Transformaciones y Clasificacion KNN

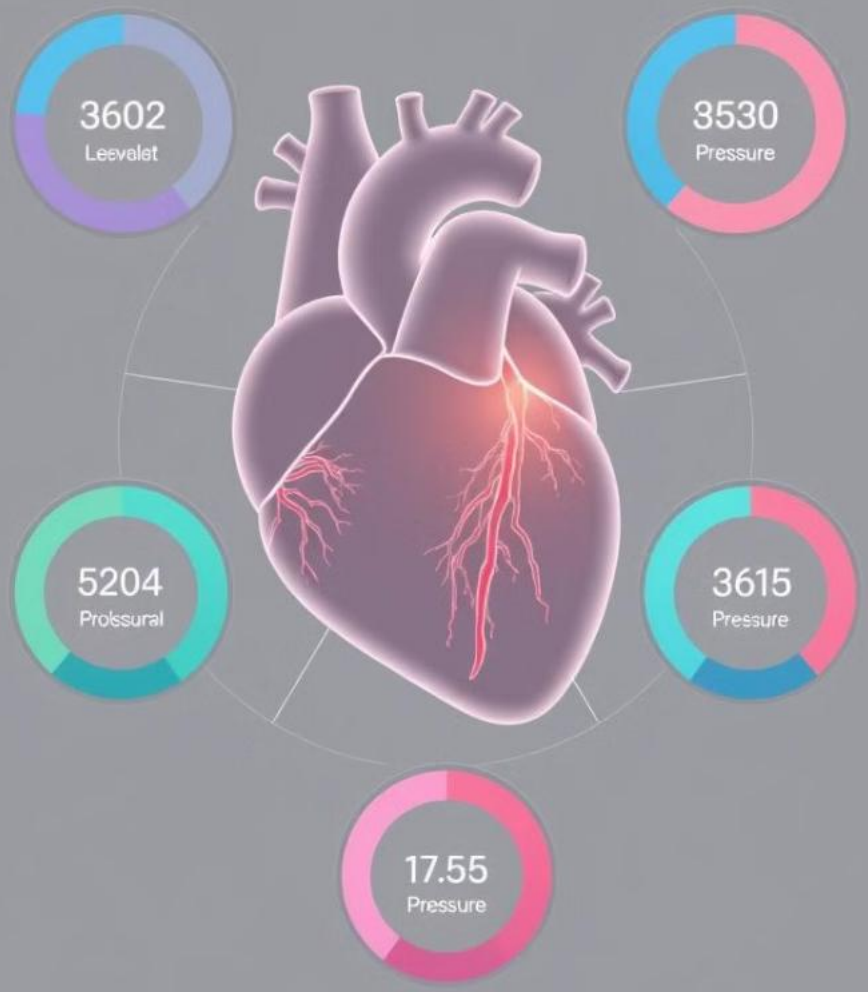
Universidad Politécnica Salesiana

Michael Israel Lata Zambrano

Computación

Inteligencia Artificial

Abril, 2025



Heart Disease Dataset

Fuente: UCI Machine Learning Repository

Enlace:

https://archive.ics.uci.edu/dataset/45/heart+diseaseq

Numero de variables: 14
14 atributos.

Observaciones : 303 registros.

registros.

Variable objetivo (TARGET): Predicción de la enfermedad cardiaca.

Descripción del Dataset

Tiene como finalidad de evaluar la predicción de la enfermedad cardiaca. Está compuesta por

Número de Variables: 14
Número de instancias(Pacientes): 303
Variable de estudio (Objetivo): TARGET
Sin enfermedad:1
Con enfermedad:2

	Id	Variable	Descripción breve	Tipo	Técnica de transformación a aplicar
0	1	edad	Años de edad	Numérica discreta	Estandarización
1	2	sexo	M (masculino) o F (femenino)	Categórica nominal	OneHotEncoding
2	3	dolor_toracico	Tipo de dolor torácico	Categórica ordinal	OrdinalEncoder
3	4	colesterol	Nivel de colesterol en sangre	Numérica continua	Estandarización
4	5	frecuencia_cardiaca_maxima	Frecuencia cardíaca máxima alcanzada	Numérica continua	Estandarización

Tabla de Diseño de Transformaciones

Id	Variable	Descripción breve	Tipo	Transformación
1	edad	Años de edad	Numérica discreta	Estandarización
2	sexo	M (0: mujer, 1: hombre)	Categórica binaria	OneHotEncoding
3	dolor_toracico	Tipo de dolor torácico	Categórica ordinal	OrdinalEncoder
4	presion_arterial	Presión arterial en reposo	Numérica continua	Estandarización
5	colesterol	Nivel de colesterol en sangre	Numérica continua	Estandarización
6	azucar_sangre	Azúcar en sangre (1: > 120 mg/dl)	Categórica binaria	OneHotEncoding
7	electrocardiograma	Resultados de electrocardiograma	Categórica nominal	OneHotEncoding
8	frecuencia_cardiaca_maxima	Frecuencia cardíaca máxima alcanzada	Numérica continua	Estandarización
9	oldpeak	Depresión ST inducida por ejercicio	Numérica continua	Estandarización
10	pendiente_ST	Pendiente del segmento ST en ejercicio	Categórica ordinal	OrdinalEncoder
11	cantidad_vasos	Cantidad de vasos coloreados	Numérica discreta	Estandarización
12	thal	Resultado de thal (3: normal, 6: defecto fijo, 7: defecto reversible)	Categórica nominal	OneHotEncoding
13	feature_13	Variable adicional 13	Desconocida	Estandarización
14	feature_14	Variable adicional 14	Desconocida	Estandarización
15	feature_15	Variable adicional 15	Desconocida	Estandarización

La tabla resume las características del dataset de **Heart Disease**, detallando las variables, sus tipos, y las técnicas de transformación necesarias para la preparación del modelo.

- Variables numéricas (como edad y colesterol): requieren estandarización para normalizar las escalas.
- Categóricas binarias (como sexo y azúcar en sangre): se transforman con OneHotEncoding para convertirlas en valores. compatibles con modelos de machine learning.
- Ordinales (como dolor torácico y pendiente ST): utilizan OrdinalEncoder para reflejar su jerarquíaVariables .
- Desconocidas: se incluyen como estimaciones técnicas, aplicando transformaciones genéricas hasta confirmar su impacto.

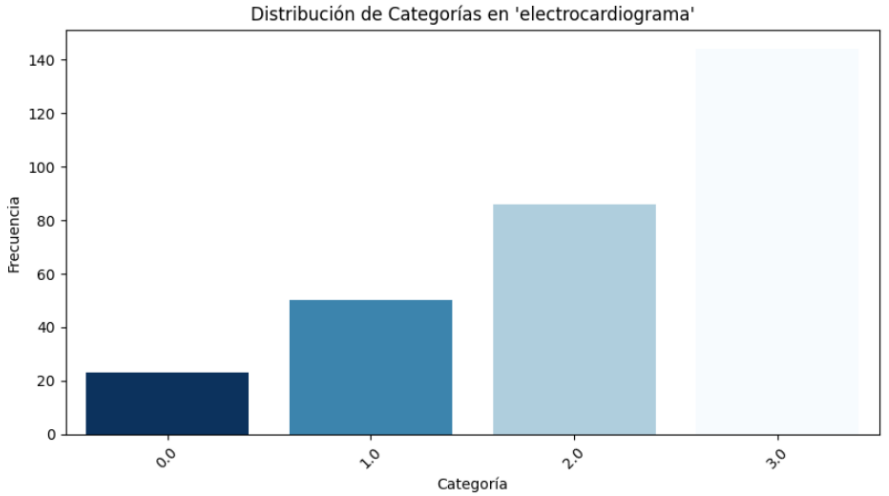
Analisis de Frecuencia Categorica

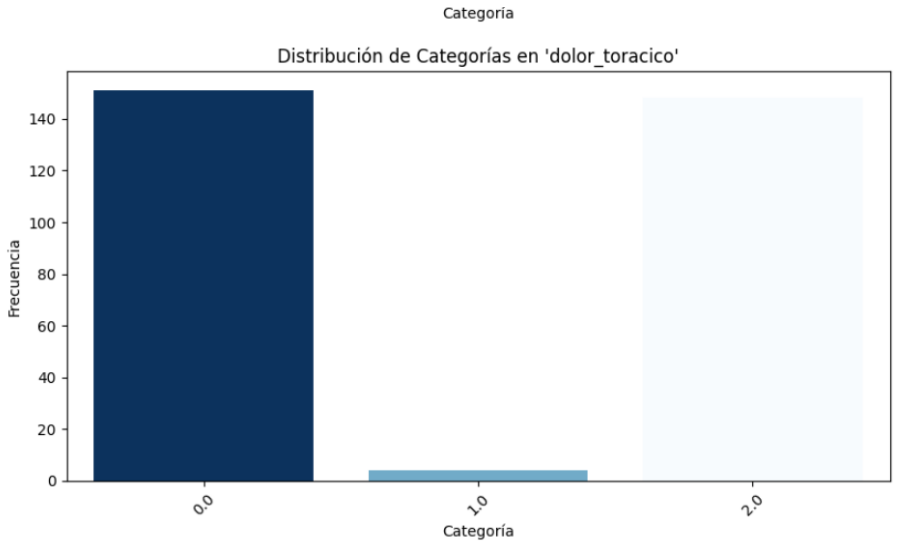
La imagen presenta tres gráficos de barras que muestran la distribución de categorías en variables categóricas del dataset de **Heart Disease**:

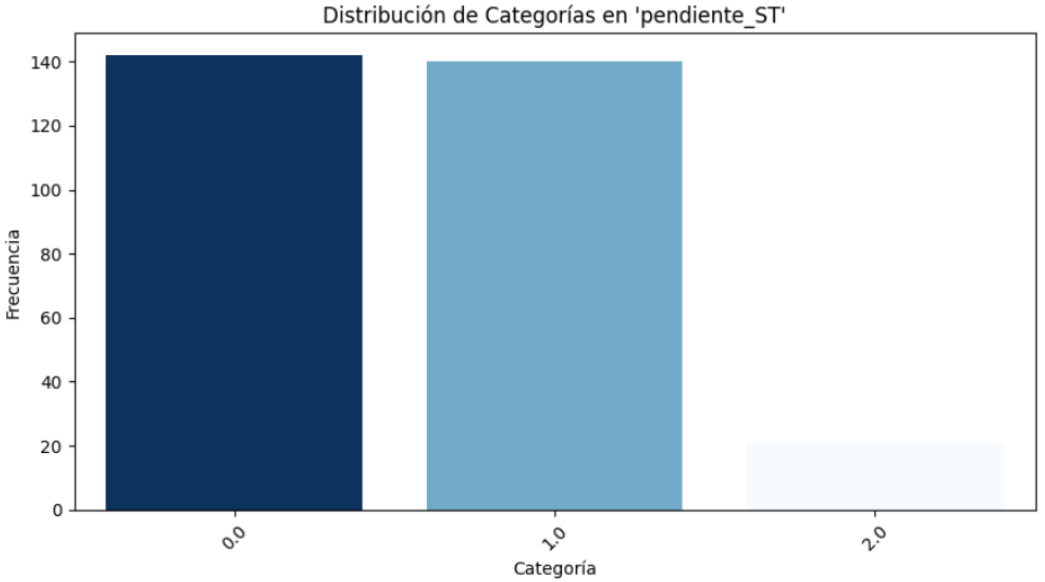
•Electrocardiograma: Se observa una distribución desigual, destacando la categoría 3.0 como la más frecuente, lo que sugiere un patrón recurrente en resultados cardíacos normales o alterados.

•Dolor torácico: La categoría 0.0 domina ampliamente, indicando que la mayoría de los pacientes no presentan dolor significativo al momento del análisis

•Pendiente ST: Las categorías 0.0 y 1.0 tienen alta frecuencia, mientras que la 2.0 aparece marginalmente, reflejando diferencias en la respuesta del corazón al esfuerzo.

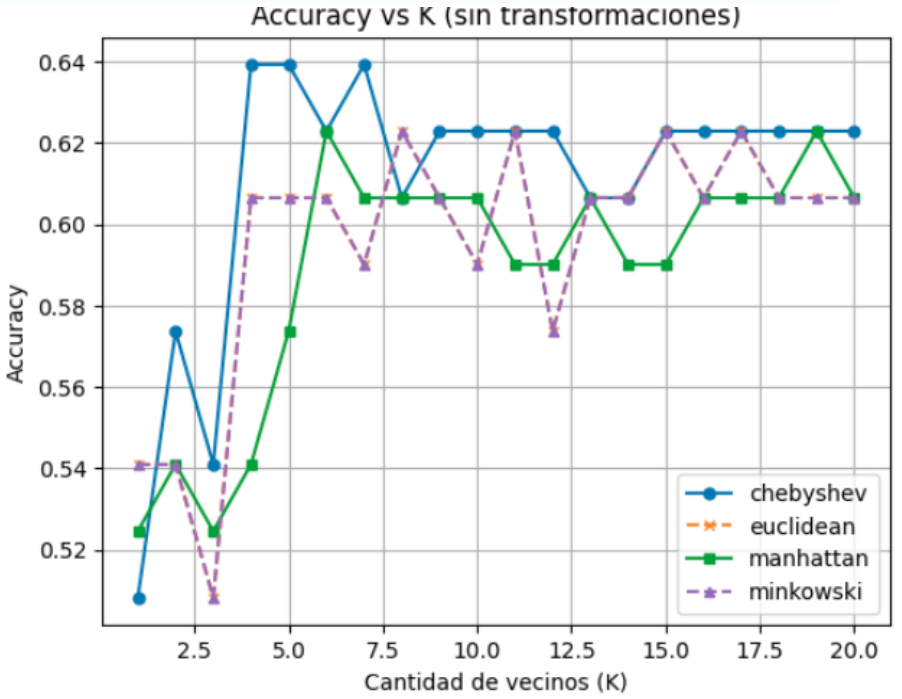


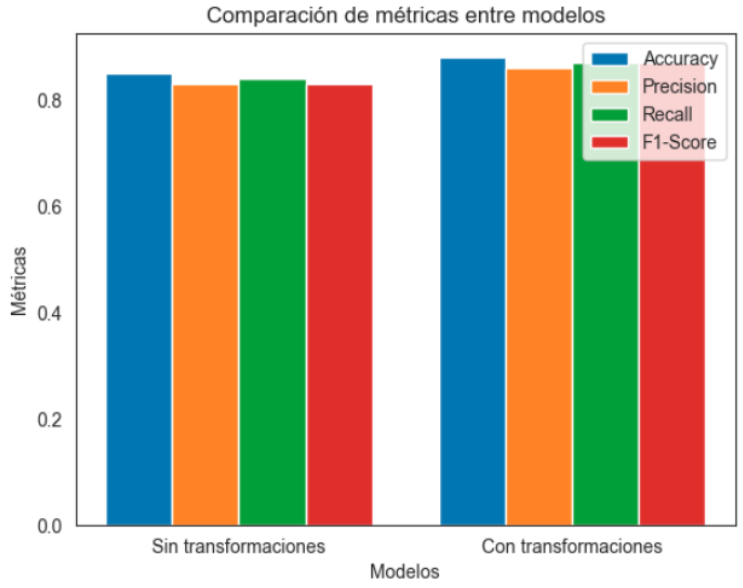


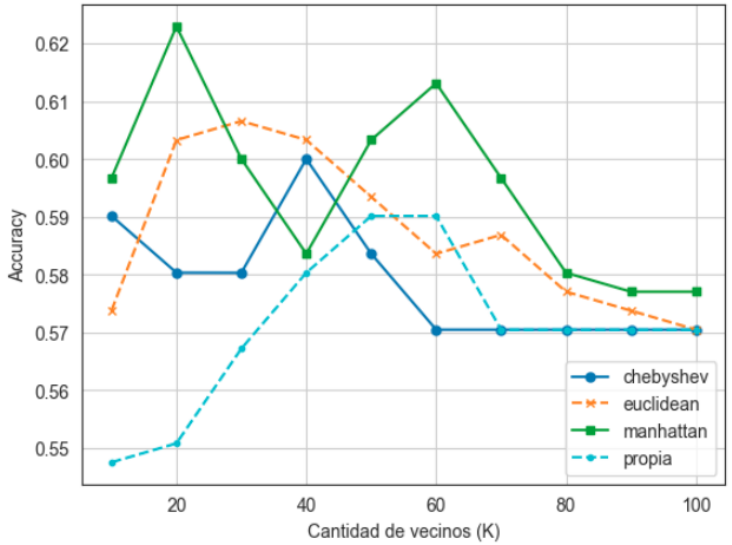


Valores K (Sin transformaciones)

El análisis muestra cómo **KNN sin transformación de datos** afecta la clasificación de enfermedad cardíaca. Destaca la influencia de **Edad y Frecuencia Cardíaca Máxima**, diferencias en **Precisión y Recall** y cómo distintas métricas de distancia impactan el modelo.

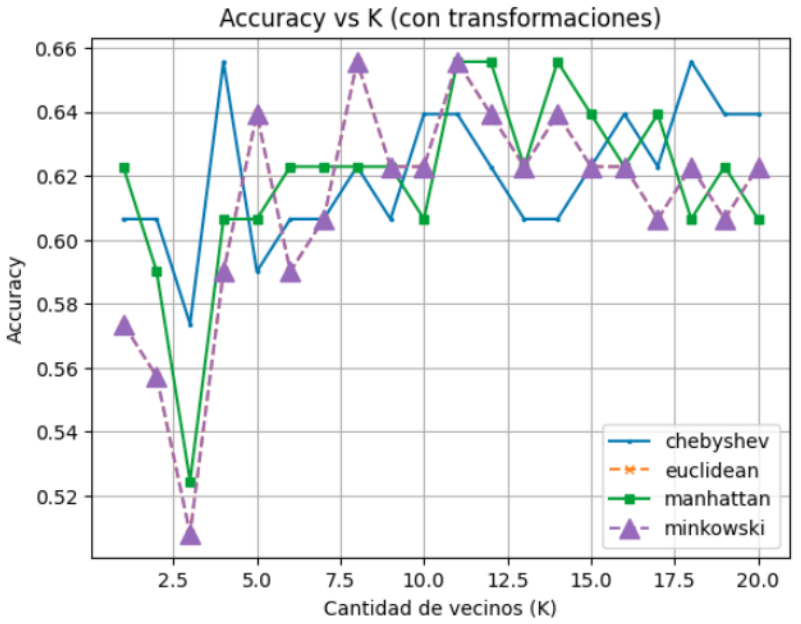


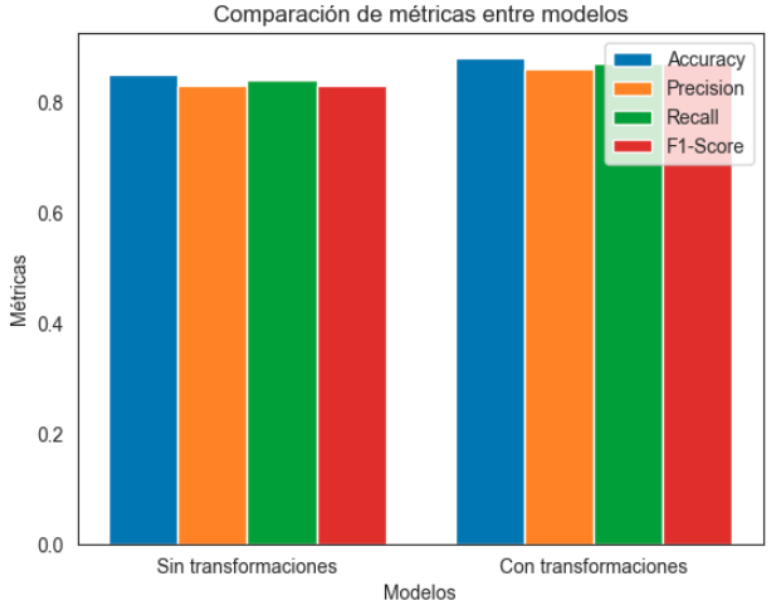


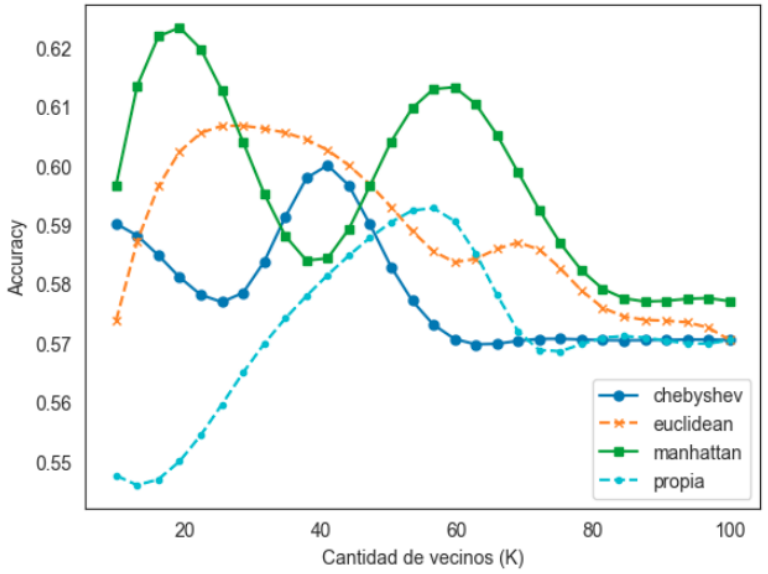


Valores K (Con transformaciones)

El modelo **KNN** optimizado con la métrica **manhattan (K=20)** demuestra el mejor rendimiento en las métricas clave: **Exactitud**, **Precisión**, **Recall** y **F1-Score**. Esto se logra gracias a las transformaciones aplicadas al dataset, que estabilizan las distribuciones de las variables, y la selección adecuada de hiperparámetros, ajustando tanto el número de vecinos como la métrica de distancia para maximizar la calidad predictiva. Este enfoque garantiza una clasificación precisa y equilibrada entre los casos positivos y negativos.







Resultados de la Calidad del Modelo

La tabla evalúa la calidad del modelo de **K-Nearest Neighbors (KNN)** bajo distintas métricas de distancia (**chebyshev, euclidean, manhattan, propia**) y valores de **K**. Los indicadores clave son:

- **ACC (Exactitud):** Representa la proporción de predicciones correctas. Los valores más altos, como en la distancia manhattan (K=20) con un 62.294%, indican mejor rendimiento global.
- **PRECISION:** Mide la proporción de verdaderos positivos sobre los positivos predichos. La distancia manhattan (K=20) muestra alta precisión, reflejando menos falsos positivos en la predicción.
- **RECALL:** Calcula la proporción de verdaderos positivos sobre todos los positivos reales. En este caso, manhattan (K=20) mantiene un recall elevado, demostrando la capacidad del modelo para identificar correctamente los casos relevantes
- **F1-SCORE:** Es el balance entre precision y recall. De nuevo, manhattan (K=20) destaca con el mejor rendimiento equilibrado (0.57496), consolidándose como la configuración más adecuada en términos generales

El análisis técnico muestra que la elección de la métrica de distancia y el valor de **K** son determinantes para optimizar la calidad del modelo, influyendo directamente en su desempeño predictivo.

	DISTANCE	K	ACC	PRECISION	RECALL	F1-SCORE
0	chebyshev	10	0.59016	0.52022	0.59016	0.53752
1	chebyshev	20	0.58036	0.48832	0.58036	0.51078
2	chebyshev	30	0.58034	0.4427	0.58034	0.49116
3	chebyshev	40	0.60002	0.47046	0.60002	0.50954
4	chebyshev	50	0.58362	0.4387	0.58362	0.48708
10	euclidean	10	0.57378	0.52992	0.57378	0.54438
11	euclidean	20	0.60328	0.5172	0.60328	0.55046
12	euclidean	30	0.60656	0.50218	0.60656	0.54376
13	euclidean	40	0.6033	0.4891	0.6033	0.52952
14	euclidean	50	0.59346	0.46816	0.59346	0.49796
20	manhattan	10	0.59672	0.5416	0.59672	0.55802
21	manhattan	20	0.62294	0.54346	0.62294	0.57496
22	manhattan	30	0.60002	0.51508	0.60002	0.54332
23	manhattan	40	0.58362	0.46722	0.58362	0.513
24	manhattan	50	0.6033	0.51864	0.6033	0.53778
30	propia	10	0.54756	0.36398	0.54756	0.43222
31	propia	20	0.55084	0.35042	0.55084	0.4237
32	propia	30	0.56722	0.37018	0.56722	0.44336
33	propia	40	0.58034	0.40298	0.58034	0.46834
34	propia	50	0.59016	0.43484	0.59016	0.49338

Predicciones

Este análisis se basa en **predicciones de pacientes sanos y enfermos** utilizando un modelo KNN con datos preprocesados.

.

Tabla de resultados: Se genera un DataFrame con la clasificación de cada paciente junto con el porcentaje de certeza del modelo, facilitando la interpretación de la precisión del sistema.

```
Modelo 'modeloKNN' cargado correctamente.

Predicciones para nuevos pacientes:
   Predicción Resultado Certeza
0        2.52   Enfermo    252%
```

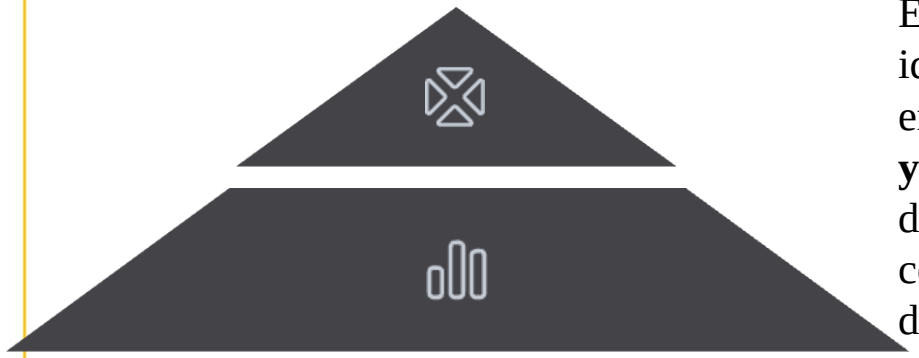
```
Tabla de Predicciones:
Paciente  Predicción Resultado Certeza
    Sano        0.36      Sano   64.0%
 Enfermo        1.20   Enfermo  120.0%
```

Predicción del modelo: Se establecen los umbrales para determinar si un paciente es clasificado como sano o enfermo, basado en un índice de probabilidad.

Transformación de datos: Las variables han sido ajustadas para coincidir con el modelo, asegurando que el escalado y la normalización no generen errores de formato.

Conclusiones

.



El análisis del **Heart Disease Dataset** permite desarrollar modelos predictivos que identifican patrones clave en la probabilidad de enfermedad cardíaca. Mediante el entrenamiento del modelo con variables como **colesterol, frecuencia cardíaca máxima y cantidad de vasos coloreados**, observamos tendencias significativas en la clasificación de pacientes. La evaluación del modelo refleja la importancia del **escalado de datos** y la correcta selección de parámetros como el umbral de clasificación, evitando sesgos hacia diagnósticos incorrectos. La matriz de confusión y las métricas de evaluación muestran la **precisión y recall** del modelo, permitiendo ajustes en hiperparámetros para mejorar su desempeño en futuras predicciones.

Referencias

.

1. Kelleher, J. D., & Mac Namee, B. (2019). *Machine Learning for Healthcare*. MIT Press.
2. Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer.
3. Müller, A. C., & Guido, S. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media.
4. Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

UNIVERSIDAD POLITÉCNICA SALESIANA ECUADOR

Más que formar Profesionales, *Transformamos Vidas*

@upsalesianaec

www.ups.edu.ec