Predicción de cardiopatías mediante Machine Learning

Christian y Laura

Sumario

Introducción

Las enfermedades coronarias son una de las principales causas de muerte en los países occidentales. Según la organización mundial de la salud, se estima que en 2019 el 32% de las muertes a escala mundial fueron provocadas por enfermedades del corazón o cerebro-vasculares. Sin embargo, está ampliamente contrastado que, en muchas de estas afecciones, los factores de riesgo conductuales tienen una influencia crucial en el desarrollo o no de la enfermedad.

En este estudio nos centramos en un tipo de enfermedad coronaria en la que la prevención juega un rol particularmente importante: las **cardiopatías**. Las cardiopatías coronarias son una enfermedad que afecta a los vasos sanguíneos que irrigan al músculo cardíaco. Su prevención y detección en etapas tempranas es de vital importancia para evitar que los pacientes acaben desarrollando la enfermedad y salvar así millones de vidas. El uso del Machine Learning para la predicción de poblaciones de riesgo usando registros sanitarios puede ser particularmente importante para paliar el desarrollo de la enfermedad.

Planteamiento del problema

El objetivo de este proyecto es, por lo tanto, desarrollar modelos de Machine Learning que puedan predecir la ocurrencia de eventos de cardiopatías coronarias a partir de datos clínicos.

Recopilación de datos

La base de datos utilizada en este proyecto proviene de un estudio de enfermedades cardíacas publicado en Sudáfrica en 1983 por investigadores del National Research Institute for Nutritional Diseases of the South African Medical Research Council.

La base de datos utilizada contiene 462 observaciones médicas de hombres en una región con una alta prevalencia de cardiopatías.

Para cada sujeto, se determinó si había desarrollado enfermedad coronaria o no (la variable “chd” que estudiaremos en detalle más adelante), junto con toda una variedad de factores de riesgo conocidos en la época. Las características incluyen la presión arterial sistólica (sbp), el consumo acumulativo de tabaco (en kg), el colesterol LDL (ldl), la adiposidad (medida de obesidad similar al IMC), la obesidad, el historial de antecedentes familiares de enfermedades cardíacas (famhist), la clasificación por tipo A (una métrica propia elaborada por los artífices de este estudio), el consumo de alcohol y finalmente la edad.

El conjunto de datos utilizado en este estudio proviene de un repositorio de Kaggle que digitalizó estos datos para su explotación mediante técnicas de ciencia de datos.

Limpieza y preparación de datos

Los datos fueron limpiados y preprocesados de la siguiente manera:

- Se pasó el dataframe original, en formato de texto (.txt) a csv (cardiovascular.csv) para poder manipularlo
- Eliminamos la variable “ind” referente al identificador de registro por individuo al observar que no aporta ninguna información relevante para la elaboración de nuestro modelo de predicción
- Tras verificación, no se observan valores nulos ni outliers que se pudieran eliminar.

Análisis exploratorio de datos

Variable	Descripción	Tipo
Ind	Identificador de registro por individuo	Cuantitativa / Discreta / Razón
sbp	Presión sistólica (mm HG)	Cuantitativa / Discreta / Razón
tobacco	Consumo de tabaco acumulado (kg)	Cuantitativa / Continua / Razón
ldl	Colesterol lipoproteico de baja densidad	Cuantitativa / Continua / Razón
adiposity	Medición corregida de obesidad	Cuantitativa / Continua / Intervalo
famhist	Presencia de antecedentes familiares de cardiopatía	Cualitativa / Dicotómica / Nominal
typea	Clasificación conductual (elaborada por autores estudio)	Cuantitativa / Discreta / Razón
obesity	Medición de la obesidad (IMC)	Cuantitativa / Continua / Intervalo
alcohol	Consumo de alcohol anual (L)	Cuantitativa / Continua / Razón
age	Edad (años)	Cuantitativa / Discreta / Razón
chd	**Diagnóstico de enfermedad coronaria**	**Cualitativa / Dicotómica / Nominal**

Las observaciones clave del análisis inicial de esta base de datos, son los siguientes:

1. **Muestra reducida sin datos nulos:** Del análisis exploratorio de los datos disponibles se desprende que pese a disponer de una muestra reducida de 462 individuos, no se observan datos nulos en el conjunto. Las estadísticas cuantitativas del conjunto y la distribución del mismo pueden observarse en la figura 1 y 3.

2. **Variable objetivo "chd":** La variable "chd" que es una variable cualitativa dicotómica (ver tabla superior) identifica a los individuos que han desarrollado enfermedad coronaria: 1 si desarrollaron cardiopatía y 0 si no es el caso. Utilizaremos "chd", por lo tanto, como nuestra variable objetivo.

3. **Desbalanceo en la distribución de individuos con y sin chd:** La distribución de los datos muestra un cierto desequilibrio en la cantidad de individuos con y sin enfermedad coronaria ("chd" de ahora en adelante). Disponemos de 160 casos positivos frente a 302 individuos sanos (fig.2), Este desbalanceo deberá ser considerado en el momento de entrenar nuestros modelos, ya que podría afectar su capacidad para predecir correctamente los casos positivos.

4. **Diferencias en las distribuciones de edad, ldl, sbp, adiposidad y tabaco entre individuos con y sin chd:** Los boxplots revelan diferencias notables en las distribuciones de varias características entre los individuos con y sin chd. Las variables como la edad, el colesterol LDL, la adiposidad, el consumo de tabaco y la presión sistólica muestran diferencias significativas (fig.4). Estas diferencias indican que estas características podrían ser importantes para la predicción de chd.

5. **Sesgo hacia la derecha en la edad de los individuos con chd:** La edad de los individuos de la muestra abarca a sujetos de los 15 a los 64 años (en consonancia con la esperanza media de vida que en Sudáfrica en 1983 era de 58,2 años) y, como cabría esperar, se aprecia un sesgo de la distribución hacia la derecha en los individuos con chd (fig6.).

6. **Mayor incidencia en los individuos con antecedentes familiares:** En cuanto a la variable categórica famhist, la gráfica muestra claramente cómo los individuos con antecedentes familiares de cardiopatías tienen una mayor incidencia de chd en comparación con aquellos sin antecedentes familiares (fig.4). Este resultado sugiere que los antecedentes familiares son un factor de riesgo significativo para la chd.

7. **Correlación significativa entre las variables "obesity" y "adiposidad":** Del estudio de las correlaciones entre las variables se desprende que existe una correlación positiva significativa (superior al 0,7) entre adiposidad y obesidad (fig.5).

8. **No se puede eliminar los outliers:** ya que la mayoría correspondían a individuos que desarrollaban enfermedad y la muestra estaba desbalanceada.

	ind	sbp	tobacco	ldl	adiposity	typea	obesity	alcohol	age	chd
count	462.000000	462.000000	462.000000	462.000000	462.000000	462.000000	462.000000	462.000000	462.000000	462.000000
mean	231.935065	138.326840	3.635649	4.740325	25.406732	53.103896	26.044113	17.044394	42.816017	0.346320
std	133.938585	20.496317	4.593024	2.070909	7.780699	9.817534	4.213680	24.481059	14.608956	0.476313
min	1.000000	101.000000	0.000000	0.980000	6.740000	13.000000	14.700000	0.000000	15.000000	0.000000
25%	116.250000	124.000000	0.052500	3.282500	19.775000	47.000000	22.985000	0.510000	31.000000	0.000000
50%	231.500000	134.000000	2.000000	4.340000	26.115000	53.000000	25.805000	7.510000	45.000000	0.000000
75%	347.750000	148.000000	5.500000	5.790000	31.227500	60.000000	28.497500	23.892500	55.000000	1.000000
max	463.000000	218.000000	31.200000	15.330000	42.490000	78.000000	46.580000	147.190000	64.000000	1.000000

Fig. 1 Tabla de estadísticas cuantitativas

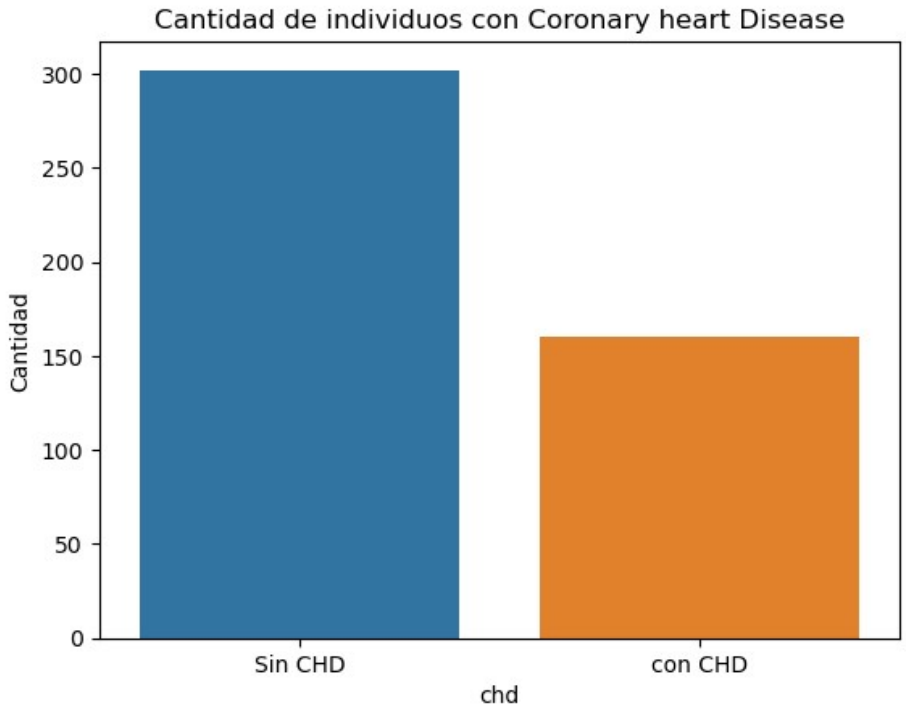

Fig. 2 Cantidad de individuos sin y con chd

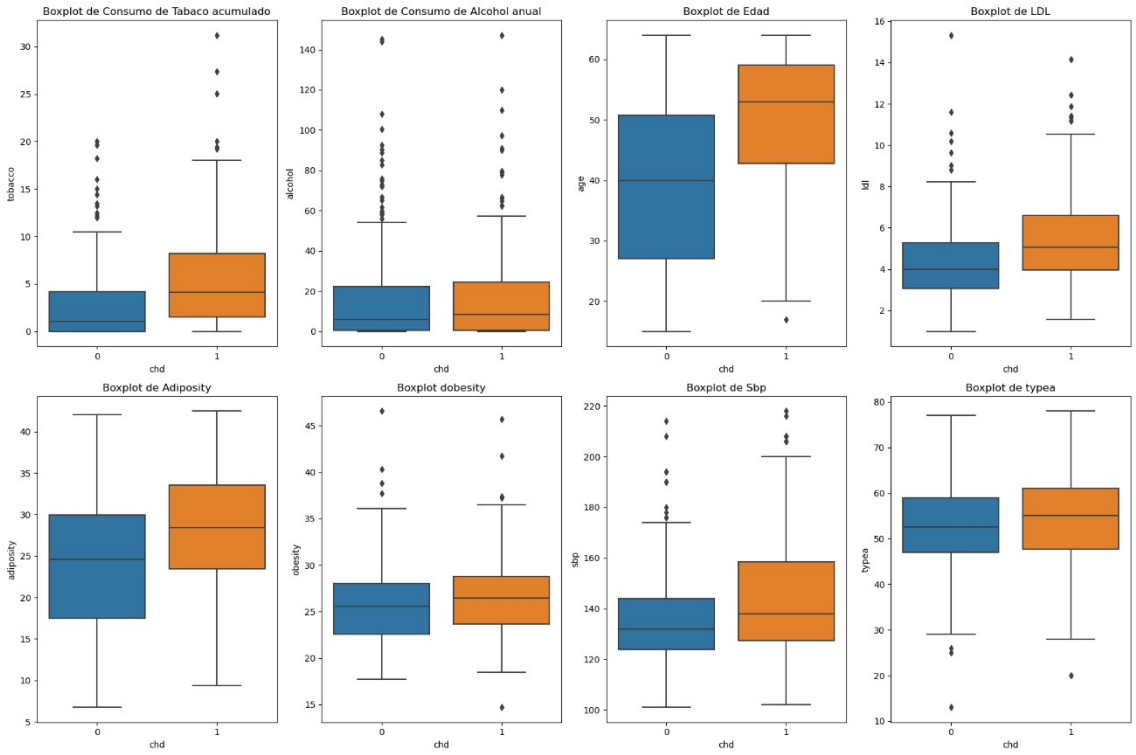

Fig. 3 Boxplot de variables numéricas

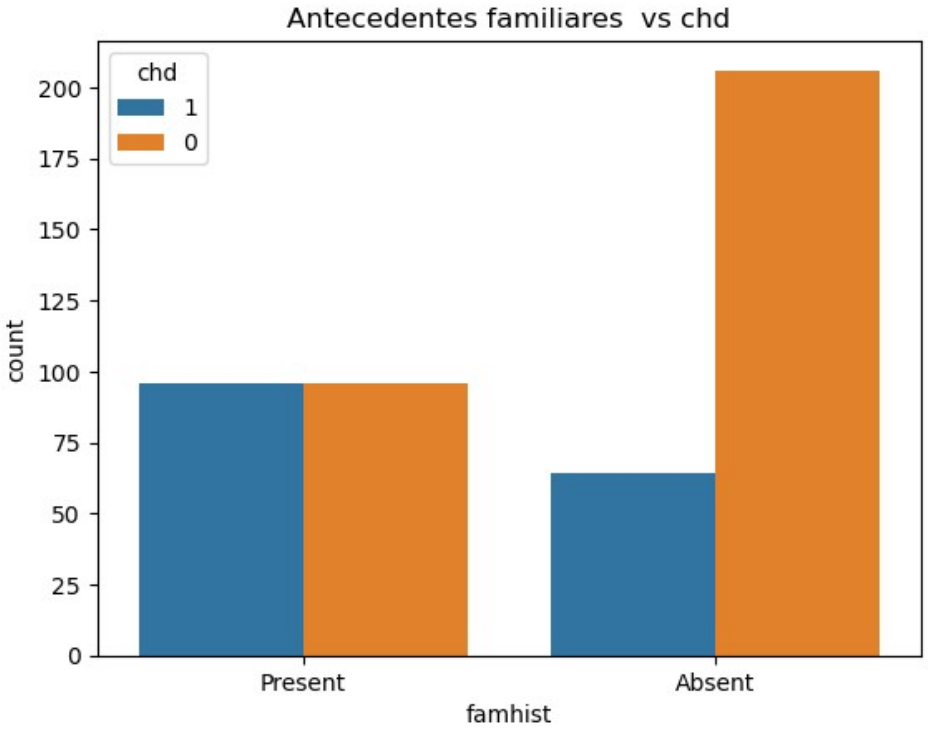

Fig. 4 Cantidad de individuos con antecedentes familiares de enfermedad coronaria sin y con chd

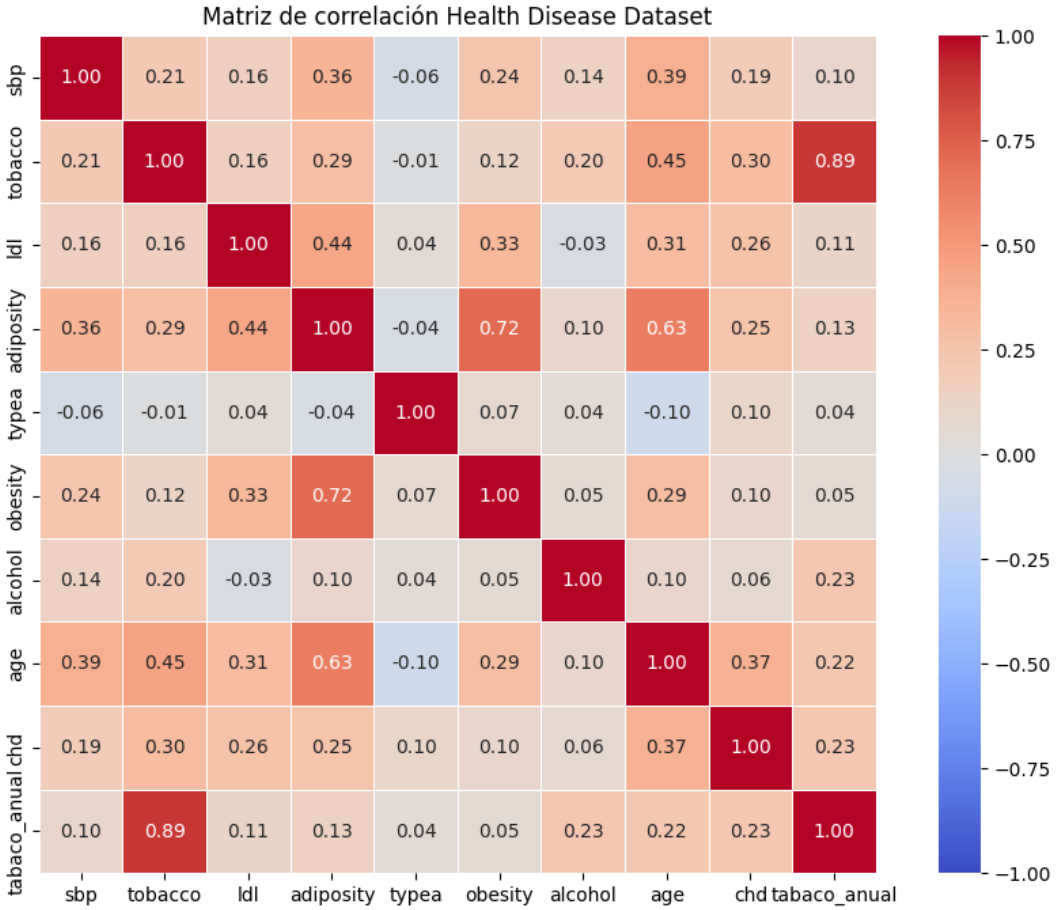

Fig. 5 Matriz de correlación de variables numéricas

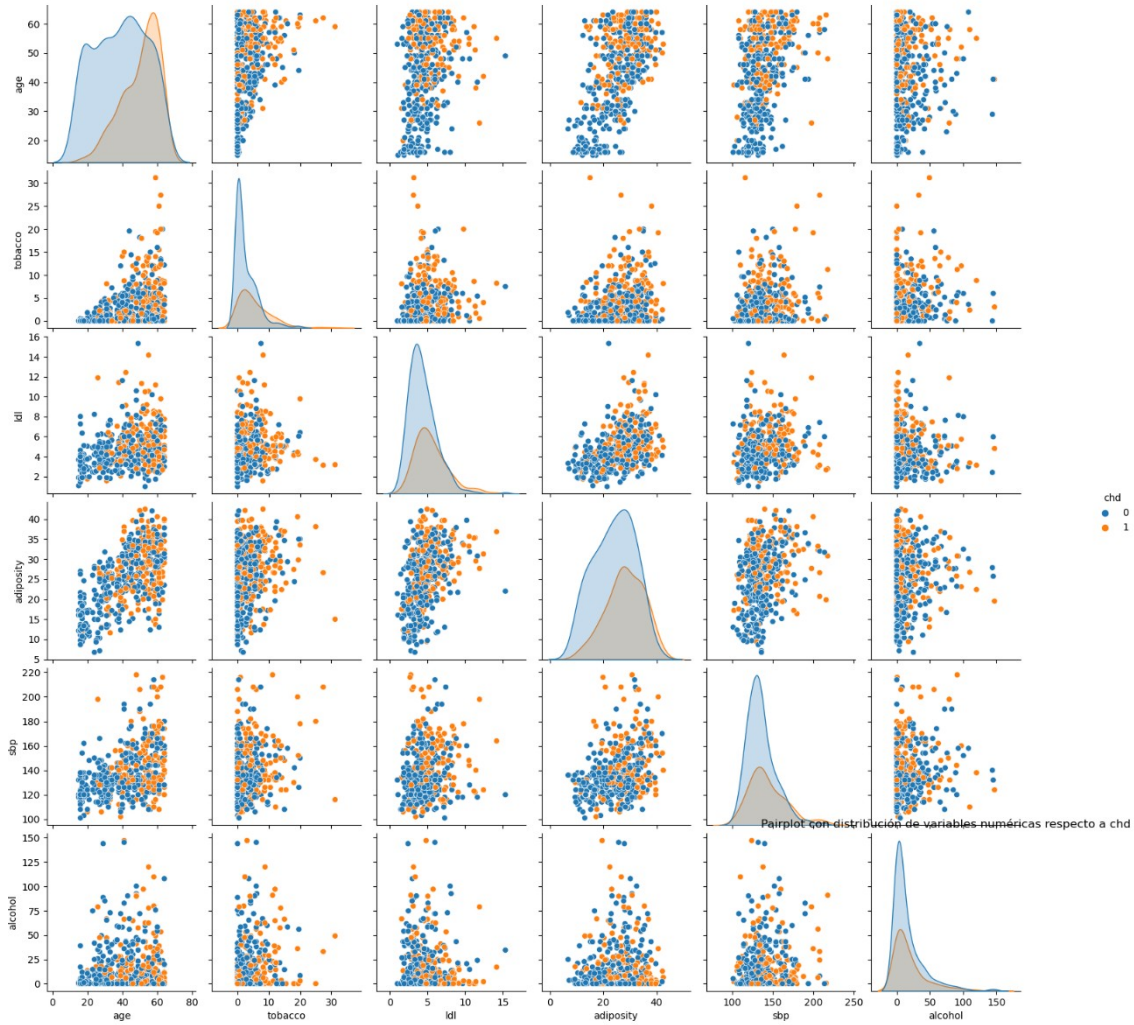

Fig. 6 Pairplot con distribución de variables numéricas respecto a chd

Ingeniería de características (Feature Engineering)

Para mejorar la calidad del modelo y la interpretabilidad de los datos, se realizaron las siguientes transformaciones y selecciones de características:

1. **Sustitución de la Variable "tobacco" por "tabaco_anual"** Se transformó la variable "tobacco" a "tabaco_anual" para evitar que la edad tenga un peso desproporcionado sobre esta variable y poder ponerla en la misma medida que "alcohol", que también se mide de forma anual. Para realizar este ajuste, se consideró que el consumo de tabaco acumulado debía ser dividido por los años fumando, definidos como la diferencia entre la edad y 15 años (pues no se registraron fumadores de 15 años o menos) Esta transformación asegura que el consumo de tabaco se mide de manera coherente con el consumo de alcohol anual, al tiempo que evita que tenga una correlación desproporcionada con respecto a chd.

2. **Eliminación de "obesity" para evitar la multicolinealidad:** La variable "obesity" fue eliminada porque mostró una alta correlación con "adiposity" en la matriz de

correlación, lo cual la hace redundante y podría añadir ruido a nuestro modelo predictivo.

3. **Transformación a boleano de "famhist":** Se utilizó GetDummies para transformar la variable famhist que contenía originalmente los valores "Present" en caso de presencia de antecedentes familiares de cardiopatía y "Absent" en caso contrario, a True y False, para poder utilizar la misma en nuestros modelos.
4. **Creación de Variables "over_median"** Se crearon variables "over_median" para las características más correlacionadas: sbp, ldl, tabaco_anual, adiposity y age. Se utilizó la mediana como umbral para evitar que el sesgo de las distribuciones afecte al conjunto. Estas variables se usarán para la creación de subgrupos de clasificación.
5. **Identificación de variables más correlacionadas con chd:** Se utilizó el coeficiente de correlación de Pearson para identificar las variables más correlacionadas con chd. El resultado es el siguiente**:**

Variable	**Coef. correlación**
chd (objetivo)	**1.000000**
age	0.372973
tobacco*	0.299718
famhist	0.272373
ldl	0.263053
adiposity	0.254121
tabaco_anual	0.228753
sbp	0.192354
typea	0.103156
Obesity*	0.100095
alcohol	0.062531

*Variables desechadas en el proceso de *feature engineering*

Las variables más correlacionadas son "age", "tobacco", "famhist", "ldl", "adiposity", "tabaco_anual" y "sbp". Sin embargo, como las correlaciones son similares y nuestra muestra reducida y desbalanceada, para entrenar nuestros modelos utilizaremos todas las variables menos obesity, que tiene una alta correlación con adiposity y tobacco, que hemos sustituido por la variable reconstruida "tabaco_anual"

Construcción y evaluación de modelos

Se desarrollaron y evaluaron tres modelos de clasificación:

- Regresión Logística (RL)
- Máquinas de soporte Vectorial (SVC)
- Redes Neuronales Artificiales (ANN)

Antes de entrenar los modelos, se aplicaron varias técnicas de preprocesamiento de datos para paliar el problema del desbalanceo entre variables y las diferencias de escala.

Se utilizaron técnicas de escalado (**StandardScaler y MinMaxScaler**) para normalizar las características numéricas, asegurando que todas estuvieran en la misma escala y evitando así que algunas variables predominen sobre otras debido a sus diferentes unidades de medida.

Además, se aplicó la técnica de sobremuestreo **SMOTE** para manejar el problema del desbalance de clases. Por el mismo motivo, se utilizó **validación cruzada** de 10 folds para evaluar el rendimiento de los modelos.

A continuación, se utilizó el método de búsqueda en cuadrícula (**GridSearch**) para ajustar los parámetros óptimos en Regresión Logística y SVM. Y para Redes Neuronales, se probaron diferentes configuraciones de **capas, dropouts, EarlyStopping y parámetros de activación** de las neuronas (relu, sigmoid y hard-sigmoid).

Las métricas utilizadas para la evaluación incluyeron:

- Exactitud (Accuracy)
- Precisión para la clase 0 (Precision 0)
- Recall para la clase 0 (Recall 0)
- Precisión para la clase 1 (Precision 1)
- Recall para la clase 1 (Recall 1)
- F1 Score para la clase 0 (F1 Score 0)
- F1 Score para la clase 1 (F1 Score 1)

Se probaron los modelos realizando diferentes combinaciones de features y ajustes de los mismos, para tratar de obtener una predicción lo más precisa posible.

1. **Todas las variables:** En un primer momento se incluyeron todas las variables del conjunto de datos
2. **9 variables seleccionadas:** Se entrenó el modelo con las 9 variables seleccionadas en el proceso de feature engineering
3. **3 variables más correlacionadas con chd:** Se utilizaron las 3 variables con mayor correlación con la variable objetivo (age, adiposity, ldl)
4. **Creación de submodelos:** Se crearon agrupaciones utilizando variables con valores por encima de la mediana para mejorar la precisión de los modelos.

Tras comprobar que los submodelos con algunas de las variables funcionaban con mayor precisión que los generales (en particular, para con las variables “ldl”, “sbp”, “famhist”, “adiposity”), se decidió utilizar el algoritmo de agrupación no supervisado **Kmeans** para la segmentación previa de poblaciones.

De la misma forma se probaron diferentes combinaciones para encontrar la óptima. Para la determinación del número de clusters óptimo se utilizó el método del codo (Elbow method):

1. **4 Clusters con 9 variables**: Los datos se agruparon en 4 clusters basados en las 9 variables seleccionadas: Un primer cluster “0” que agrupa a individuos con baja probabilidad de desarrollar enfermedad coronaria (10,1%) y el resto de individuos agrupados en otros tres clusters (“1”, “2”, “3”).

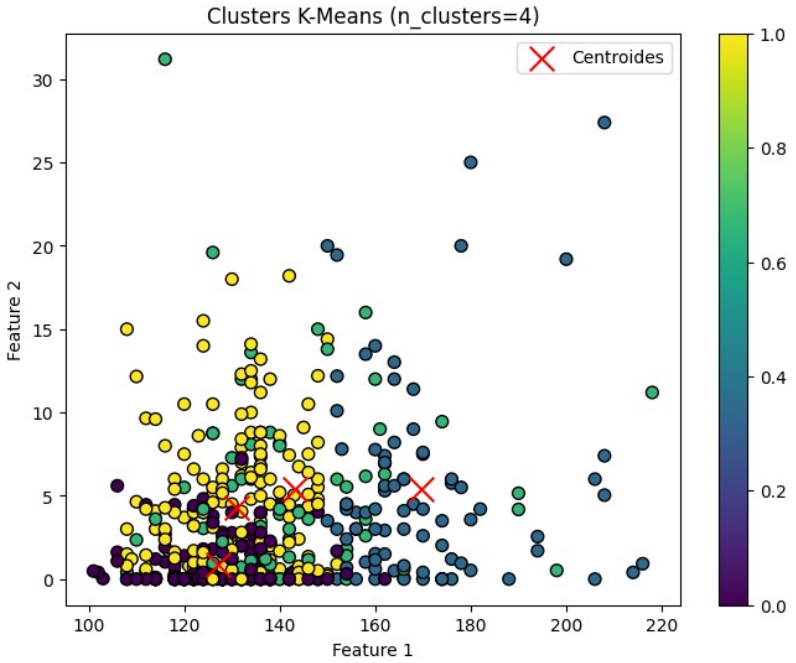

Fig. 7 Gráfico de dispersión para KMeans de 4 clusters

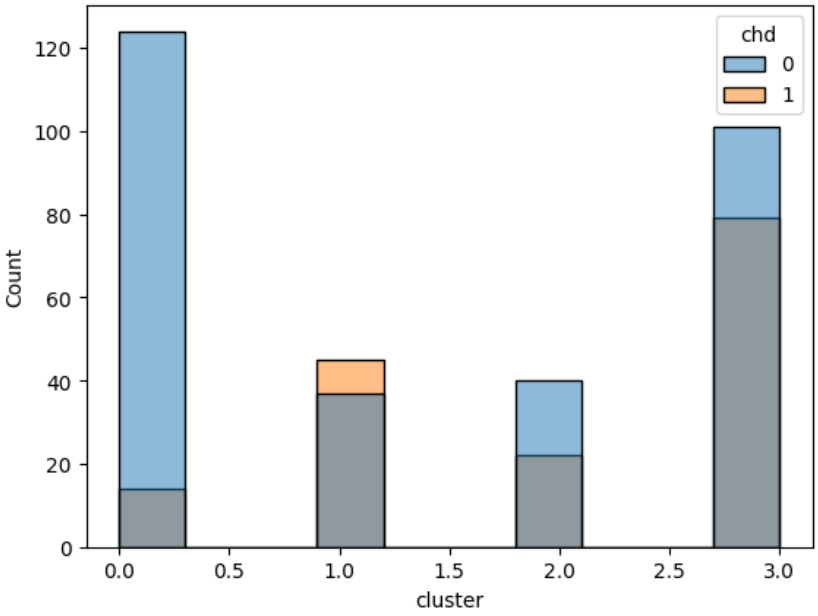

Fig. 7.1 Histograma para KMeans de 4 clusters

2. **3 Clusters con 7 variables**: En este caso, se utilizaron las 7 variables más correlacionadas y se obtuvieron 3 clusters: Un primer cluster “2” que agrupa a individuos con baja probabilidad de desarrollar enfermedad coronaria (13,7%) y una combinación de los otros dos clusters (“0” y “1”).

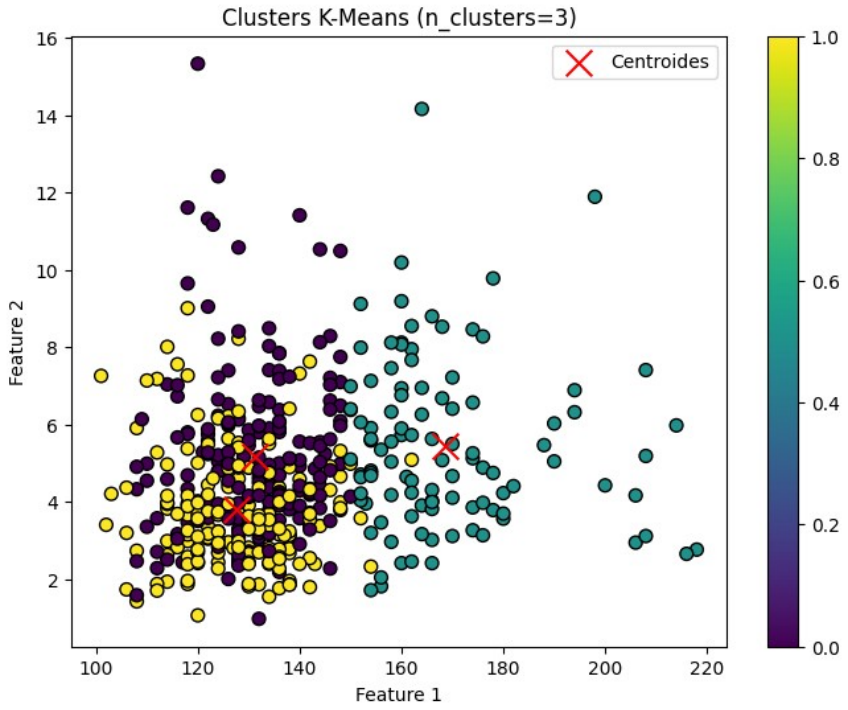

Fig. 8 Gráfico de dispersión para KMeans de 3 clusters

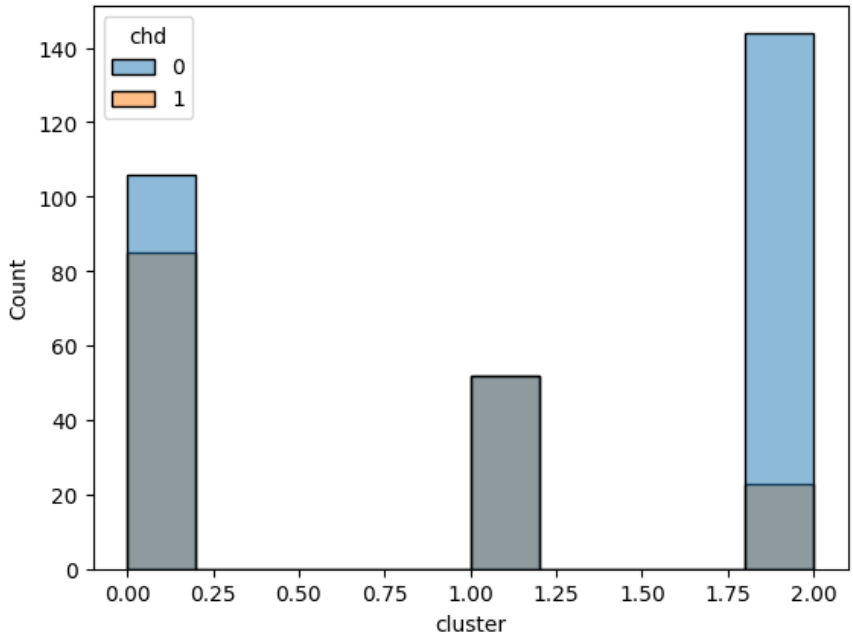

Fig. 8.1 Histograma para KMeans de 4 clusters

Resultados y discusión

La segmentación previa mediante Kmeans permitió mejorar la precisión de los modelos al identificar subgrupos con características afines dentro de la población de estudio.

De entre los subgrupos analizados se observa que la segmentación previa en tres grupos, uno con baja probabilidad de desarrollar “chd” y la aplicación del modelo de Redes Neuronales para la combinación de los otros dos grupos, resulta el más efectivo para la predicción de cardiopatías utilizando datos clínicos.

Cabe mencionar que la técnica de sobremuestreo SMOTE mejoró significativamente las métricas de rendimiento, dado el desbalanceo de la muestra utilizada.

Modelos	Accuracy	F1 Score (0)	F1 Score (1)
RL	0,66	0,71	0,6
SVM	0,64	0,66	0,63
ANN	**0,71**	**0,74**	**0,68**

fig. 9 Métricas predicción modelo final con segmentación por clusters

Conclusión

Los resultados muestran que pese a tratar con una base de datos limitada y desbalanceada, mediante técnicas de Machine Learning se consiguió elaborar un modelo de predicción de cardiopatías, ciertamente perfectible, pero que se podría seguir desarrollando para elaborar un modelo funcional de detección temprana.

Trabajo futuro

Dados los resultados, se sugiere explorar dos posibles vías de desarrollo de este modelo predictivo:

1. **Segmentación para métricas por encima de la mediana**: Para poblaciones con valores por encima de la mediana en las métricas “ldl”, “sbp”, “famhist”, “adiposity” se podría aplicar un modelo específico para cada una de estas variables, ya que, como se ha visto en los modelos elaborados previamente, dan buenos resultados. Y para el resto de poblaciones aplicar Kmeans.

2. **Ampliar el conjunto muestral y explorar parámetros**: Convendría probar el modelo propuesto a muestras más amplias para aumentar su precisión, así como ampliar el testeo con diferentes parámetros y estructuras de redes neuronales para mejorar todavía más la precisión del modelo.

Licencia

Creative Commons

Bibliografía

- Data:

https://www.kaggle.com/datasets/yassinehamdaoui1/cardiovascular-disease/data

- Documentación:

https://journals.co.za/doi/pdf/10.10520/AJA20785135_9894 # (Paper del estudio)

https://great-northern-diver.github.io/loon.data/reference/SAheart.html (Contexto y Descripción variables)

https://www.who.int/es/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)#:~:text=Las%20enfermedades%20cardiovasculares%20(ECV)%20son,las%20muertes%20a%20escala%20mundial

https://www.health.gov.za/wp-content/uploads/2022/05/Global-Adult-Tobacco-Survey-GATS-SA_FS-Populated__28-April-2022.pdf

https://apps.who.int/healthinfo/systems/surveydata/index.php/catalog/71

https://datosmacro.expansion.com/demografia/esperanza-vida/sudafrica

https://data.who.int/es/countries/710

Anexos:

- Código EDA
- Código Modelos
- Tabla de seguimiento de pruebas de modelos