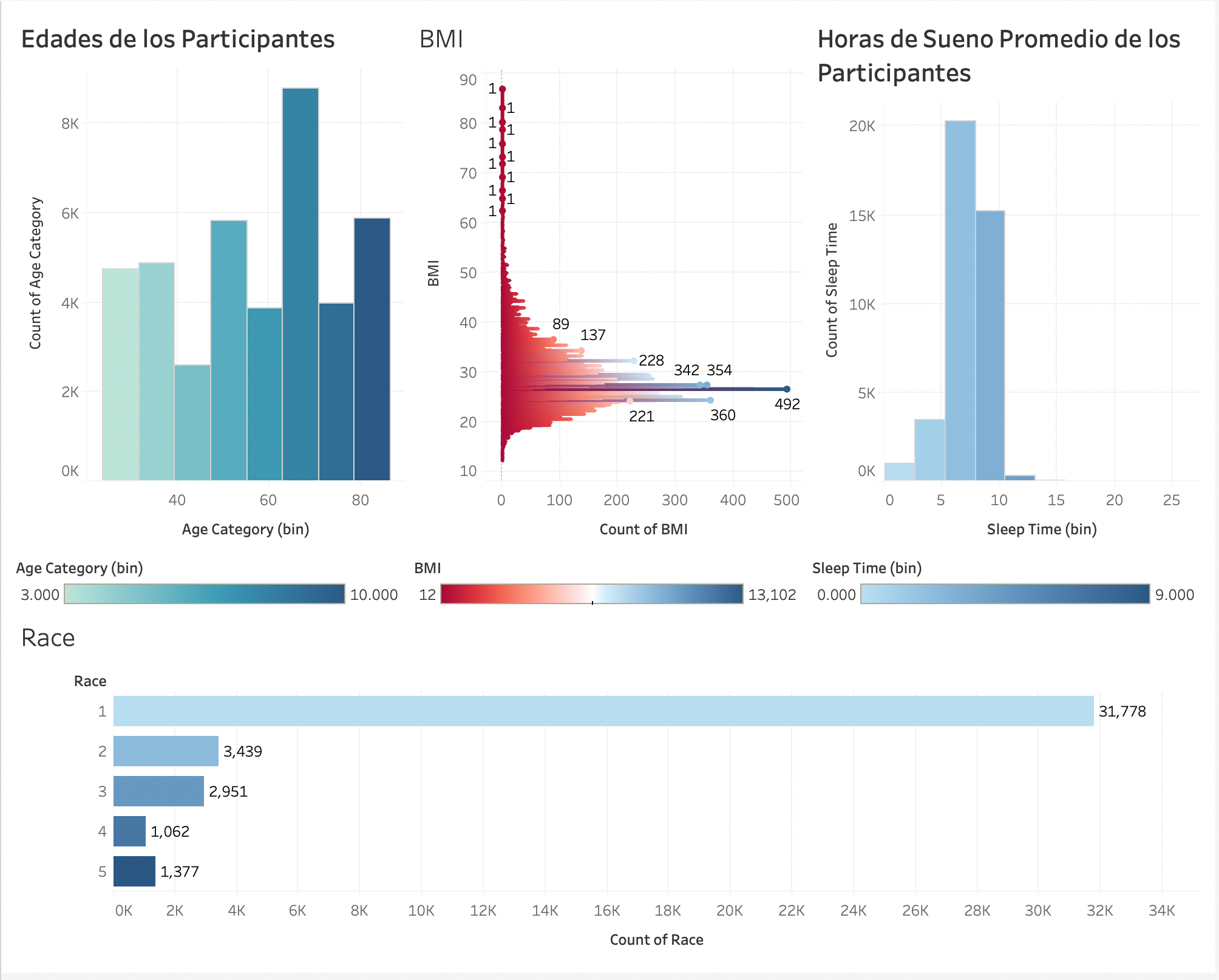
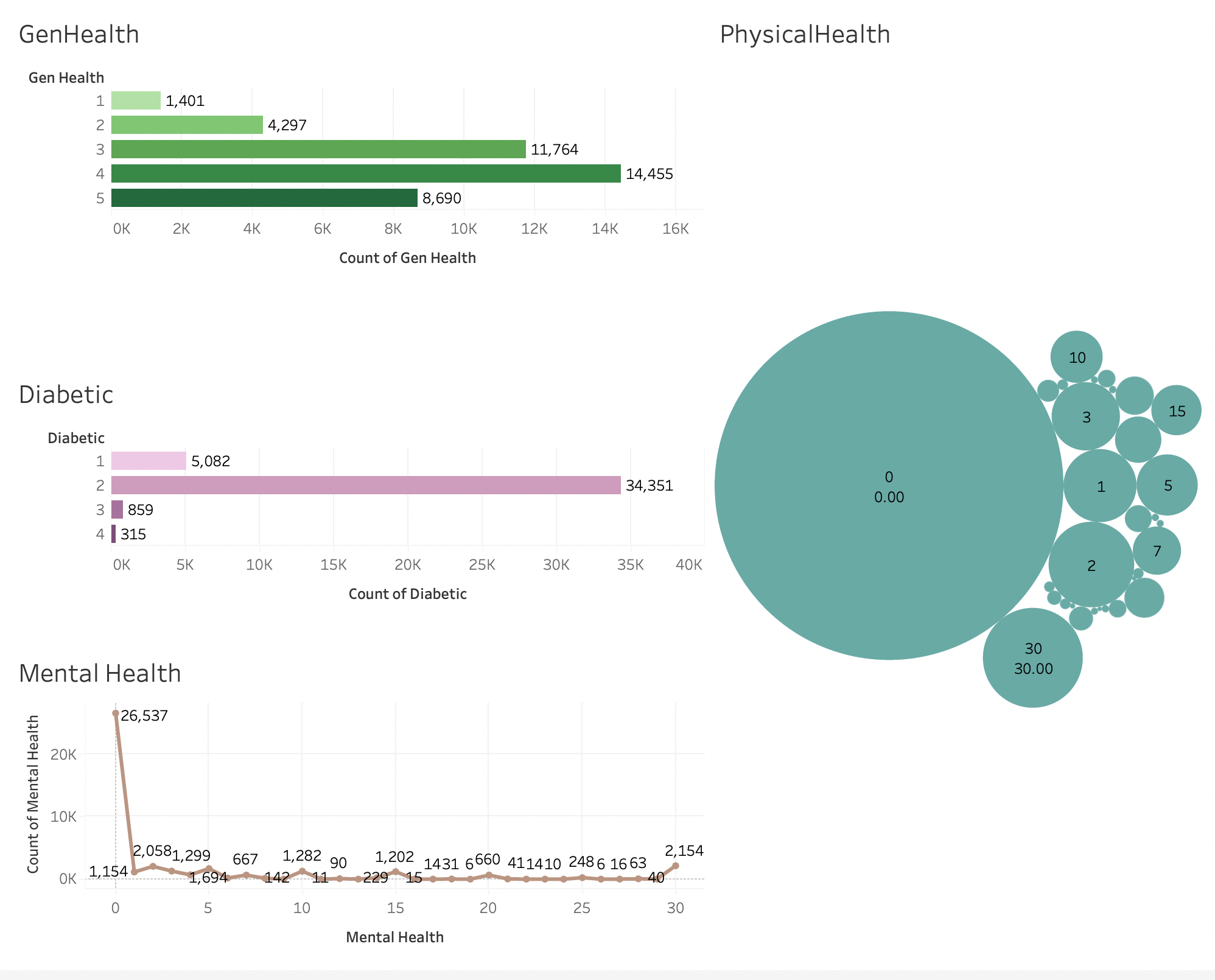


## **Dashboard Unificado:**



## **Datos Depurados:**





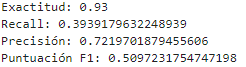
## **Dashboard Unificado:**



# **Modelos de clasificación:**

## **KNN:** Hecho por Martín Ubaque

El algoritmo de KNN es un algoritmo de aprendizaje automático supervisado que basa su funcionamiento en la distancia entre vectores. En este modelo, cada dato es un vector con sus valores siendo los valores de los atributos que lo componen, y estos se utilizan para entrenar el modelo. A partir de aquí, cuando se introduzca un nuevo dato, este medirá la distancia entre este y sus k vecinos más cercanos para decidir así si dicho valor pertenece a una u otra categoría basándose en la categoría de dichos vecinos. Para la creación de este modelo, utilizamos como hiper parámetros 3 vecinos k y la forma euclidiana para medir la distancia entre vecinos. La razón por la que escogimos estos fue el hecho de que con estos valores se obtenía el mejor resultado en la matriz de confusión:



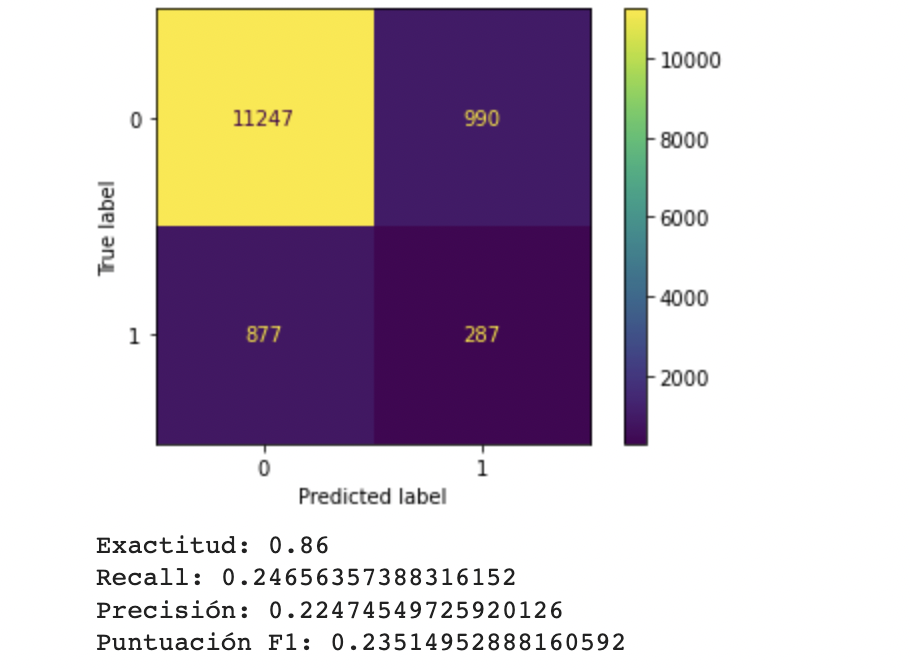
Si bien es cierto que estos valores son muy bajos para ser un modelo aceptable, son los más óptimos a la hora de utilizar este modelo. La razón del uso de dichos hiper parámetros es el hecho de que, una cantidad muy grande de k vecinos va a causar que las diferencias entre sus atributos importen menos, pues va a aceptar más distancia entre vectores, volviendo inútiles dichas diferencias. Sin embargo, una cantidad muy pequeña va a causar lo contrario, volviendo las similitudes entre los vectores casi invisibles, evitando así el reconocimiento de patrones y causando errores. Dicho esto, y teniendo en cuenta que, para este caso, todas las entidades estudiadas tenían datos muy similares en su mayoría (causando así que, a pesar de sus diferencias, los vectores, en general, tuvieron una corta distancia entre sí) se utilizó un k relativamente pequeño, pues la cercanía entre datos que no eran generalmente similares causaría que fuera difícil para el modelo diferenciar entre categorías. Además de esto, se convirtieron los datos a valores numéricos tanto para facilitar el uso de los mismos datos con los demás modelos y para facilitar los cálculos entre distancias de vectores.

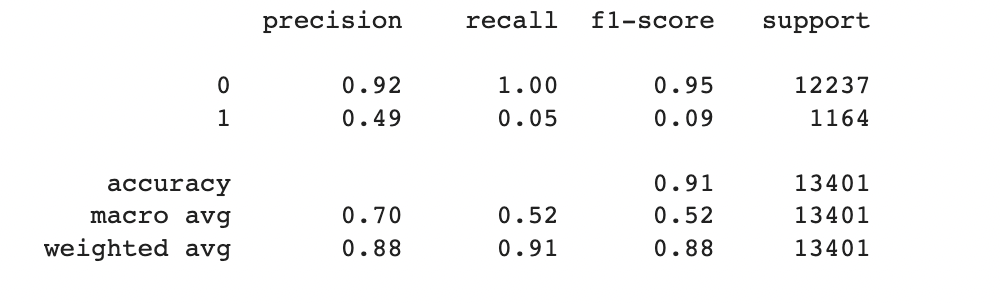
Con los resultados obtenidos con este modelo, se puede concluir que este no es lo suficientemente útil para los fines de SaludAlpes. Con este modelo, se puede detectar fácilmente si un paciente NO tiene/tendrá una enfermedad cardiaca, más no facilita el diagnóstico de un paciente positivo, siendo así inutil para los fines de la empresa.

## **Árboles de decisión:** Hecho por Diego Alejandro Tovar

Árboles de decisión es un algoritmo de aprendizaje supervisado el cual es no paramétrico. Lo que diferencia a este algoritmo es su multifuncionalidad, ya que sirve tanto para tareas de regresión como de clasificación. Como se puede ver en la imagen superior, este árbol es jerárquico, y por ende consta de un nodo raíz, el cual en este caso es “GenHealth”, o la salud general, después van los nos internos, que son 14, y por último los nodos hoja, los cuales son 16.

Este modelo funciona bien, en parte, es por que en la depuración de datos se ilimiron aquellos que no cumplian con el rango estipulado, por ejemplo que BMI fuera mayor a 99, o que el physical health o mental health no estuviera entre 0 y 30, lo cual resultó en que no hubieran muchos outliers y que estos no fueran exagerados, sin que, como se puede ver en el dashboard, la mayoría de datos estuvieran concentrados.

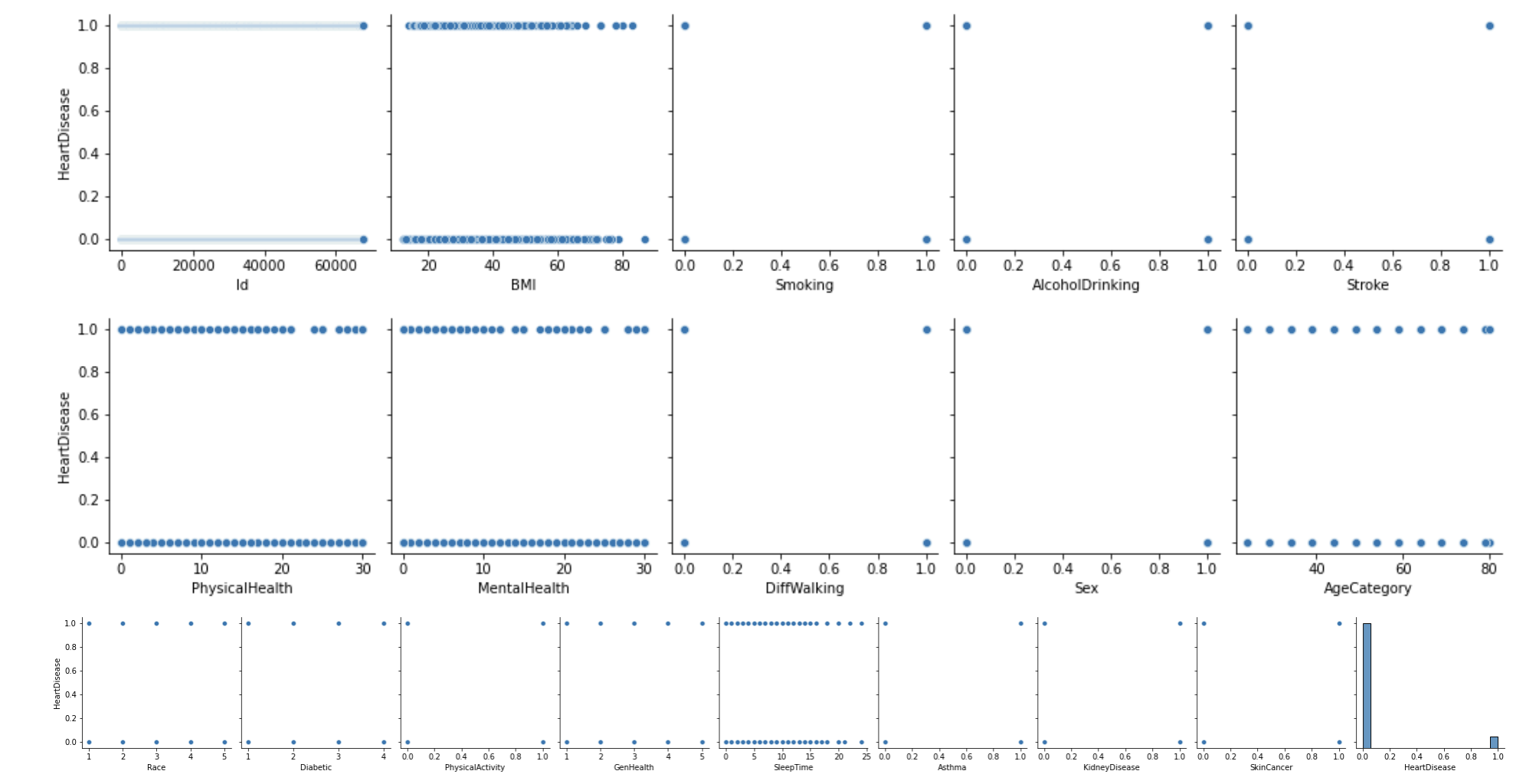


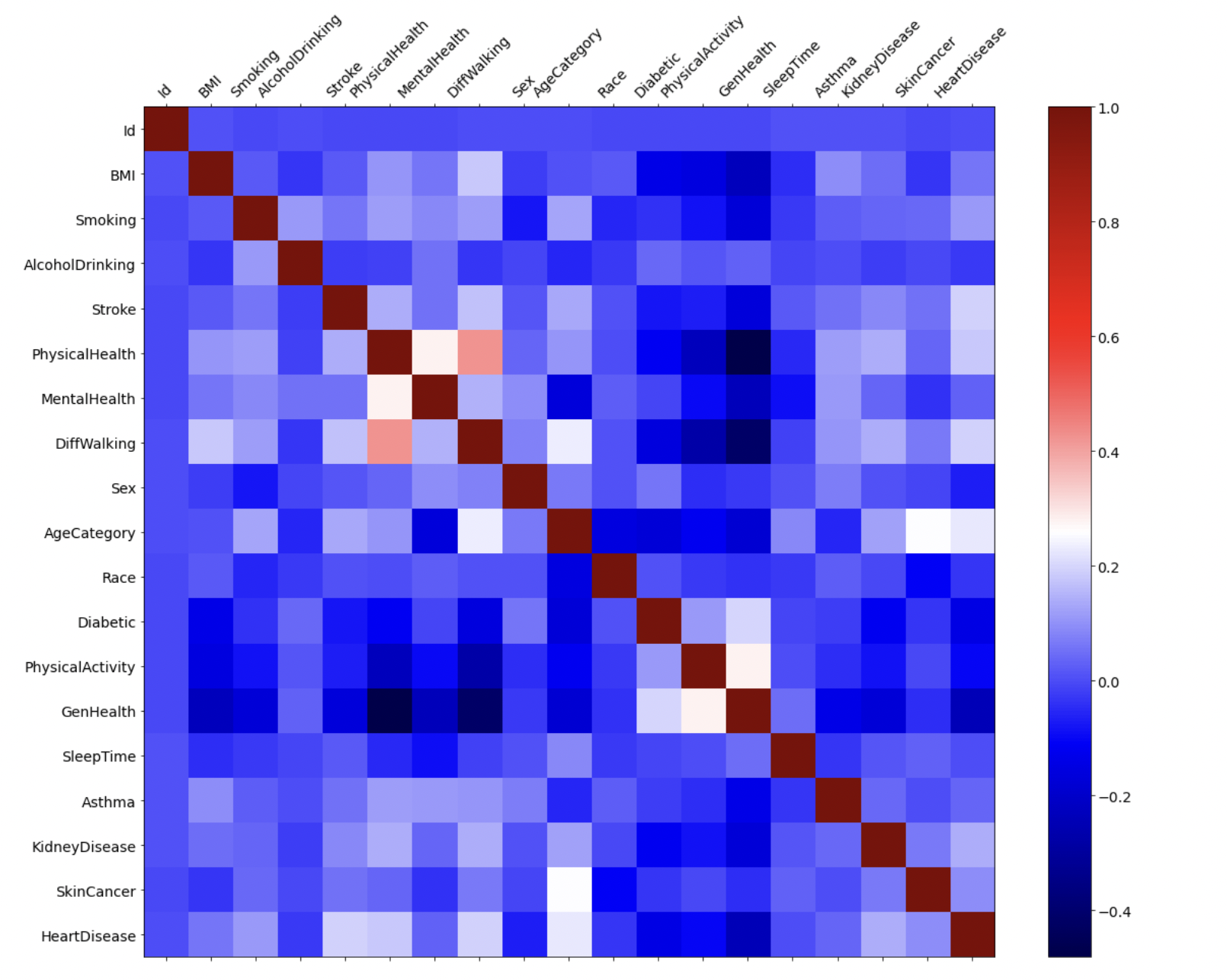


## **Regression:**

Los algoritmos de regresión son algoritmos de aprendizaje supervisado usados para generar una recta la cual proporciona la tendencia de un conjunto de datos. Teniendo en cuenta que habían datos continuos, como el BMI, y datos discretos, como la raza, se usó regresión lineal y regresión logística respectivamente.

Primero, se hizo un diagrama de dispersión, para así poder buscar las relaciones entre las variables. Para hacer esto se seleccionó la variable de “HeartDisease” para poder ver las relaciones de las diferentes variables con esta.



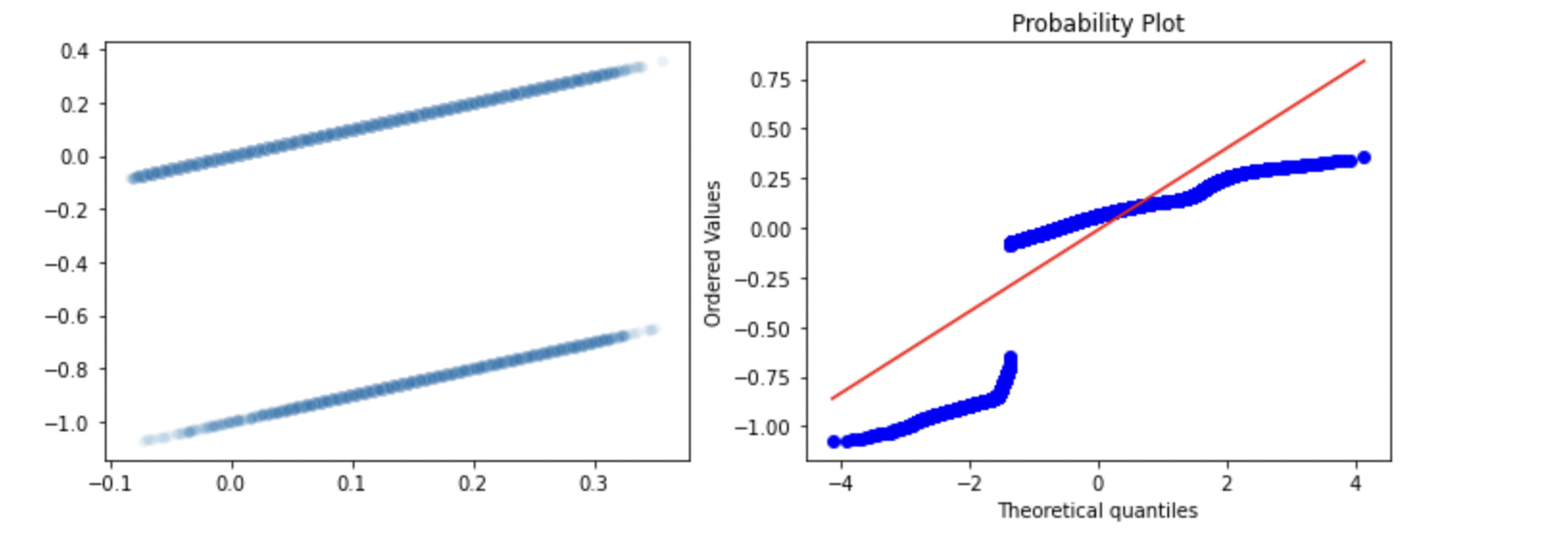
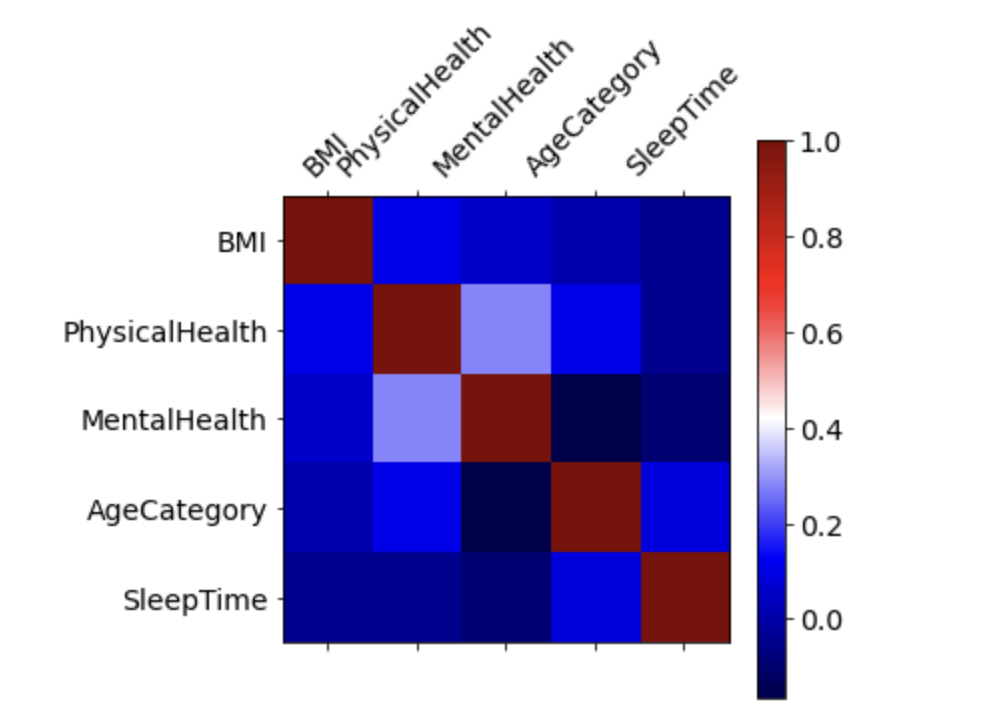
Por otro lado, también se busco las columnas que tiene relación específicamente de correlación con “HeartDisease”. La gráfica que se arrojó se muestra a continuación

Finalmente, se visualiza la regresión lineal de cada dimensión. 

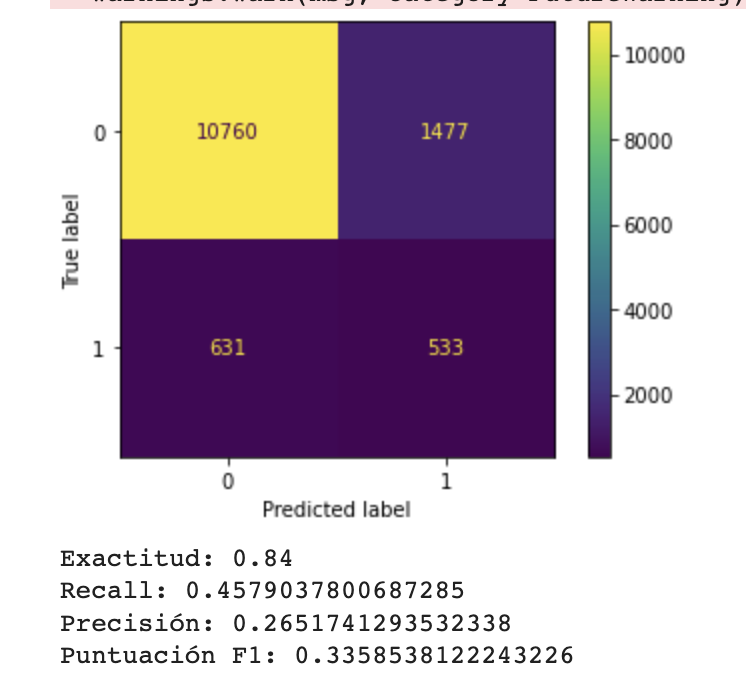
Debido a que este es un modelo de regresión se encontraron las variables del coeficiente de determinación y la raíz del error cuadrático. Esto se encuentra a continuación.

|  |  |
| --- | --- |
| **Coeficiente de Determinación** | **Raíz del Error Cuadrático Medio** |
|  |  |

A continuación se muestra la matriz de calor con las variables que se pensó que más influyen, y la regresión final calculada.



## **Bayesiano Ingenuo Gaussiano:**



El algoritmo Bayesiano Ingenuo Guassiano es un algoritmo que se basa en un clasificador probabilístico fundamentado en el teorema de Bayes. Funciona como una algoritmo rápido y escalable, el cual calcula las probabilidades condicionales para las combinaciones de atributos y el atributo de objetivo (Bayesiano ingenuo de Oracle). La idea es que con base a los distintos datos de entrenamiento, pueda crear una probabilidad independiente, la cual sea replicable. Una suposición de la variante gaussiana del Naive Bayes, es que los atributos siguen una distribución normal.

# 

# **Comparación:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Exactitud** | **Recall** | **Precisión** | **F1** |
| **KNN** | .93 | .3939 | .7220 | .5097 |
| **Árboles de Decisión** | .86 | .2466 | .2247 | .2351 |
| **Árboles de Decisión con Cross Fold Validation (Resampling)** | .91 | .05 | .41 | 0.09 |
| **Naive Gaussian Bayessian** | .84 | .4779 | .2652 | .3359 |

|  |  |  |
| --- | --- | --- |
|  | **Coeficiente de Determinación** | **Raíz del Error Cuadrático Medio** |
| **Regresion Lineal** | 0.0759 | 0.2692 |

# 

# **Resultados:**

Debido a la comparación entre las diferentes métricas (Exactitud, Recall, Precisión y F1), de los datos depurados que se usaron en los 4 algoritmos que hicimos (KNN, Árboles de decisión, Regresión Lineal y Bayesiano Ingenuo Gaussiano), nos dimos cuenta que **el mejor modelo es el KNN**. Esto es debido a que en las métricas, tiene los puntajes más altos (más cerca a 1).

# **Problemas:**

Como se puede ver, hay clases que son mucho más numerosas que otras. Por ejemplo, hay muchas más personas blancas que cualquier otra raza, pero teniendo en cuenta que a la hora de hablar de problemas de corazón, los hombres Negros tienen un 70% más de problabilidad de insuficiencia cardiaca comparado con los hombres blancos, y que las mujeres Negras tienen una probabilidad de 50% más que las mujeres blancas en tener insuficiencia cardiaca, al estos datos de raza Negra no ser tan numerosos como el de raza blanca, se pueden crear sesgos, reflejados en el arbol de decisiones.

# **Referencias:**

<https://my.clevelandclinic.org/health/articles/23051-ethnicity-and-heart-disease#:~:text=Black%20men%20have%20a%2070,be%20hospitalized%20for%20heart%20failure>.

https://www.ibm.com/docs/es/spss-modeler/saas?topic=mining-oracle-naive-bayes