Tecnológico de Monterrey

INSTITUTO TECNOLÓGICO DE ESTUDIOS SUPERIORES DE MONTERREY

Escuela de Ingeniería y Ciencias - Ingeniería en Ciencia de Datos y Matemáticas

Evidencia Final 2

Proyecto de aprendizaje no supervisado

Modelación del aprendizaje con inteligencia artificial (Gpo 302) - TC2034.302

Profesora: Dra. María Valentina Narváez Terán

Equipo 3

Annette Pamela Ruiz Abreu - A01423595

Leslie Ramos Gutiérrez - A01562461

Rodrigo González Zermeño - A00572213

Sarah Dorado Romo - A01540946

Monterrey, Nuevo León.

11 de junio de 2023

Índice

Introducción

La tecnología ha logrado grandes avances en diversos campos, y el área de la salud no es la excepción. La creación e implementación de agentes inteligentes en diversas áreas ha abierto nuevas puertas y cambiado la forma en que se resuelven los problemas de la vida cotidiana. Los agentes inteligentes son sistemas capaces de realizar tareas específicas con autonomía y adaptabilidad, basándose en algoritmos y técnicas de inteligencia artificial. Estos son capaces del razonamiento y del aprendizaje así como el humano. Este reporte tiene como objetivo explorar la fiabilidad de diversos modelos de aprendizaje no supervisado para detectar patrones en una base de datos basándose en diversas variables cuantitativas y cualitativas que nos ayuden a prevenir paros cardiacos. Finalmente, se presentarán los resultados de la implementación de estos modelos, las ventajas de su implementación y las implicaciones éticas del uso de agentes inteligentes en el ámbito de salud.

Problema

Según la Secretaría de Salud, en México, cerca de 177 mil personas murieron por un paro cardíaco en el año 2021. (Secretaría de Salud, 2022) Según los Centros para el Control y Prevención de Enfermedades (CDC), en Estados Unidos cada año mueren alrededor de 2000 jóvenes menores de 25 años por paro cardíaco repentino. (May & Menon, 2023) De acuerdo a la Organización Panamericana de la Salud (OPS), a nivel mundial, las enfermedades cardiovasculares son la mayor causa de muerte. Más de tres cuartas partes de estas muertes se producen en países de bajos y medianos ingresos, donde los casos siguen aumentando. (OPS, 2021) Todas estas cifras y estadísticas nos indican que los paros cardíacos son un gran problema en todo el mundo. Después de la pandemia de COVID-19, la salud es una gran preocupación de varios ciudadanos.

Objetivo

Para poder contribuir a la prevención o detección temprana de posibles paros cardíacos, se crearán diversos modelos de aprendizaje no supervisado como mezcla gaussiana, clustering jerárquico y DBSCAN que podrán detectar patrones entre las variables que nos ayudarán a tomar decisiones para prevenir los paros cardíacos.

Justificación

Elegimos un dataset llamado “Stroke Prediction Dataset”, el cual tiene los datos de pacientes de un hospital y si tuvieron un paro cardíaco o no. Lo elegimos por varias razones; para empezar, el dataset tiene un “label” o una categoría (si tuvo un paro cardíaco o no). Esta categoría no se usará para el entrenamiento del modelo, pero puede usarse para calificar el modelo de aprendizaje no supervisado. En segundo lugar, el dataset tiene varias variables de respuesta el modelo puede usar y tiene 5000 registros. Finalmente, lo escogimos porque como se mencionó, los paros cardíacos son problemas de salud muy graves que causan la muerte de miles de personas al año. Con este conjunto de datos podremos encontrar patrones en función de los parámetros de entrada como el sexo, la edad, diversas enfermedades y el tabaquismo. Los paros cardíacos son enfermedades comunes que le pueden suceder a cualquier persona y es muy importante identificar las variables que más influyen en que esto le suceda a una persona para poder informar a la gente y tomar medidas para prevenirlos.

Descripción de los datos

Nombre: Stroke Prediction Dataset

Dataset: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Repositorio del proyecto:

https://github.com/PameRuiz25/Equipo3_AgentesInteligentes_Reto.git

Cuaderno en Google Colab:

https://colab.research.google.com/github/PameRuiz25/Equipo3_AgentesInteligentes_Reto/blob/main/Evidencia2/Equipo3_E2_AgentesInteligentes.ipynb

Descripción inicial:

El dataset se descargó de Kaggle, una de las plataformas web más grandes del mundo para científicos de datos con millones de bases de datos. El origen y cómo fueron muestreados los datos es desconocido.

- Cantidad de datos: 5110
- Cantidad de columnas: 12
- Variables:

Nombre	**Descripción**	**Tipo de Dato**	**Valores Nulos**	**Posibles Valores**
id	Identificador de paciente	integer	0	$67 \leq x \leq 72940$
gender	Sexo del paciente	category	0	Female, Male, Other
age	Edad del paciente. Las edades con punto decimal se refieren a los meses de los infantes	float	0	$0.08 \leq x \leq 82$
hypertension	Si ha tenido hypertension	boolean	0	0, 1
heart_disease	Si ha tenido una enfermedad cardiovascular	boolean	0	0, 1

ever_married	Si ha estado casado	boolean	0	Yes, No
work_type	El tipo de trabajo que tiene. Si es niño,	category	0	Private, self-employed, children, Govt_job, Never_worked
Residence_type	El lugar en donde vive	category	0	Urban, rural
avg_glucose_level	Nivel de glucosa	float	0	$55.12 \leq x \leq 271.74$
bmi	Índice de masa corporal	float	201	$7.854067 \leq x \leq 97.6$
smoking_status	Si ha fumado	category	0	never smoked, Unknown, formerly smoked, smokes
stroke	Si ha tenido un paro cardíaco	boolean	0	0, 1

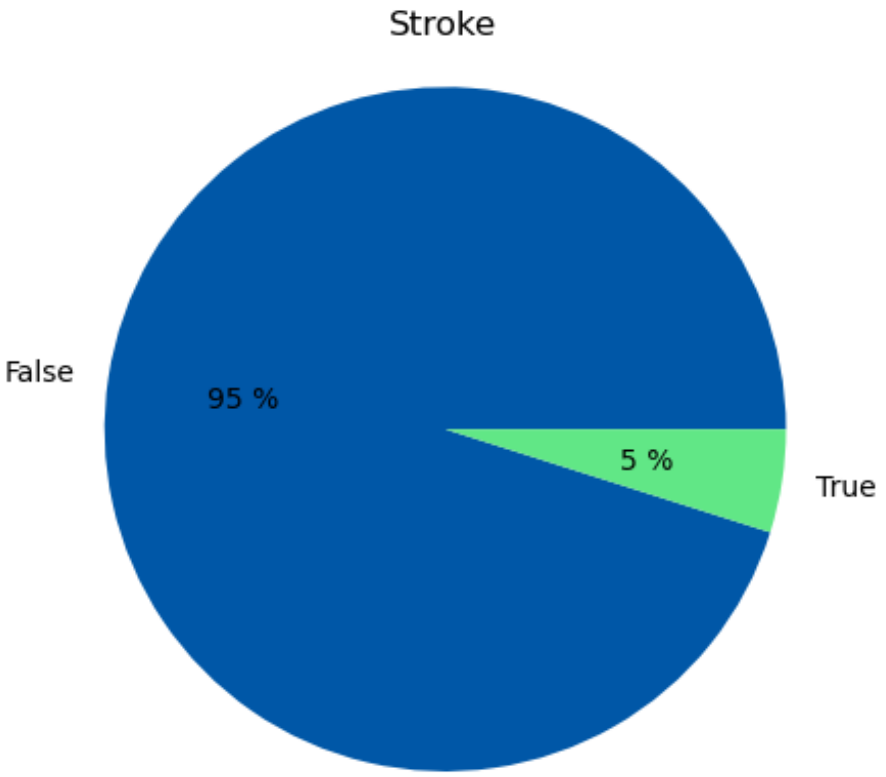


Figura 1: Distribución de valores de paros cardíacos

Metodología

Paleta de colores:

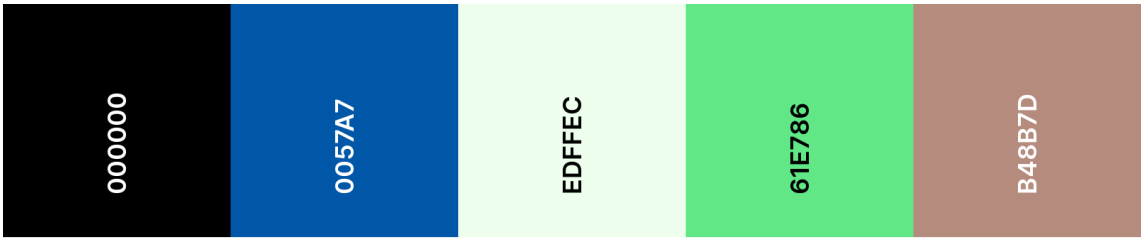


Figura 2: Paleta de colores para las gráficas

Limpieza inicial:

1. Quitamos “Otro” en género.
2. Llenamos las filas vacías con la mediana de “bmi”.
3. Transformamos los tipos de datos.
4. Creación de categorías (binning) para la visualización de datos.

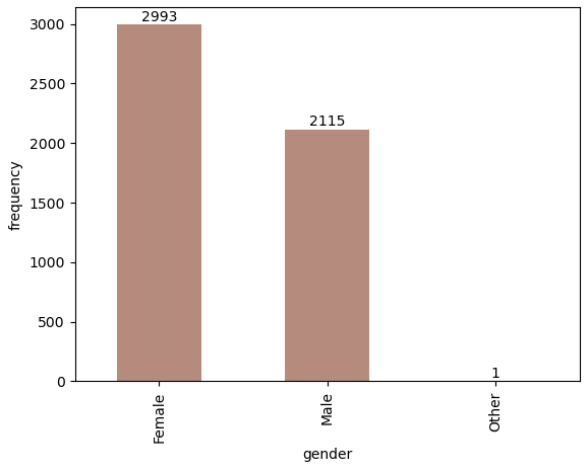
Después de la limpieza, tenemos 4908 filas y 11 columnas.

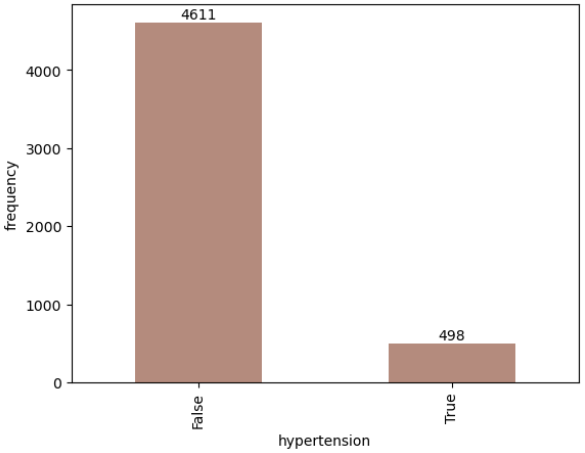
- Estadística descriptiva

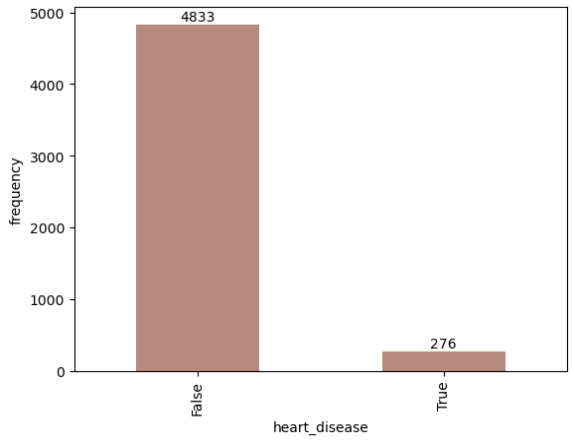
	age	avg_glucose_level	bmi
count	4909.000000	4909.000000	4909.000000
mean	42.865374	105.305150	28.893237
std	22.555115	44.424341	7.854067
min	0.080000	55.120000	10.300000
25%	25.000000	77.070000	23.500000
50%	44.000000	91.680000	28.100000
75%	60.000000	113.570000	33.100000
max	82.000000	271.740000	97.600000

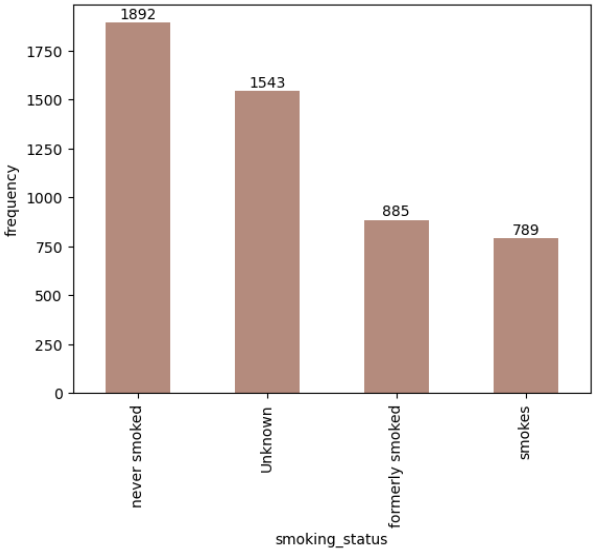
	mode
gender	Female
age	78.0
hypertension	False
heart_disease	False
ever_married	True
work_type	Private
Residence_type	Urban
avg_glucose_level	93.88
bmi	28.7
smoking_status	never smoked
stroke	False

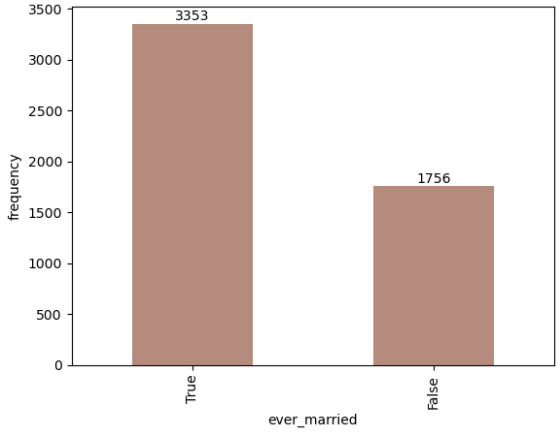
Figura 3: Tablas de estadísticas descriptivas

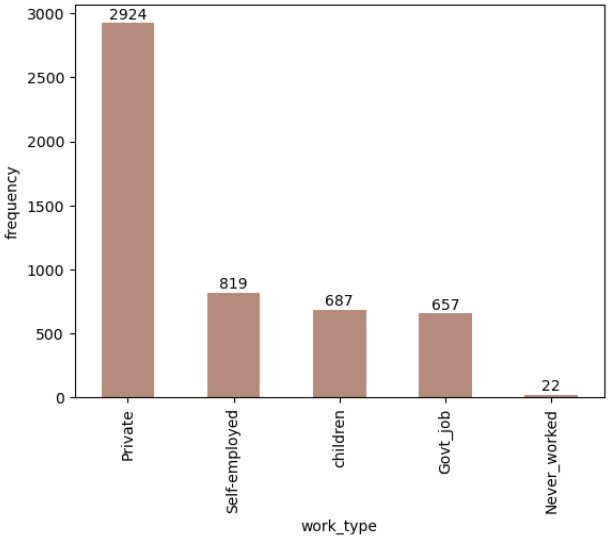

2993
2115
1
frequency
Female
Male
Other
gender


4611
498
frequency
False
True
hypertension


4833
276
frequency
False
True
heart_disease


1892
1543
885
789
frequency
never smoked
Unknown
formerly smoked
smokes
smoking_status


3353
1756
frequency
True
False
ever_married


2924
819
687
657
22
frequency
Private
Self-employed
children
Govt_job
Never_worked
work_type

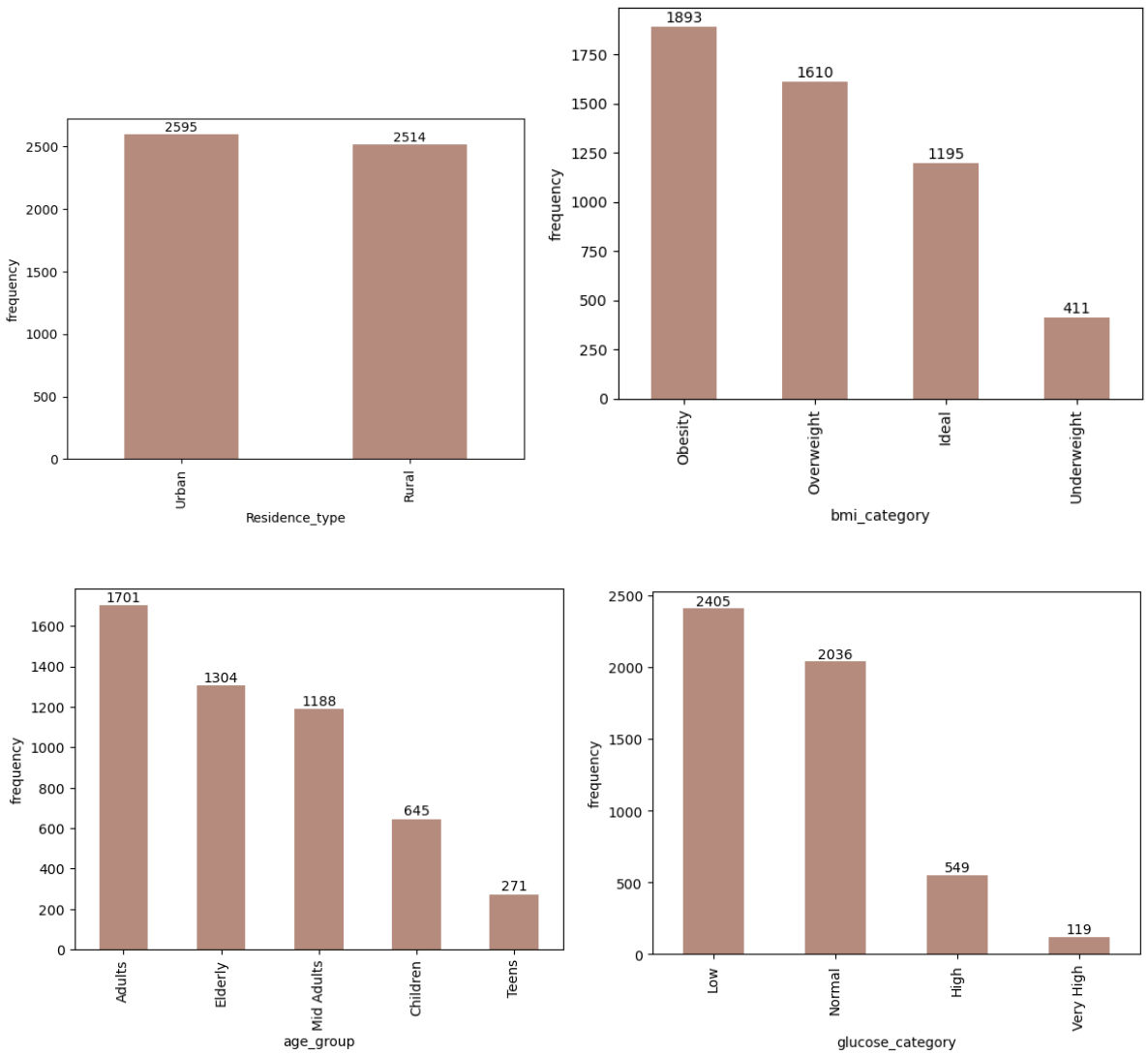


Figura 4: Gráficas de frecuencia de las variables categóricas

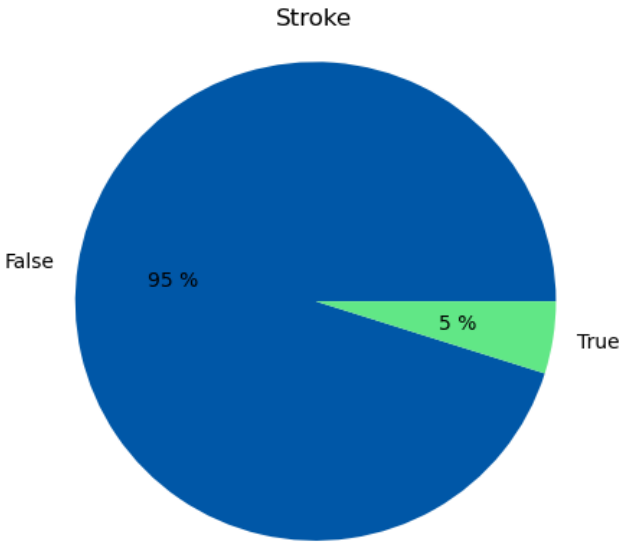


Figura 5: Distribución de valores de paros cardíacos

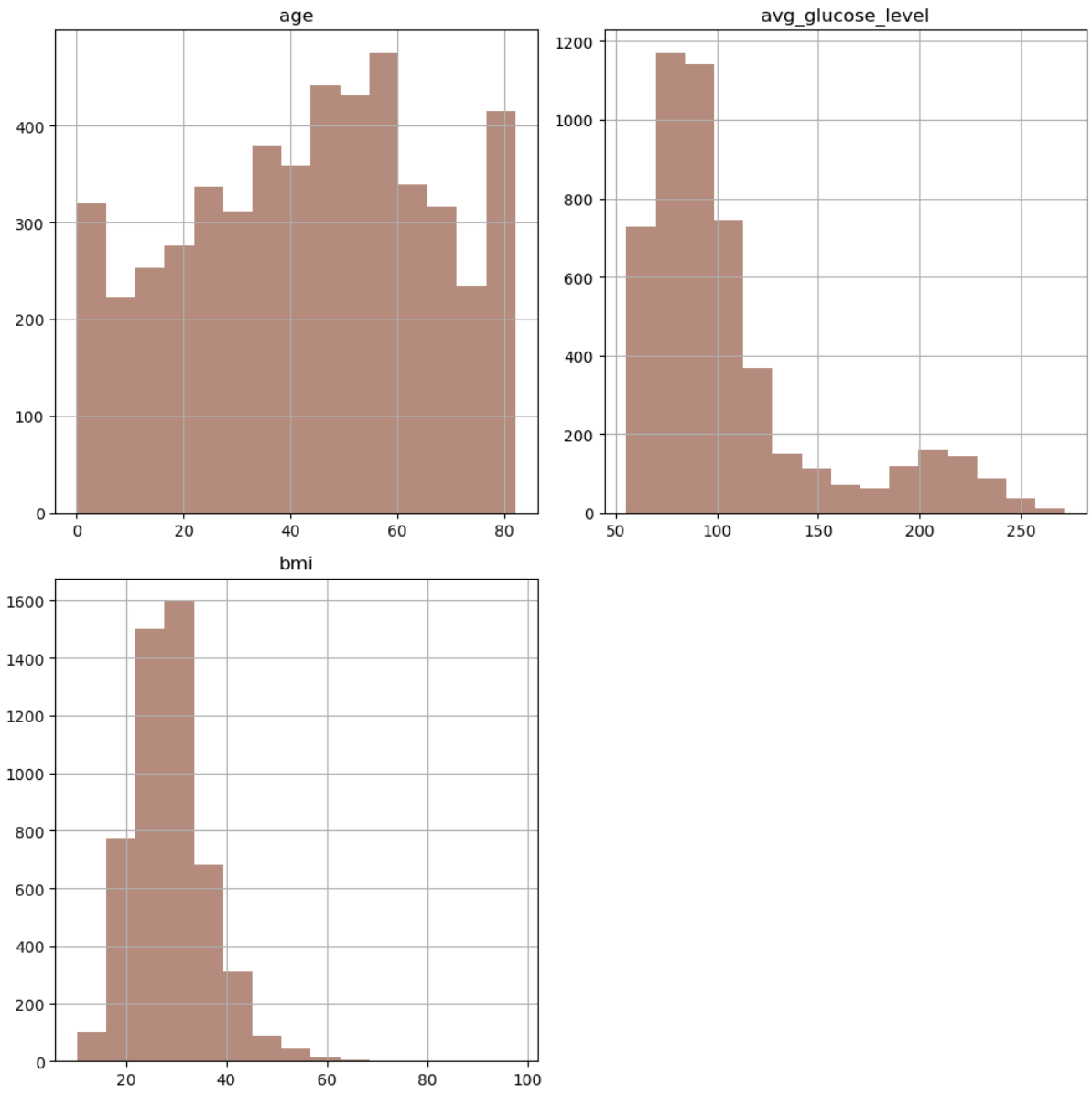


Figura 6: Histogramas de variables cuantitativas

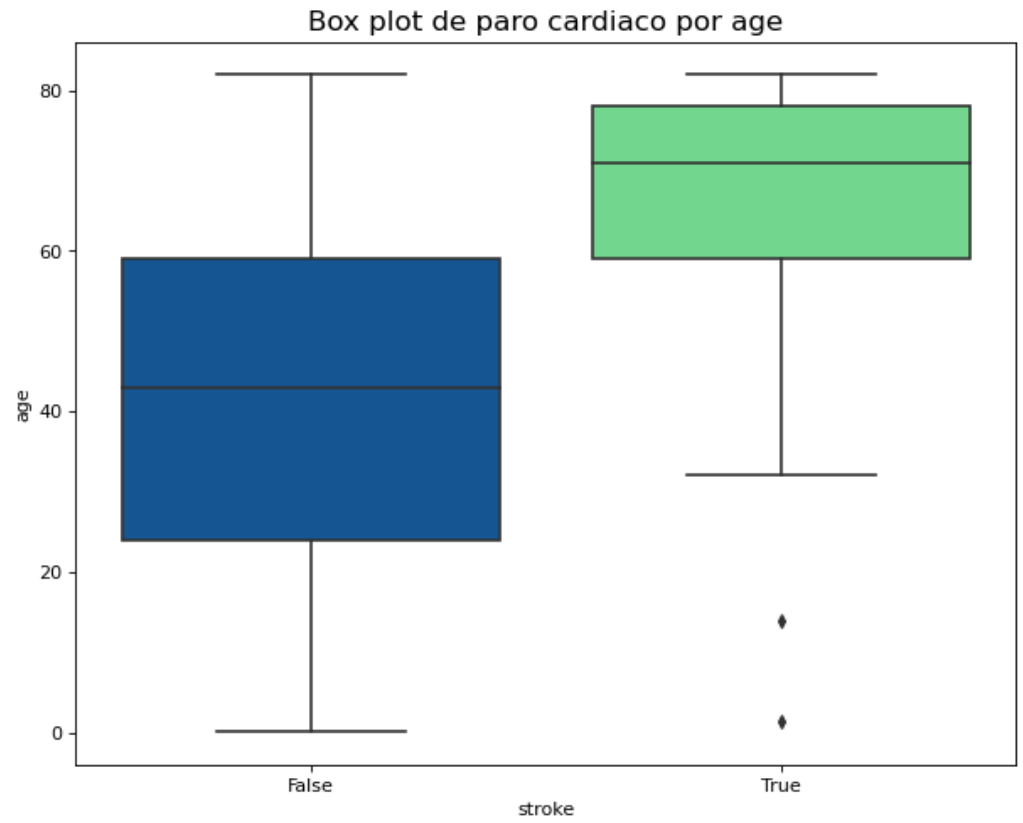


Figura 7: Diagrama de caja y bigotes de paro cardiaco por edad

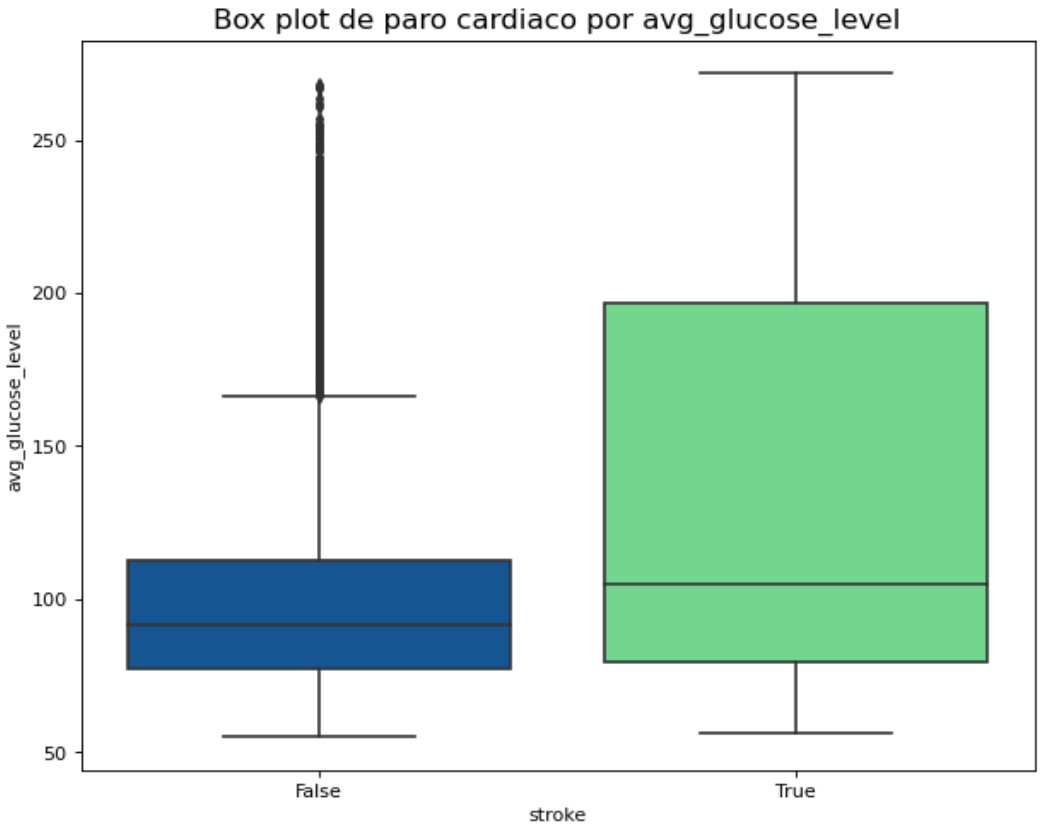


Figura 8: Diagrama de caja y bigotes de paro cardiaco por nivel promedio de glucosa

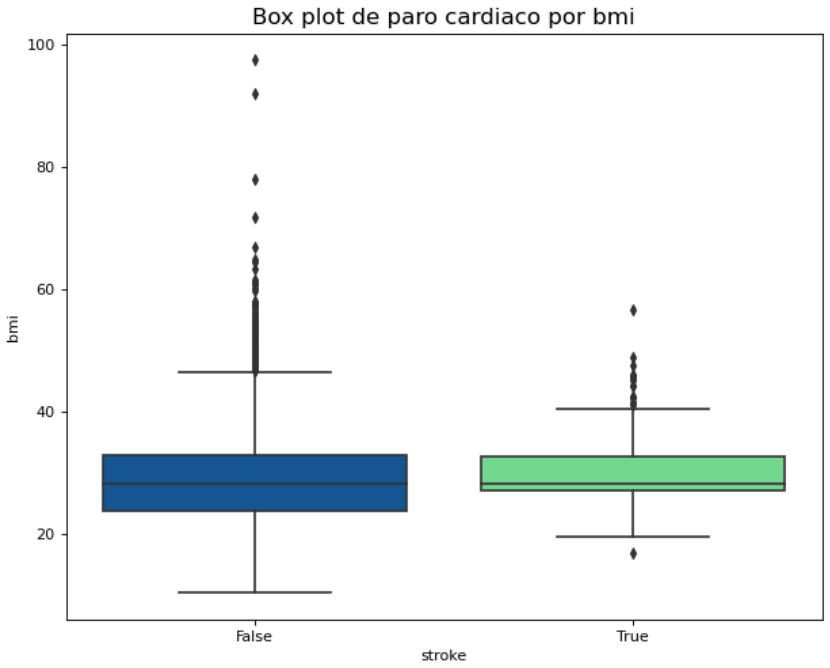


Figura 9: Diagrama de caja y bigotes de paro cardiaco por bmi

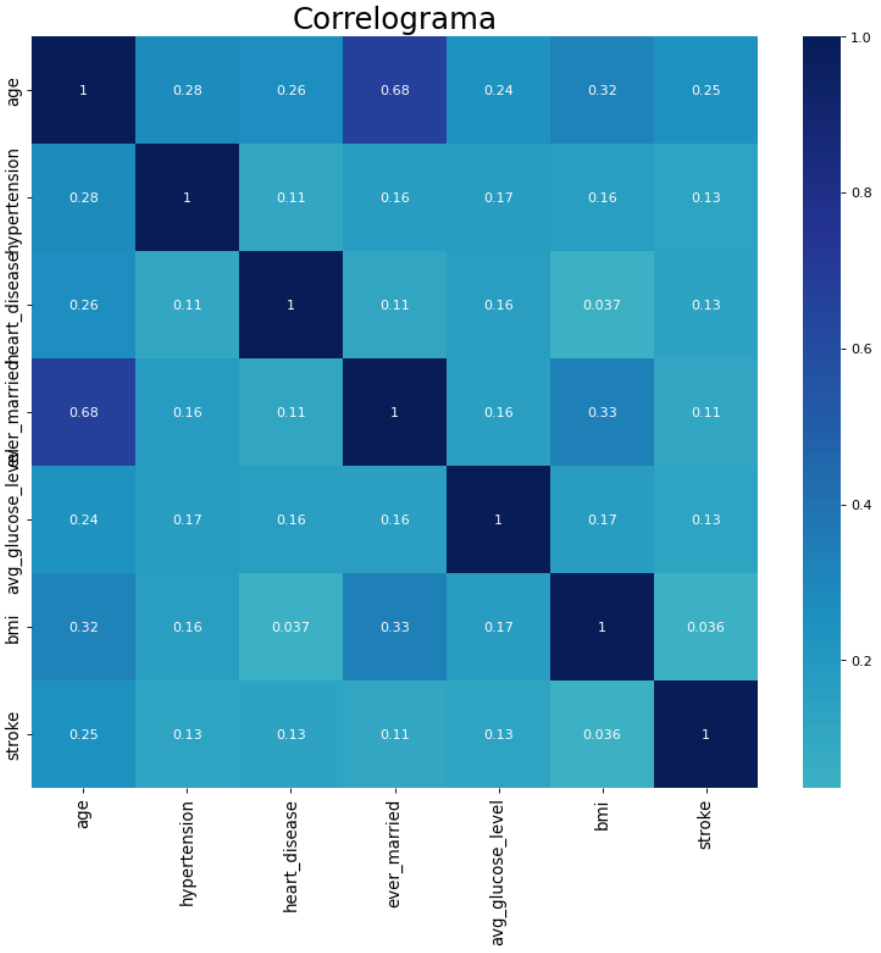


Figura 10: Correlograma de todas las variables

Análisis de paros cardíacos por edad

		stroke
age_group	**stroke**	
Children	False	644
	True	1
Teens	False	270
	True	1
Adults	False	1690
	True	11
Mid Adults	False	1129
	True	59
Elderly	False	1127
	True	177

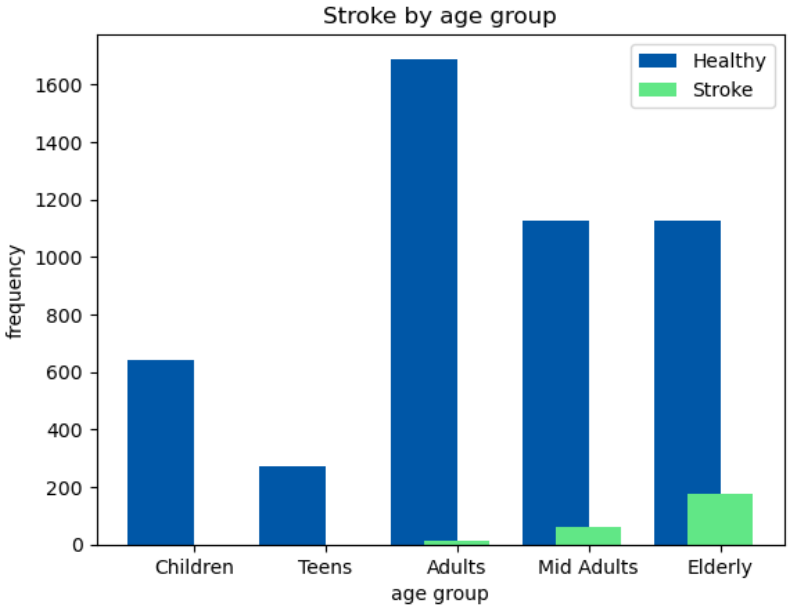


Figura 11: Paros cardíacos por grupo de edad

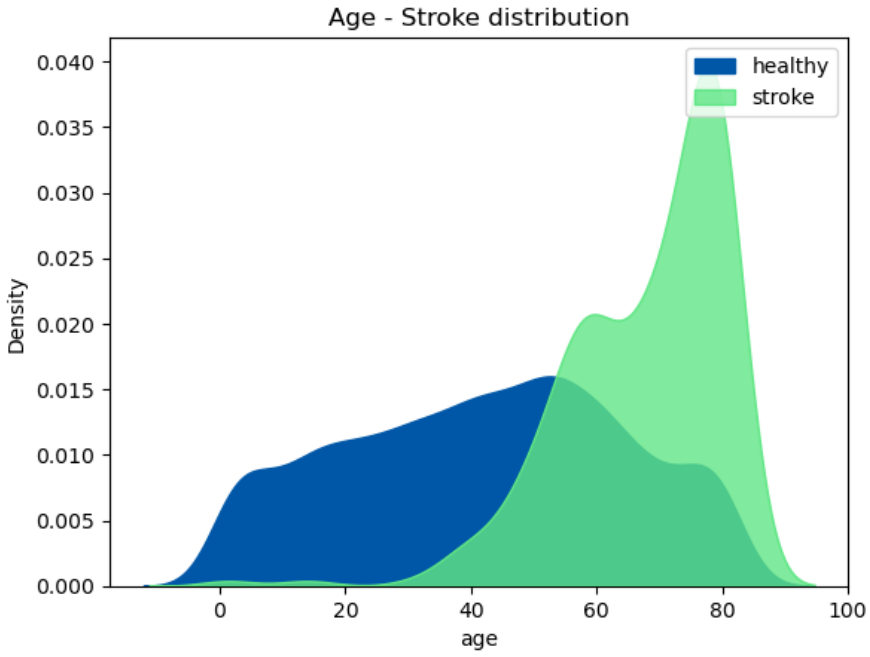


Figura 12: Distribución de paros cardíacos por edad

Con ayuda de las gráficas podemos confirmar nuestra hipótesis de que entre mayor seas de edad, más probabilidad hay de que sufras un paro cardíaco.

Análisis de paros cardíacos por nivel de glucosa

		stroke
glucose_category	**stroke**	
Low	False	2316
	True	89
Normal	False	1965
	True	71
High	False	478
	True	71
Very High	False	101
	True	18

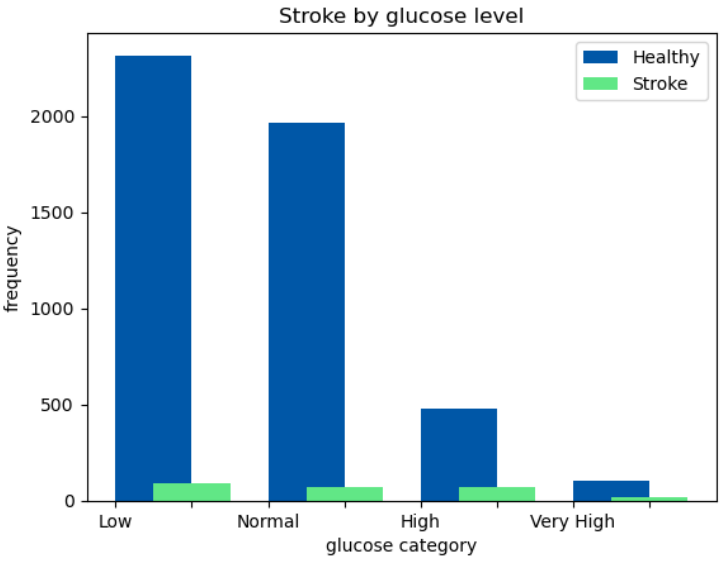


Figura 13: Paros cardíacos por nivel de glucosa

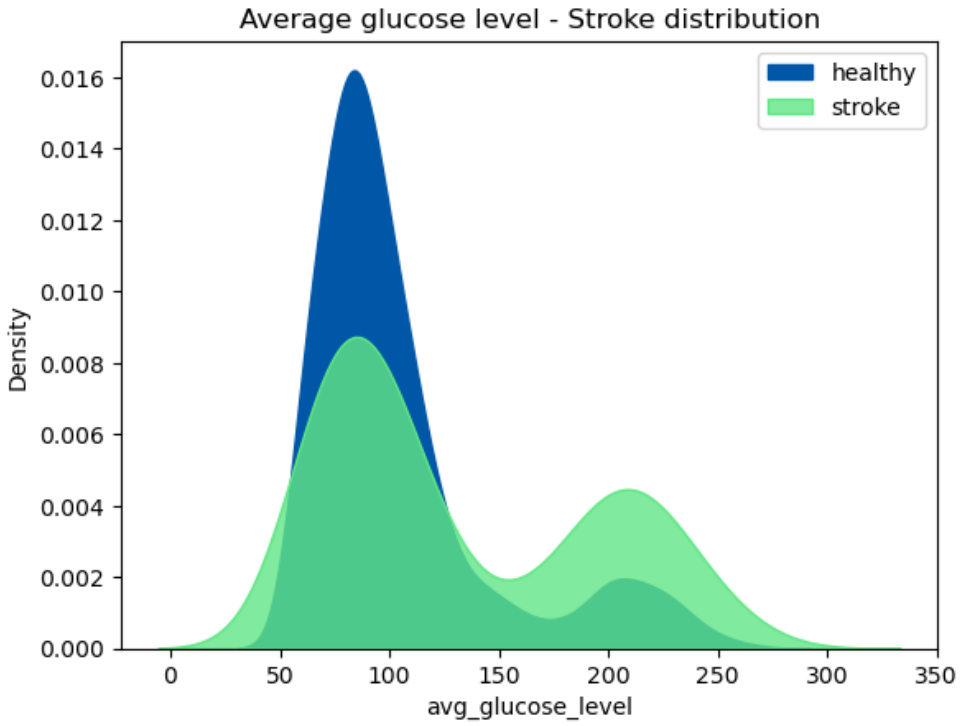


Figura 14: Distribución de paros cardíacos por nivel de glucosa

Lamentablemente, no podemos concluir si hay una relación entre el nivel de glucosa y el riesgo de sufrir un paro cardíaco; ya que en las gráficas no existe algún sesgo en relación con la glucosa y los paros cardíacos.

Análisis de paros cardíacos por índice de masa corporal

		stroke
bmi_category	**stroke**	
Underweight	False	410
	True	1
Ideal	False	1158
	True	37
Overweight	False	1495
	True	115
Obesity	False	1797
	True	96

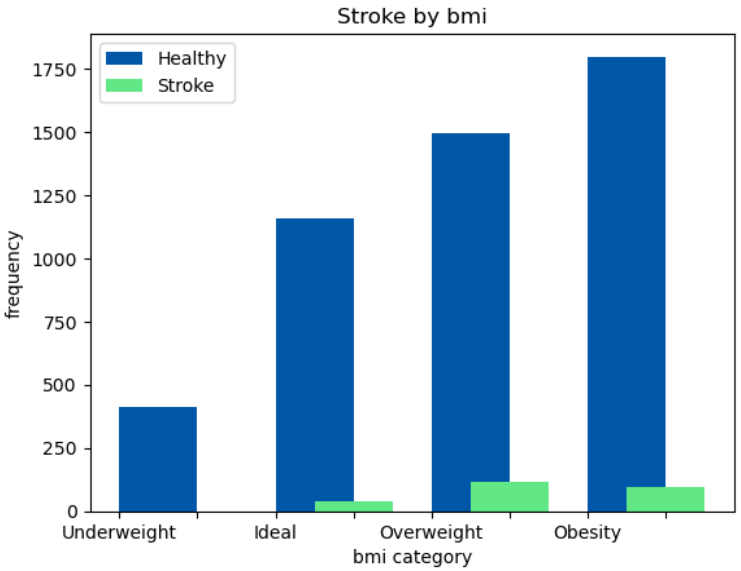


Figura 15: Paros cardíacos por índice de masa corporal

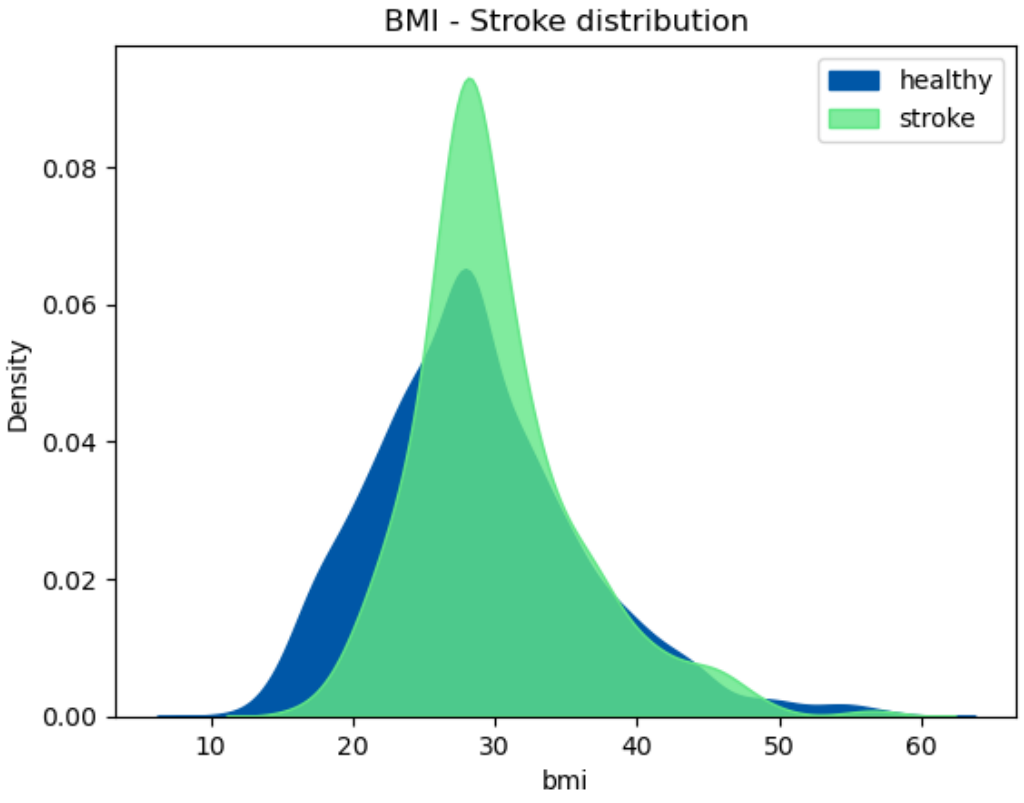


Figura 16: Distribución de paros cardíacos por índice de masa corporal

Análisis de paros cardíacos por variables categóricas

		stroke
gender	**stroke**	
Female	False	2853
	True	141
Male	False	2007
	True	108

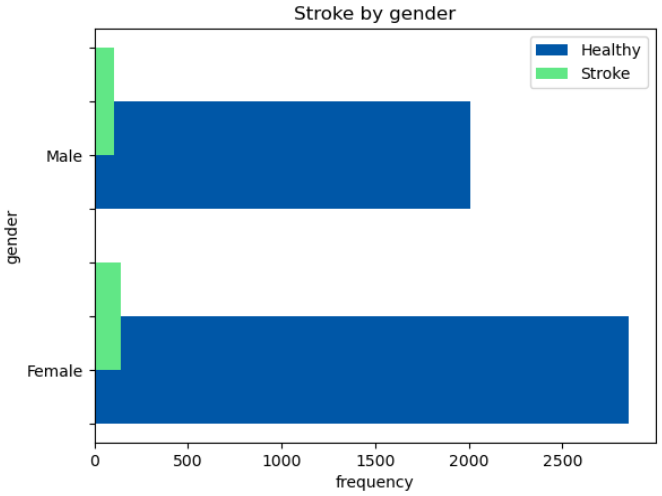


Figura 17: Análisis de paros cardíacos por género

		stroke
hypertension	**stroke**	
False	False	4428
	True	183
True	False	432
	True	66

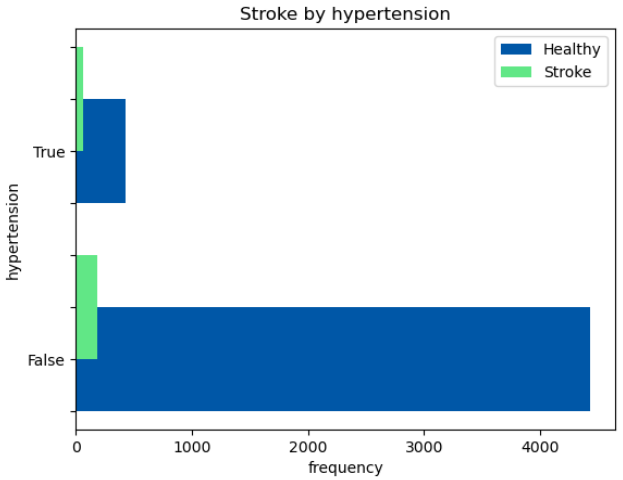


Figura 18: Análisis de paros cardíacos por hipertensión

		stroke
heart_disease	**stroke**	
False	False	4631
	True	202
True	False	229
	True	47

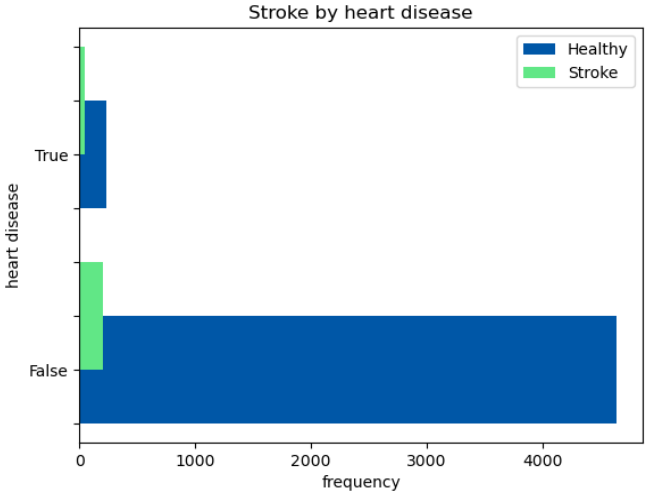


Figura 19: Análisis de paros cardíacos por problemas de corazón

		stroke
ever_married	**stroke**	
False	False	1727
	True	29
True	False	3133
	True	220

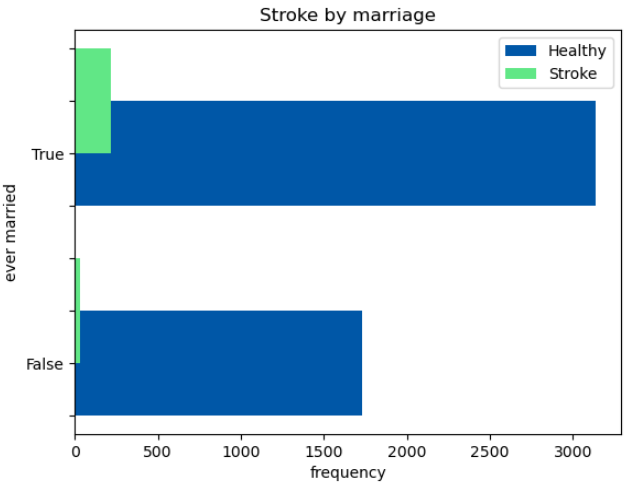


Figura 20: Análisis de paros cardíacos por matrimonio

		stroke
work_type	**stroke**	
Govt_job	False	624
	True	33
Never_worked	False	22
Private	False	2775
	True	149
Self-employed	False	754
	True	65

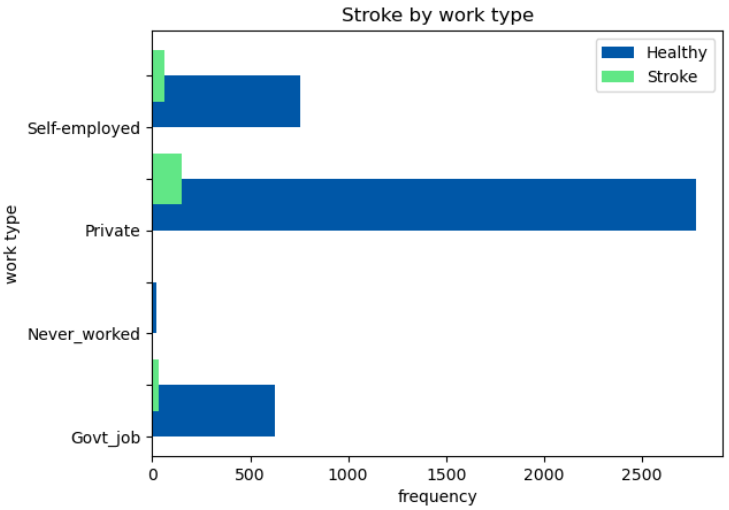


Figura 21: Análisis de paros cardíacos por tipo de trabajo

		stroke
Residence_type	**stroke**	
Rural	False	2399
	True	114
Urban	False	2461
	True	135

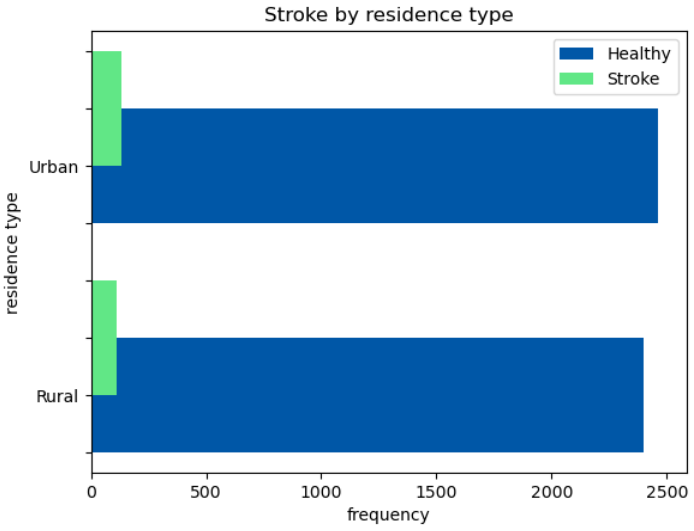


Figura 22: Análisis de paros cardíacos por lugar de residencia

		stroke
smoking_status	stroke	
Unknown	False	1497
	True	47
formerly smoked	False	814
	True	70
never smoked	False	1802
	True	90
smokes	False	747
	True	42

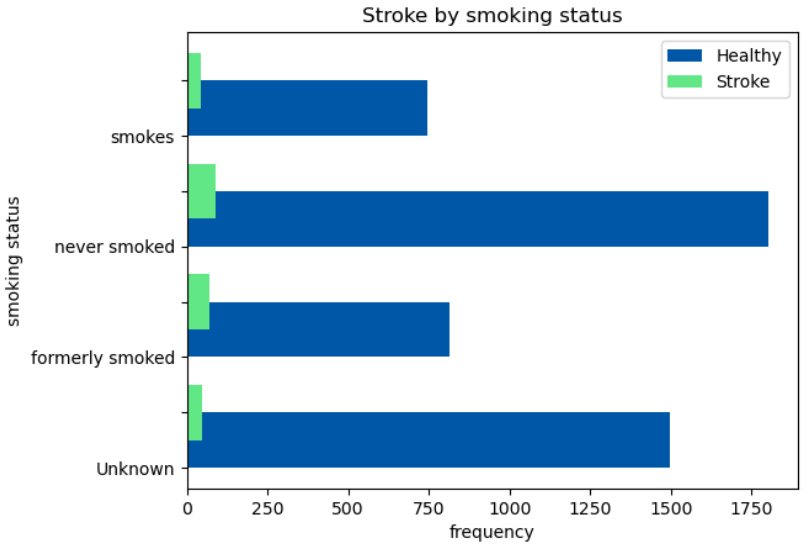


Figura 23: Análisis de paros cardíacos por historial de fumador

Como se ve en la figura 24, el género no impacta en las probabilidades de sufrir un paro cardíaco.

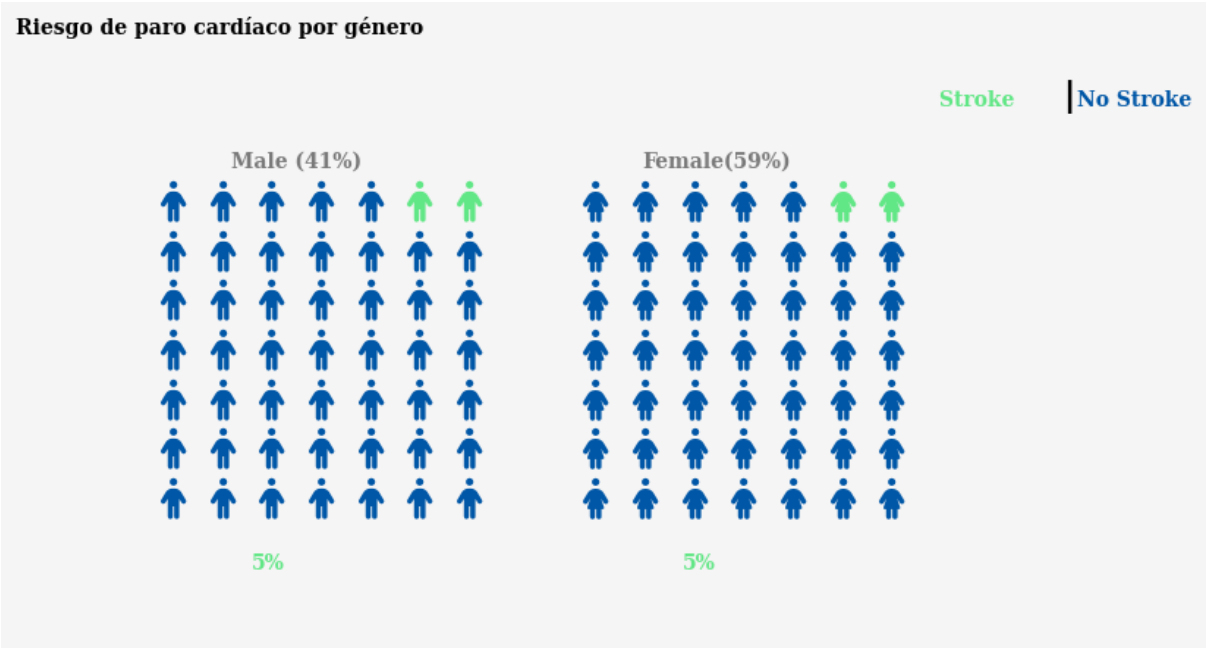


Figura 24: Comparación de probabilidades de paros cardíacos por género

En las figuras 25 y 26 podemos observar que tener hipertensión triplica las probabilidades de sufrir un paro cardíaco y tener enfermedades de corazón las cuadruplica.

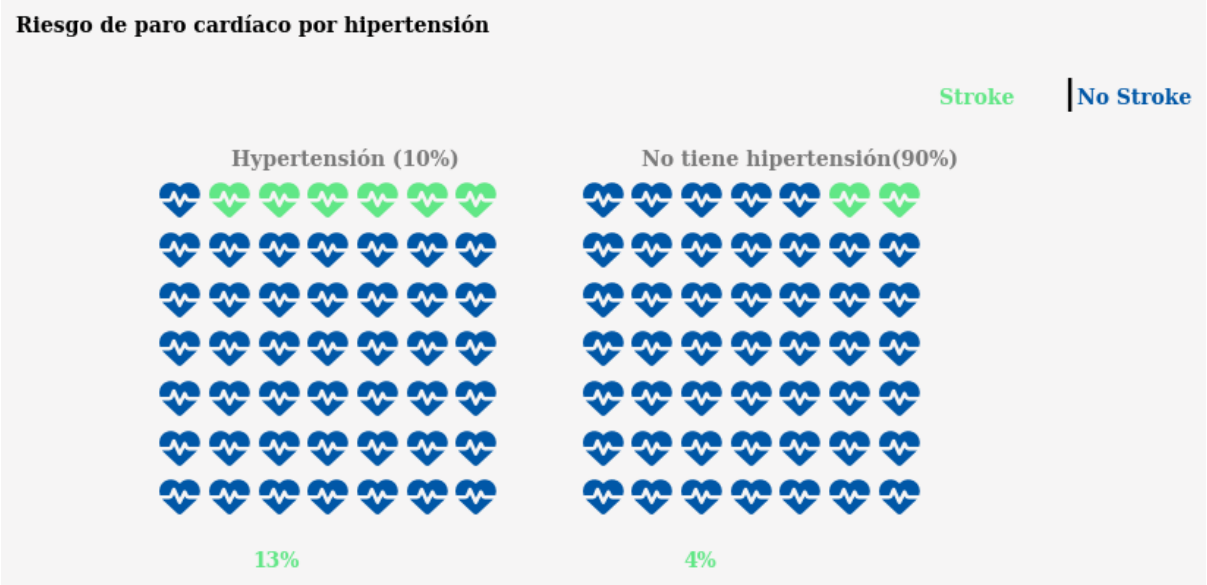


Figura 25: Comparación de probabilidades de paros cardíacos por hipertensión

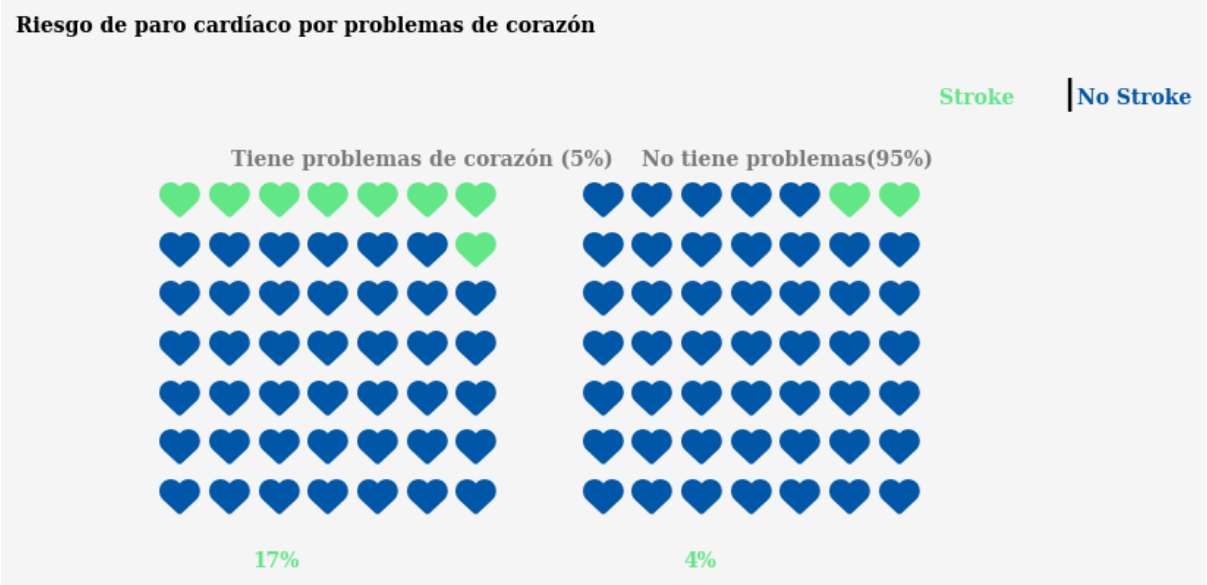


Figura 26: Comparación de probabilidades de paros cardíacos por problemas de corazón

La figura 27 es extremadamente interesante; ya que demuestra que estar casado triplica las probabilidades de sufrir un paro cardíaco. Esto probablemente se deba al estrés de estar casado y compartir tu vida con otra persona. Si la columna tuviera la variable de hijos, seguramente veríamos resultados parecidos.

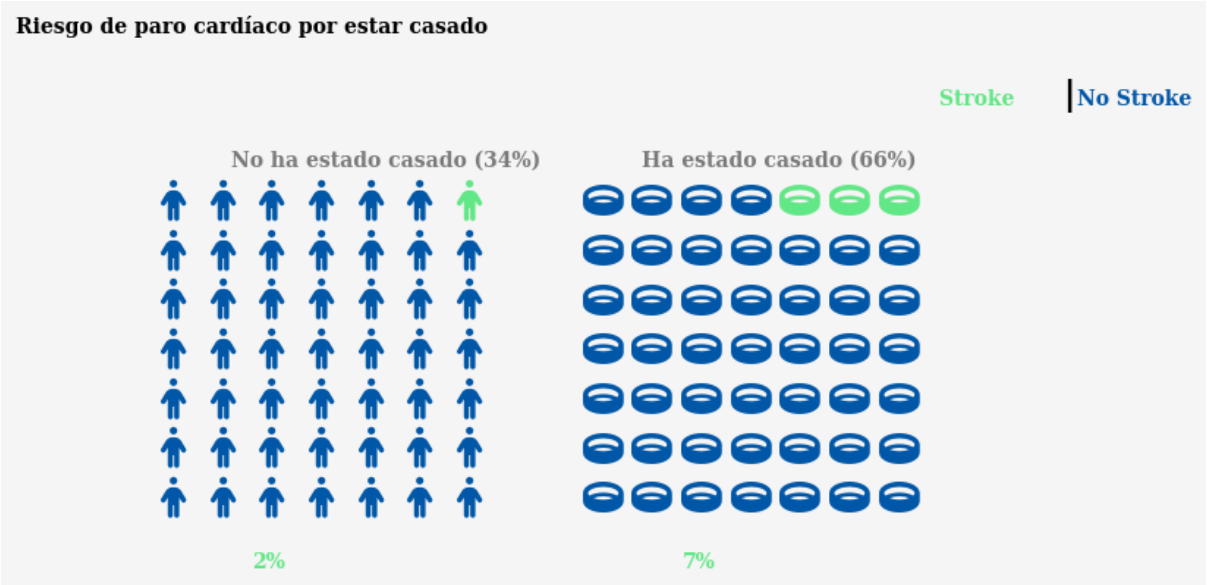


Figura 27: Comparación de probabilidades de paros cardíacos por estar casado

Al hacer todos estos análisis, es evidente que algunas de las variables analizadas sí aumentan las probabilidades de sufrir un paro cardíaco; sin embargo, no hay alguna variable que tenga una gran correlación con la categoría de interés (sufrir un paro cardíaco).

Resultados

Clustering jerárquico

El clustering jerárquico es un método de agrupamiento que se basa en la similitud de los datos para organizarlos en una estructura jerárquica, ayuda a entender y visualizar la similitud entre los objetos y la reacción entre los subconjuntos formados.

Primer clustering

En el primer clustering jerárquico se usaron todas las variables excepto la variable "stroke"; ya que esta es la categoría que se supone no conocemos.

	registros
0	3486
1	1623

	stroke
Cluster	
0	234
1	15

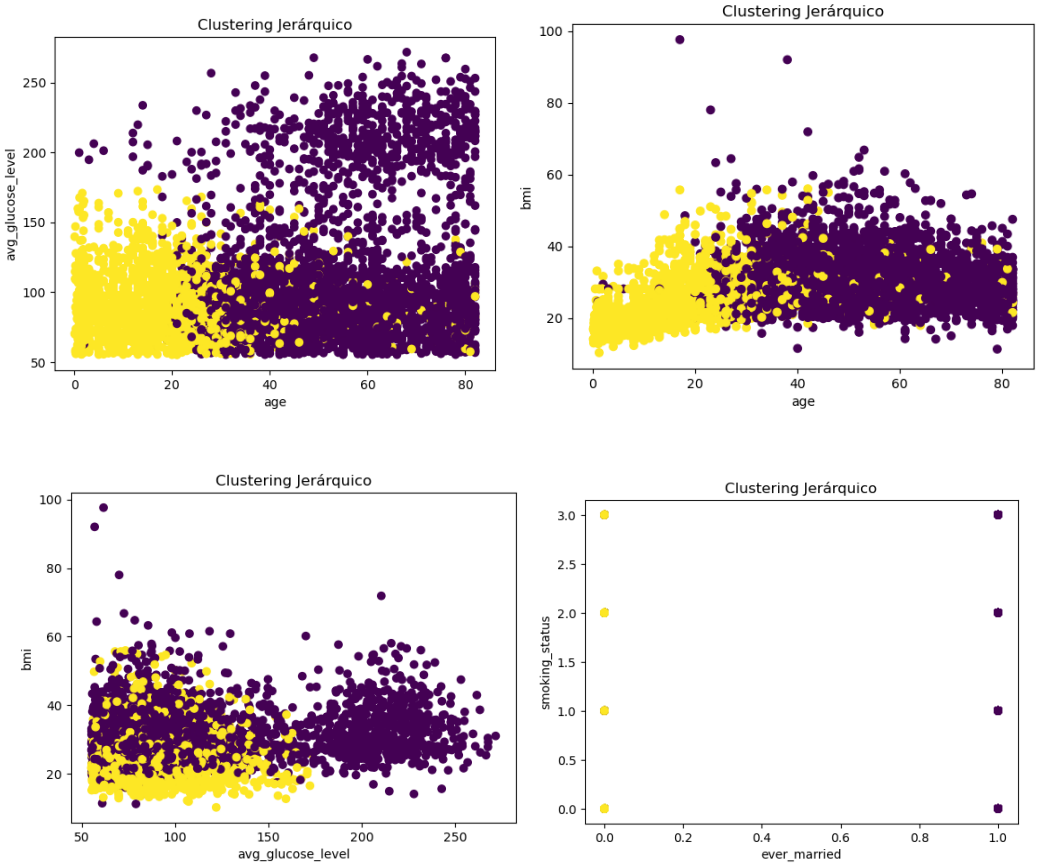


Figura 28: Gráficos del primer clustering jerárquico

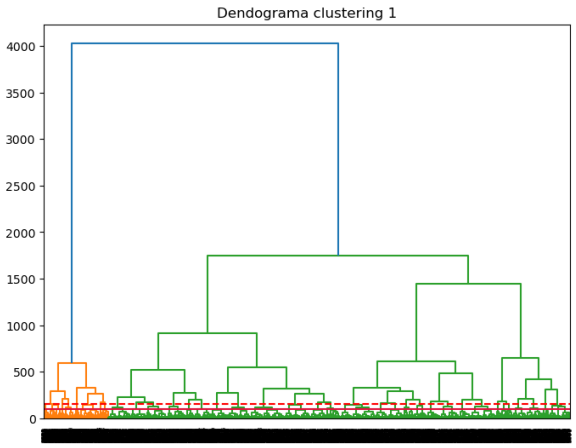


Figura 29: Dendograma del primer clustering jerárquico

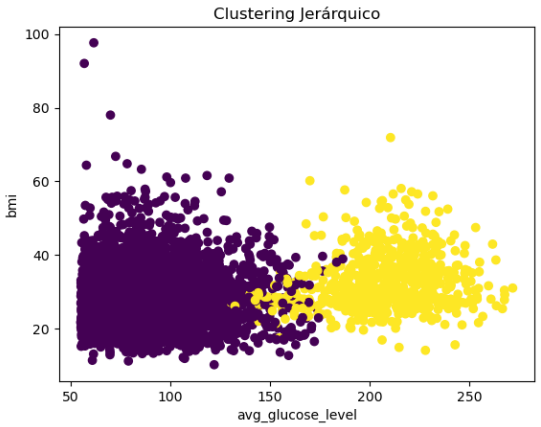
Este primer clustering solo logró una puntuación de silueta de 0.18. Esta puntuación indica que el clustering tiene cierto grado de estructura, pero no es muy fuerte. Es posible que algunos objetos estén bien agrupados y separados de otros, pero también puede haber superposiciones o puntos ambiguos que dificultan una agrupación clara.

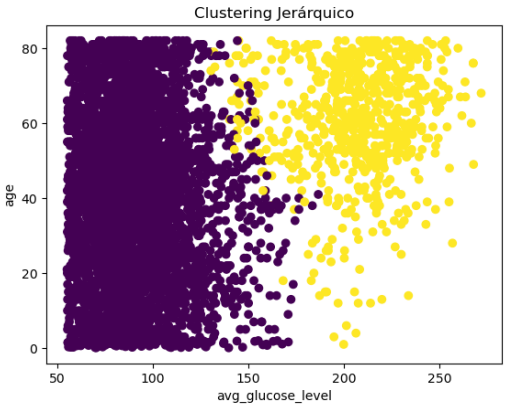
Segundo clustering

Como el resultado del primer clustering no fue muy prometedor, decidimos solamente usar las columnas cuantitativas para ver si eso mejoraba la agrupación.

	registros
0	4428
1	681

Cluster	stroke	probabilidad de sufrir un paro
0	157	0.035456
1	92	0.135095





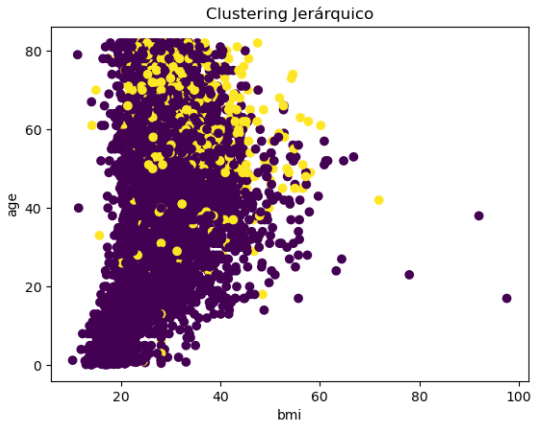


Figura 30: Gráficos del segundo clustering jerárquico (solo variables cuantitativas)

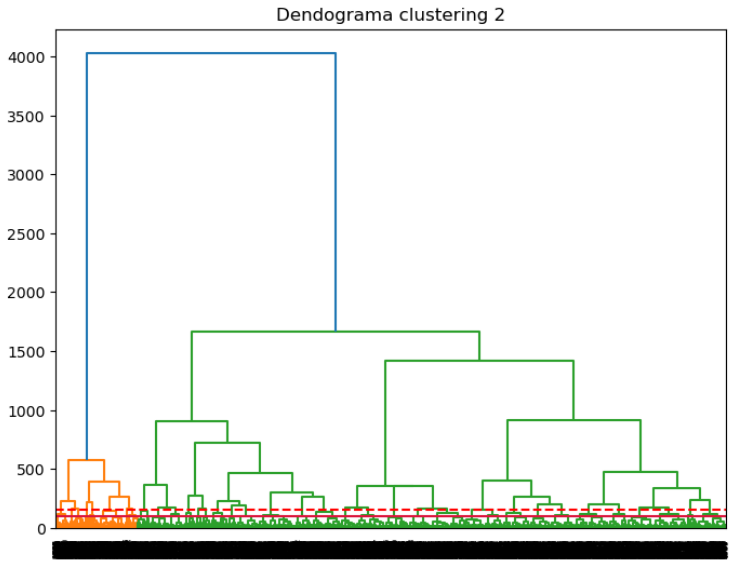


Figura 31: Dendograma del segundo clustering jerárquico

Este clustering logró una puntuación de silueta de 0.43, lo cual significa que los datos están bien agrupados. Sin embargo, es importante recalcar que aún puede haber cierta superposición o proximidad a la frontera entre los clusters.

Tercer cluster

Para el tercer y último cluster probamos la combinación de diferentes columnas hasta encontrar la que tuviera la puntuación de silueta más alta. Decidimos utilizar las columnas de “age”, “bmi”, “avg_glucose_level”, “heart_disease”.

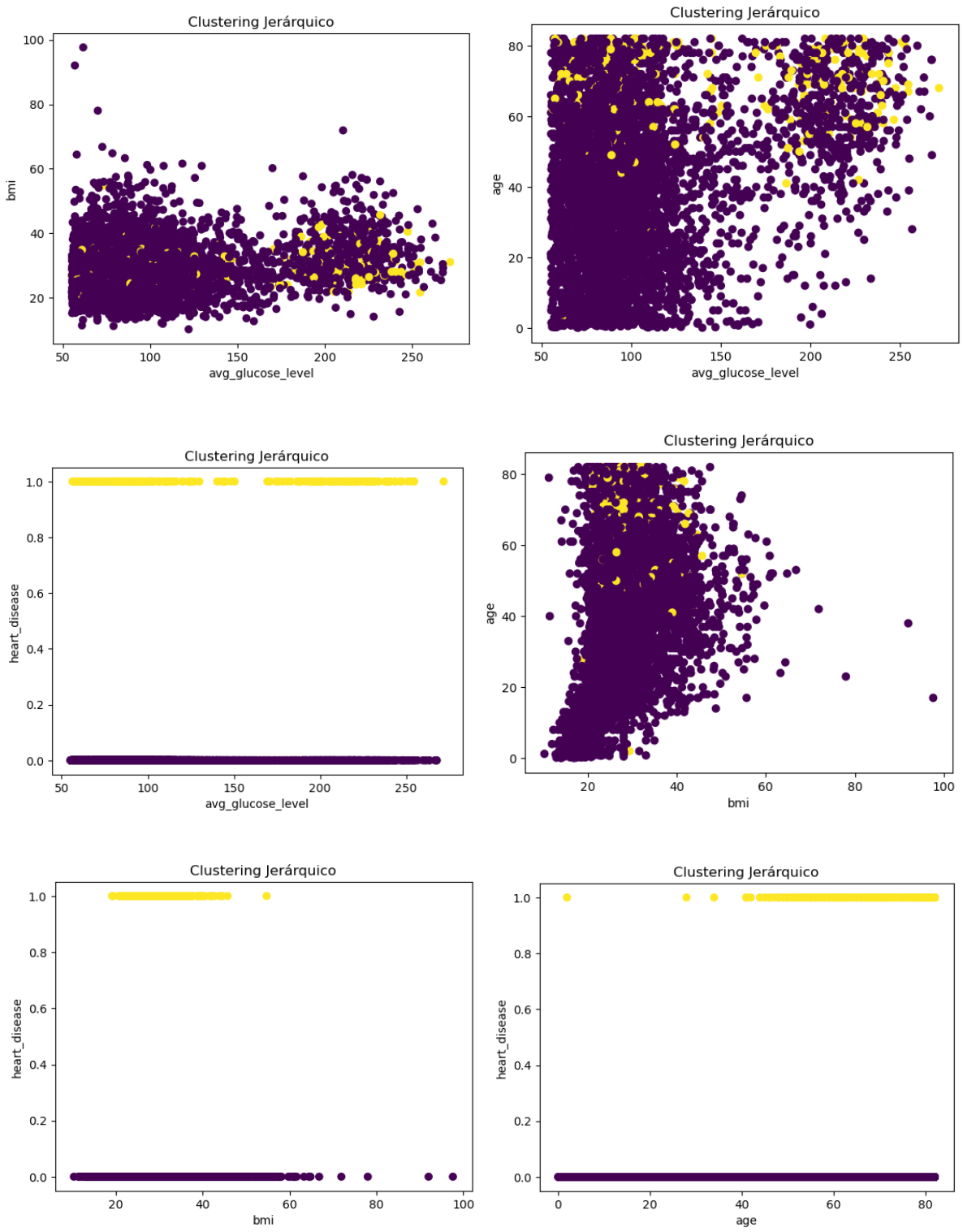


Figura 32: Gráficos del tercer clustering jerárquico

Esta agrupación logró una puntuación de silueta de 0.59; por ende, los objetos dentro de cada cluster están relativamente cerca unos de otros y están bien separados de los objetos en otros clusters.

Mezcla gaussiana

Un modelo de mezcla gaussiana es una distribución ensamblada a partir de distribuciones gaussianas multivariadas ponderadas. Los factores de ponderación asignan a cada distribución diferentes niveles de importancia. El modelo resultante es una superposición de curvas en forma de campana. Se pueden usar para detectar anomalías. (Benites, 2022)

Para determinar el número de componentes y el tipo de covarianza del modelo, evaluamos los valores BIC y escogemos el que tenga el menor valor.

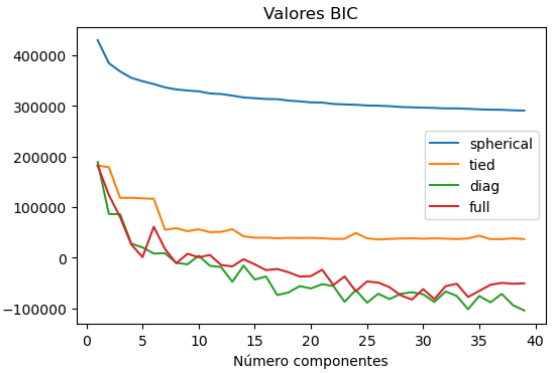


Figura 33: Gráfica de valores BIC de la mezcla gaussiana

```
GaussianMixture
GaussianMixture(covariance_type='diag', n_components=34, random_state=123)
```

Figura 34: Modelo de mezcla gaussiana

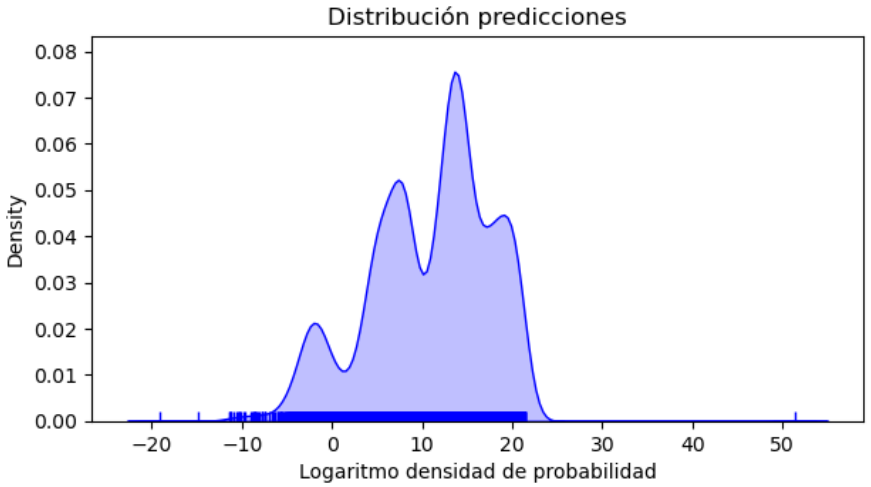
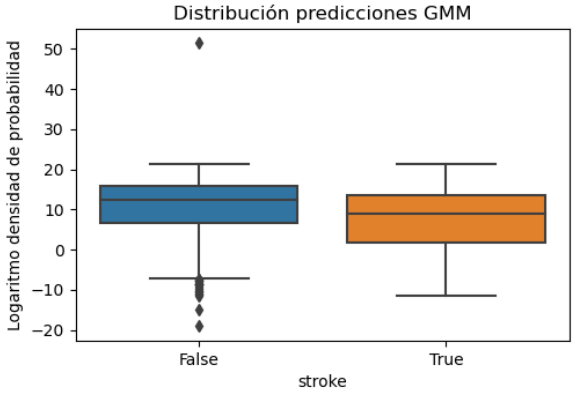


Figura 35: Gráfica de probabilidad relativa según la cual dicha variable aleatoria tomará determinado valor

Como se ve en la figura 35, la distribución del logaritmo de probabilidad en el grupo de paros cardíacos (positivos) es inferior. También podemos observar que el modelo es bueno identificando cuando una persona no tendrá un paro cardíaco, pero no es bueno identificando cuando una persona sí tendrá un paro cardíaco. Una de las principales limitaciones del uso de modelos GMM como detectores de anomalías es que consideran que los datos siguen distribuciones normales multivariante.



clasificacion	0	1
stroke		
False	4702	158
True	230	19

Probabilidad de falso positivo: 0.95 Probabilidad de falso negativo: 0.05

Probabilidad de verdadero negativo: 0.89 Probabilidad de verdadero positivo: 0.11

Figura 36: Gráfica de probabilidad relativa de cada clasificación de paro cardíaco y tabla de resultados

Como este modelo obtuvo una puntuación de silueta de -0.39, podemos concluir que los puntos dentro de cada cluster no están muy cerca unos de otros y que los puntos de

diferentes clusters están más cerca entre sí que los puntos dentro de los clusters. Es decir, no podemos confiar en estos resultados.

DBSCAN

Este tipo de algoritmos para clustering, a diferencia de los métodos más populares se encarga de formar clusters que son más difíciles de encontrar o tienen formas complejas. Este algoritmo está basado en la densidad de las regiones en el espacio, pues asume que ahí se encuentran los clusters y los puntos de ruido en las regiones de menor densidad(Kumar, 2021).

Para aplicar este modelo fue necesario realizar One Hot Encoding, ya que los datos en su mayoría eran booleanos o categóricos.

Las features utilizadas fueron ‘smoking_status’, 'ever_married', 'stroke', 'gender', 'glucose_category' y 'heart_disease'. Para los parámetros se utilizó 0.9 para el valor de epsilon y un mínimo de 30 muestras por feature.

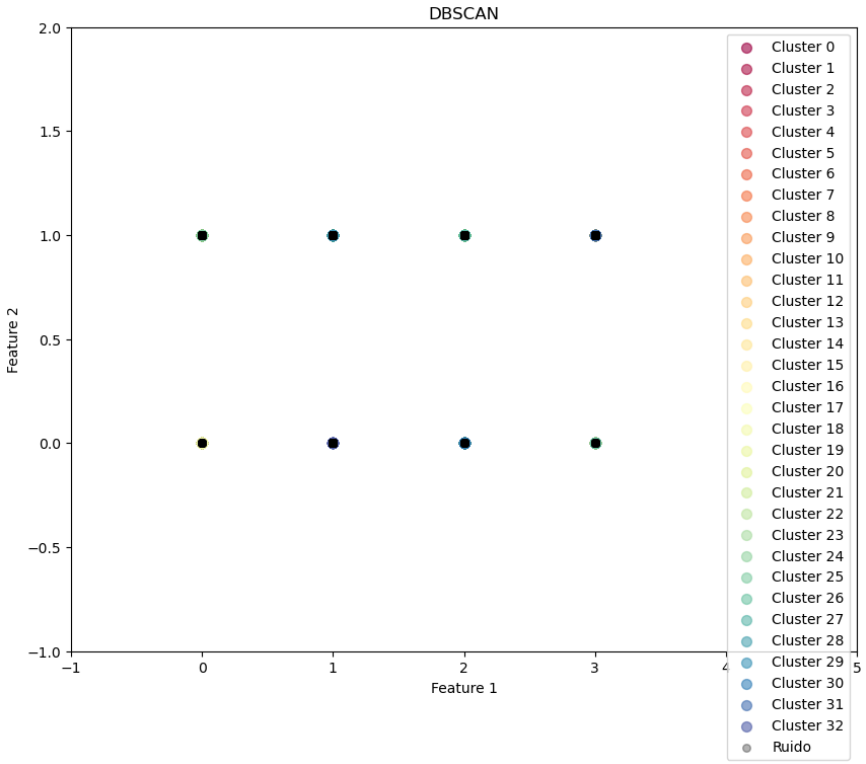


Figura 37: Gráfico de clustering DBSCAN

En total se encontraron 33 clusters. Sin embargo como se puede observar no hay agrupaciones claras de puntos que indiquen alguna relación significativa. Esto puede ser ya que los datos categóricos aumentan la dimensionalidad lo que no es recomendado para modelos de aprendizaje no supervisado (Kumar, 2021).

Este modelo obtuvo un coeficiente de silueta de -0.2167 lo que indica asignaciones imprecisas.

Comparación de modelos

Clustering jerárquico 1	Clustering jerárquico 2	Clustering jerárquico 3	Mezcla gaussiana	DBSCAN
0.18	0.43	0.59	- 0.39	- 0.2167

En la tabla podemos observar que evidentemente el mejor modelo es el tercer clustering jerárquico; el cual obtuvo un puntaje de silueta de 0.59. Este resultado indica que este algoritmo ha logrado una agrupación efectiva de los datos en comparación con los otros algoritmos mencionados.

Por otro lado, la "Mezcla gaussiana" tiene un valor de puntaje de silueta de -0.39, lo que indica una menor calidad de agrupación en comparación con los demás algoritmos. Un puntaje negativo sugiere que los puntos podrían estar asignados incorrectamente a los clústeres o que los clústeres tienen una superposición considerable.

Para mejorar estos modelos es fundamental explorar diferentes configuraciones y evaluar cómo afectan los resultados. Además, para futuros proyectos o mejoras de este proyecto sería recomendable aplicar una reducción de dimensionalidad mediante PCA y utilizar técnicas de validación cruzada para evaluar y comparar diferentes configuraciones y modelos. Esto permitirá seleccionar el mejor modelo y evitar el sobreajuste al evaluar el rendimiento en datos no vistos.

Conclusiones

En este reporte se han puesto en práctica modelos de aprendizaje no supervisado como K-means, clustering jerárquico y mezcla gaussiana para predecir paros cardíacos. Estos modelos son útiles para identificar patrones y subgrupos dentro de los datos de los pacientes, lo que puede ser valioso para la detección temprana y la predicción de eventos cardíacos.

Sin embargo, es importante considerar las consecuencias éticas asociadas con el uso de estos modelos en la vida real. En primer lugar, es fundamental garantizar la privacidad y la protección de los datos de los pacientes. La utilización de información médica sensible implica la necesidad de contar con medidas de seguridad y consentimiento adecuadas para asegurar que se respeten los derechos y la confidencialidad de los individuos. Además, la implementación de estos modelos en hospitales reales requiere que sean muy precisos y que sean validados para no tener predicciones o grupos de datos incorrectos. Como todos los modelos tienen un márgen de error, no podemos confiar completamente en los resultados que estos arrojan y se deben utilizar otras herramientas médicas para tomar decisiones porque en el área de salud si hay un error, corremos el riesgo de arruinar una vida humana. Otra preocupación ética importante es la posibilidad de bias (sesgos) y discriminación en la predicción de paros cardíacos. Los modelos de aprendizaje automático pueden verse afectados por sesgos inherentes en los datos de entrenamiento, lo que podría conducir a resultados injustos o discriminatorios, especialmente si se utilizan en diferentes poblaciones o grupos minoritarios. Es fundamental abordar y mitigar estos sesgos para garantizar la equidad en el diagnóstico y el tratamiento médico.

Reflexiones Individuales

Annette Pamela Ruiz Abreu

Considero que los modelos de aprendizaje no supervisado son muy útiles e interesantes cuando tenemos datos sin categorías; sin embargo, en el caso del dataset utilizado en esta entrega, no es muy útil la implementación de estos modelos y es más fácil utilizar métodos supervisados. Para realmente apreciar las funcionalidades de estos modelos podríamos utilizar una base de datos de un hospital e intentar encontrar patrones o grupos que nos ayuden a ofrecerles tratamientos más específicos. El modelo no supervisado que más me gusta es K Means porque se me hace el más fácil de interpretar. Algo que no me encanta de los modelos no supervisados es que no hay una "respuesta correcta"; es decir, estos modelos agrupan con base en medidas de similitud, pero no podemos saber inmediatamente qué significan estos grupos o si los agrupó correctamente. Otra consideración importante es que aunque es bueno tener muchos datos, no siempre se deben usar todos para entrenar al modelo. Finalmente, al no tener mucha experiencia con modelos de aprendizaje, es muy fácil que me equivoque al implementar los de aprendizaje no supervisados.

Leslie Ramos Gutierrez

Al trabajar con modelos de aprendizaje no supervisado note principalmente dos cosas, la primera es que la implementación de estos métodos tiende a ser bastante sencilla y se puede encontrar mucha información al respecto, aunque la interpretación suele ser más complicada y para nada "obvia" ya que si no entiendes lo que las gráficas y las diferentes métricas de evaluación te están diciendo, no hay manera fácil de entender si esta bien el modelo o si no hizo un buen trabajo; lo segundo que comprendí es que aunque suelen ser muy útiles, no se podrán implementar en todos los casos, por lo que aprender a usar modelos de aprendizaje supervisado es fundamental y necesario, ya que muchas de las bases de datos

existentes cuentan con variables categóricas, las cuales impiden una buena implementación de los modelos no supervisados.

Rodrigo González Zermeño

En esta etapa del proyecto se empezaron a implementar los métodos que hemos estado aprendiendo anteriormente en las clases, pero en esta etapa ayudó a ver como realmente se pueden llegar a implementar en diferentes escenarios, y cuál puede ser el mejor para implementar en diferentes ocasiones, dependiendo del objetivo que se tenga en un principio, aunque este puede llegar a cambiar después del análisis inicial que se hagan de los datos o de los modelos preliminares. Los datos nos pueden ayudar a buscar patrones en el área de salud y siento que es una gran oportunidad ya que se recolectan muchos datos y puede llegar a ser de mucha utilidad saber como se puede prevenir algunos casos de enfermedades, en este caso cardiovasculares.

Sarah Dorado

Durante este proyecto utilizamos varios tipos de machine learning, lo que me permitió entender mucho más como funciona cada uno y para qué casos es el más adecuado. Para nuestro dataset elegido en particular, no llegamos a muchas conclusiones solamente con los métodos de aprendizaje no supervisado. Existen varias razones para que esto suceda, por ejemplo.Intente programar el método ‘Density-Based Spatial Clustering of Applications with Noise’. Pero al tener tantos datos categóricos y booleanos existían demasiadas dimensiones lo que no es recomendado al aplicar algún método de clustering. Incluso intenté utilizar One hot encoding pero esto no mejoró el resultado. Es por esto que me doy cuenta que para este tipo de dataset es mucho más útil aplicar modelos supervisados. Ya que en los métodos que aplicamos como K-nearest neighbor y Logistic regression se obtuvieron resultados bastante buenos con un alto accuracy para las predicciones. Aunque los modelos de ML supervisados

y no supervisados no den los mismos resultados o permitan los mismos análisis, considero que para este dataset el tener modelos predictivos muy acertados ya nos permite llegar a muchas conclusiones. Además de crear una herramienta que permita predecir si un paciente puede sufrir de alguna condición según sus datos.

Referencias

Benites, L. (2022). Modelo de mezcla gaussiana: definición simple. *Statologos*. https://statologos.com/modelo-de-mezcla-gaussiana/

Bhuvanchennoju. (2021). 📣💡🎨 Data Storytelling 🎯AUC focus on 🩸strokes. *Kaggle*. https://www.kaggle.com/code/bhuvanchennoju/data-storytelling-auc-focus-on-strokes

May, L., & Menon, S. (2023). *Paro cardíaco repentino en personas jóvenes*. HealthyChildren.org. https://www.healthychildren.org/Spanish/health-issues/injuries-emergencies/sports-injuries/Paginas/Sudden-Cardiac-Death.aspx

OPS. (2021). *Las enfermedades del corazón siguen siendo la principal causa de muerte en las Américas*. OPS/OMS | Organización Panamericana de la Salud. https://www.paho.org/es/noticias/29-9-2021-enfermedades-corazon-siguen-siendo-principal-causa-muerte-americas

Kumar, V(2021). *Tutorial for DBSCAN Clustering in Python Sklearn.* *https://machinelearningknowledge.ai/tutorial-for-dbscan-clustering-in-python-sklearn/*