1. Exploración Inicial y Análisis Descriptivo (EDA)

El objetivo de este apartado se basa en el análisis de la estructura del conjunto de datos mitbih_train.csv el cual contiene registros de latidos cardíacos con 187 valores por señal y una etiqueta. A través de un (EDA), se pretende entender dicho conjunto de datos y ser capaz de identificar patrones, diferenciar entre clases, y descubrir posibles anomalías relevantes de la señal ECG para los apartados posteriores.

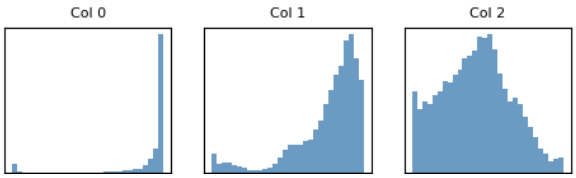
Estructura del dataset

- El conjunto contiene 87.554 registros y 188 columnas (187 valores de señal + 1 etiqueta).
- Todas las señales están normalizadas entre 0 y 1, sin valores nulos ni errores.
- Las clases están desbalanceadas: la clase 0 (latido normal) representa más del 80 % de los datos.
- Al transformar el problema a clasificación binaria (normal o anómalo), se obtiene una distribución más manejable (82 % vs. 18 %) pero aun asi altamente desbalanceada.

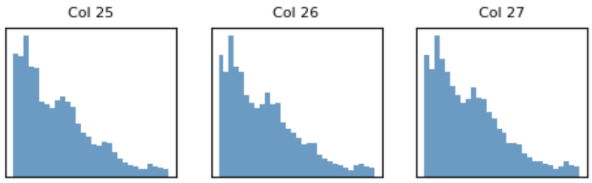
Tipos de columnas según su comportamiento

Se ha observado que las columnas pueden agruparse según su forma de distribución:

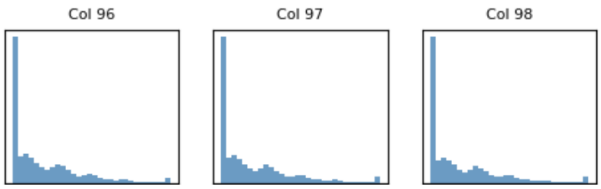
- Columnas 0 a 3: Estas posiciones muestran formas de distribución muy distintas al resto y variables también entre ellas. Esto sugiere que corresponden a los primeros instantes del latido cuyas formas probablemente sean por motivo de picos iniciales o porque algunas señales empiezan antes o después.



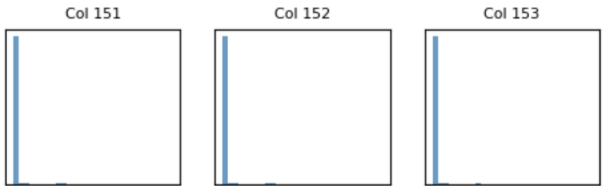
- Columnas 4 a 89 aproximadamente: En esta franja, la distribución de los valores se vuelve más regular y similar entre columnas. Resalta una concentración de valores bajos con colas a la derecha. Estas posiciones parecen corresponder a una fase activa del latido que es relativamente estable y repetitiva.



- Columnas 90 a 139 aproximadamente: Aunque las distribuciones siguen teniendo una forma similar (pico en 0 y cola hacia la derecha de manera descendente), se observa que la cantidad de valores mayores que 0 disminuye progresivamente.



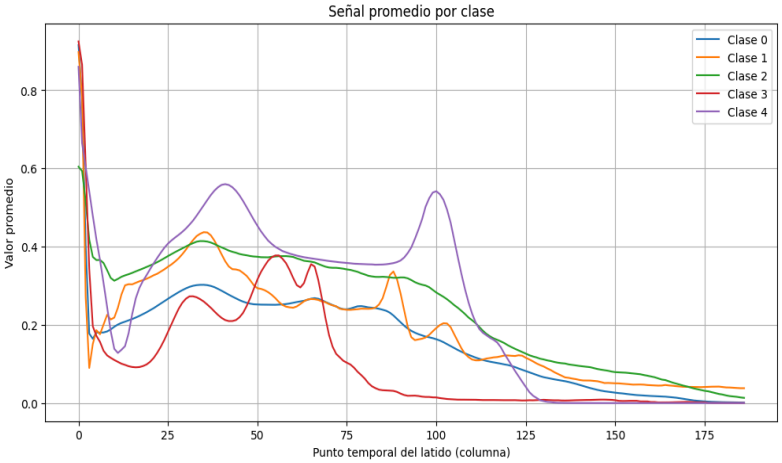
- Columnas 140 en adelante: A partir de esta sección, la gran mayoría de valores son exactamente 0 y no contienen información relevante en la mayoría de los registros. Muy probablemente sean parte del relleno al final del latido.

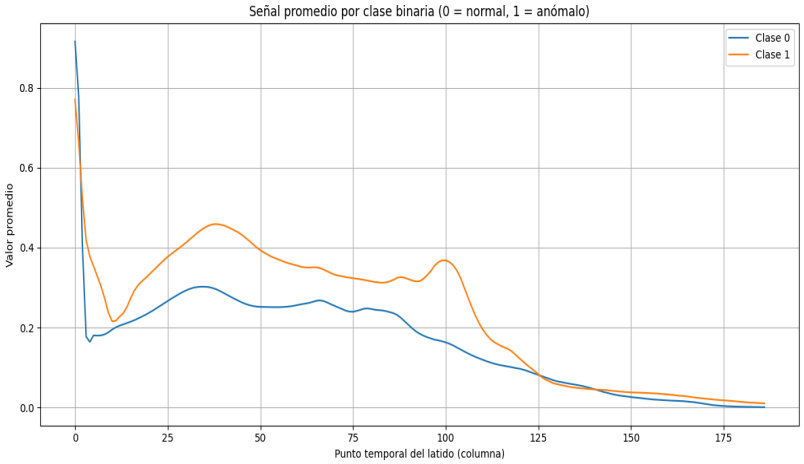


Estadísticas descriptivas globales

Los resultados estadísticos obtenidos (media, desviación, simetría...) confirman numéricamente lo observado en el análisis anterior.
Las primeras columnas muestran más variabilidad e inestabilidad, las intermedias son más estables y las finales tienen muy poca actividad, alineándose con la evolución temporal del latido.

Análisis por clase

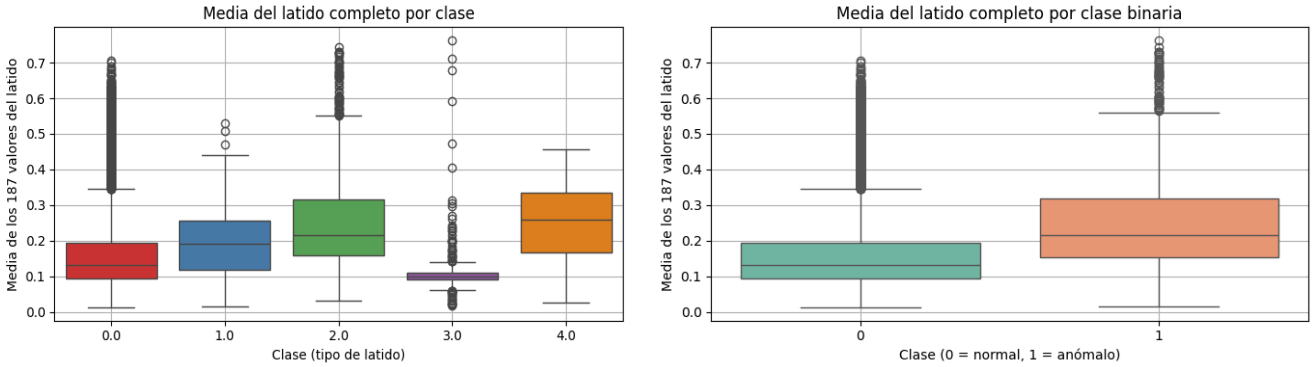




Todas las curvas comienzan con un valor alto en la columna 0, reflejando el pico inicial típico de un latido. En la gráfica de la izquierda (clasificación multiclase), se observa que la clase 0 (latido normal) tiene una señal suave, sin picos pronunciados, lo que indica estabilidad. Las clases 1 y 2 presentan más oscilaciones, con subidas y bajadas a lo largo del tiempo. Especialmente la clase 1 muestra picos evidentes alrededor de las columnas 35 y 95. La clase 3 es claramente diferenciable, con una forma de señal más baja y particular. La clase 4, correspondiente a ruido, mantiene valores más elevados e irregulares durante casi todo el latido, lo que puede indicar interferencias.

En la gráfica de la derecha, con clasificación binaria, también se aprecia un pico inicial común en ambas clases. Sin embargo, la clase 0 mantiene una señal media más baja y estable a lo largo del tiempo, mientras que la clase 1 conserva valores más altos, con picos alrededor de las columnas 35 y 100. Esto puede reflejar un carácter anómalo. A partir de la columna 125, ambas todas las clases de ambas gráficas descienden progresivamente hacia valores cercanos a cero, coincidiendo con la pérdida de información útil.

Media del latido completo por clase



En la clasificación multiclase, los latidos normales (clase 0) presentan medias más bajas y estables, mientras que las clases 1, 2 y 3 muestran mayor variabilidad y medias más altas. Por otra parte, la clase número 3 vemos que tiene valores mucho más concretos y con una media mucho más baja como ya se observaba en las imágenes del apartado anterior.

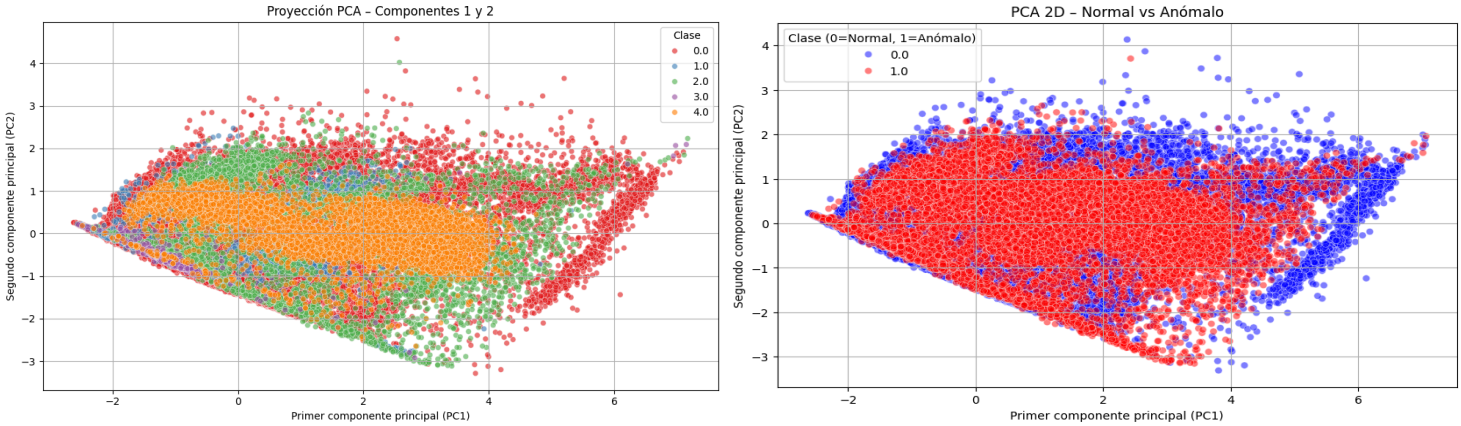
En la clasificación binaria, los latidos normales (clase 0) siguen mostrando valores promedio más bajos que los anómalos (clase 1) corroborando también las gráficas anteriores.

2. Reducción de Dimensionalidad (PCA)

Antes de aplicar PCA, se ha realizado un filtrado para eliminar columnas con muy poca variabilidad. Usando un umbral de varianza del 2 %, se han eliminado 45 columnas, reduciendo el conjunto de 187 a 142 columnas. Este paso se justifica debido a que se observó en el análisis exploratorio anterior que hay muchas columnas (especialmente a partir de la 140) con valores casi siempre iguales a cero, por lo que apenas aportan información de ningún tipo. Eliminar estas columnas permite que el PCA se enfoque en las zonas activas y con variación real de la señal, mejorando el estudio.

Por otro lado, la reducción de dimensionalidad mediante PCA ha mostrado resultados coherentes y adecuados para el análisis:

Con 2 componentes, se conserva el 55,65 % de la varianza total y con 3 componentes, la varianza explicada asciende a 62,68 %. Estos valores indican que el PCA está capturando una buena parte de la información relevante del conjunto de datos, y son adecuados para los fines exploratorios.



En la clasificación binaria, el solapamiento entre clases normales y anómalas es también muy evidente. Aunque la proyección ayuda a visualizar cierta estructura, no se observa una separación clara entre ambas clases.

En conclusión, predomina un fuerte solapamiento entre categorías, lo que indica que no son linealmente separables. Aun así, pueden existir patrones útiles, aunque su interpretación directa a través de las visualizaciones resulta compleja.

3. Modelos de Clasificación

El objetivo de este análisis es seleccionar el modelo de clasificación binaria más adecuado para diferenciar entre latidos normales (clase 0) y latidos anómalos (clase 1).
Desde el punto de vista clínico, la prioridad no es únicamente la precisión global, sino especialmente minimizar los falsos negativos (FN): aquellos casos en los que un latido anómalo es clasificado erróneamente como normal, ya que podrían ocultar una situación peligrosa y que cuando llegue a detectarse finalmente, esta situación puede que se haya agravado. Por tanto, aunque todas las métricas se tendrán en cuenta, el foco principal se sitúa en maximizar el recall de la clase 1 y reducir los FN al mínimo posible

	Regresión Logística	Árbol de Decisión	Clasificador Bayesiano
Accuracy	0.8401	0.9598	0.7829
Precision (0)	0.9466	0.9762	0.9147
Recall (0)	0.8551	0.9753	0.8135
F1 (0)	0.8985	0.9757	0.8611
Precision (1)	0.5248	0.8818	0.4153
Recall (1)	0.7684	0.8858	0.6359
F1 (1)	0.6237	0.8838	0.5025
Macro avg F1	0.7611	0.9298	0.6818
Falsos Negativos (FN)	874.0	431.0	1374.0
Falsos Positivos (FP)	2626.0	448.0	3379.0
True Negatives (TN)	15492.0	17670.0	14739.0
True Positives (TP)	2900.0	3343.0	2400.0

Análisis métrico por métrico

- **Accuracy**: El modelo de Árbol de Decisión es el más preciso (0.9598), seguido por la Regresión Logística (0.8401) y finalmente el Bayesiano (0.7829).
- **Precisión clase 0:** El árbol de decisión obtiene la mejor precisión (0.9762), seguido por la regresión logística (0.9466) y el bayesiano (0.9147) pero muy de cerca.
- **Recall clase 0:** El árbol también lidera (0.9753), con la regresión en 0.8551 y el bayesiano en 0.8135.

- **F1-score clase 0:** De nuevo destaca el árbol (0.9757) con un valor muy cercano a la perfección y obviamente por encima de la regresión (0.8985) y del bayesiano (0.8611).
- **Precisión clase 1**: El árbol de decisión ofrece la mayor precisión (0.8818), lo que indica que la mayoría de las predicciones de “anómalo” realmente lo son. El modelo bayesiano queda lejos (0.4153), lo que refleja una alta tasa de falsos positivos.
- **Recall clase 1**: Aquí se refleja directamente la capacidad para detectar latidos peligrosos. El árbol de decisión detecta el 88.58 % de los casos, la regresión el 76.84 % y el bayesiano tan solo el 63.59 %.
- **F1-score clase 1:** Combinando precisión y recall, el árbol de decisión vuelve a liderar con 0.8838, confirmando sus buenos resultados.
- **Falsos Negativos (FN)**: Esta es la métrica crítica desde el punto de vista clínico. El árbol de decisión comete solo 431 errores graves, mientras que la regresión comete 874 y el bayesiano 1374.
- **Falsos Positivos (FP)**: Aunque el árbol de decisión también es el que menos falsos positivos genera (448), la diferencia con el resto no es tan abismal como en FN pero aún así sigue siendo el mejor.
- **Macro F1 promedio**: De nuevo, el árbol de decisión obtiene la media más alta de F1-score entre ambas clases..

Conclusión

A la vista de todas las métricas, el modelo de Árbol de Decisión demuestra ser el mejor en términos de precisión, recall, F1-score y, especialmente, número de falsos negativos. Prácticamente en todos los campos. No solo cumple con los requisitos de rendimiento, sino que se alinea perfectamente con el criterio clínico de gran importancia que comentamos previamente: Que se clasifiquen como normales los mínimos latidos anómalos como sea posible.

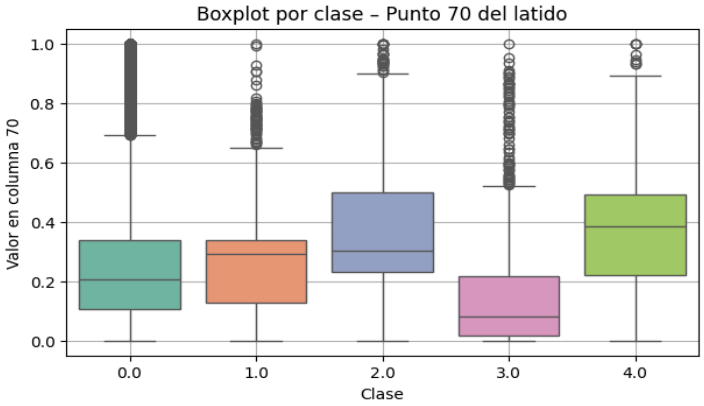
Además, tanto este como los demás modelos han sido entrenados con los hiper parámetros óptimos, determinados tras evaluar múltiples configuraciones posibles, a continuación la configuración final del mejor modelo, el modelo de Árbol de decisión.

- *max_depth=20*: se refiere a la profundidad máxima de 20 niveles, permitiendo al árbol captar relaciones complejas sin llegar al sobreajuste.
- *class_weight='balanced':* ajusta automáticamente los pesos de las clases según su frecuencia. Esto compensa el desbalance entre latidos normales y anómalos, mejorando la detección de los anómalos al ser menos frecuentes y tener menos proporción.
- *random_state=42:* parámetro que siempre se suele usar con este valor en los modelos para que los resultados sean reproducibles.

Adicionalmente, es importante destacar el impacto que tiene el desbalanceo de clases en los resultados. Aunque tanto en el Árbol de Decisión como en la Regresión Logística se ha utilizado el parámetro *class_weight='balanced'* para suavizar este efecto, el desbalance sigue influyendo en el rendimiento del modelo, especialmente en las métricas asociadas a la clase 1 (latidos anómalos), que en general son más bajas que las de la clase 0 (latidos normales). Esta situación es aún más evidente en el Clasificador Bayesiano, que no permite ajustes de pesos por clase de la misma manera que los otros modelos. Como consecuencia, este modelo presenta un desempeño mucho más débil a la hora de identificar correctamente los latidos anómalos, lo cual se refleja en una menor precisión, recall y F1-score para la clase 1. Esto subraya la importancia de tener en cuenta el desbalanceo de cualquier conjunto de datos a la hora de un estudio de cualquier tipo.

4. Probabilidades Condicionales y Teorema de Bayes

Análisis del Patrón: Valor en columna 70 entre 0.02 y 0.22 (clases 0 vs 3)

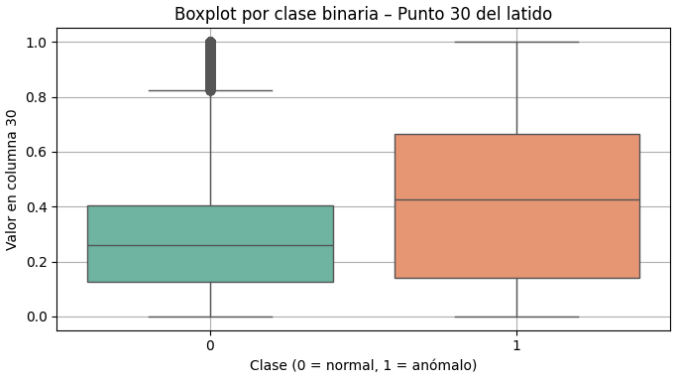


Este patrón se ha diseñado observando la anterior imagen y en concreto la separación entre las clases 0 y 3 en el boxplot, donde se aprecia que la clase 3 tiende a tener valores considerablemente más bajos y agrupados en la columna 70. El objetivo es ver si esta franja concreta permite detectar latidos de clase 3. Resultados obtenidos:

	P(clase = 0)	P(clase = 3)	P(patrón | clase = 0)	P(patrón | clase = 3)	P(patrón)	P(clase = 0 | patrón)	P(clase = 3 | patrón)
Valor (%)	99.12	0.88	46.46	47.89	46.47	99.1	0.9

A pesar de que el patrón aparece en una proporción similar en ambas clases, la probabilidad final está completamente dominada por la clase 0 debido a su aplastante mayoría. La clase 3 representa menos del 1 % de los casos, y eso hace que incluso cuando se cumple el patrón, la probabilidad de que se trate de un latido de clase 3 sea solo del 0.9 %. Desde un punto de vista clínico, este patrón no es útil para detectar latidos anómalos tipo 3.

Análisis del Patrón: Valor en columna 30 mayor de mayor de 0.5 (clasificación binaria)



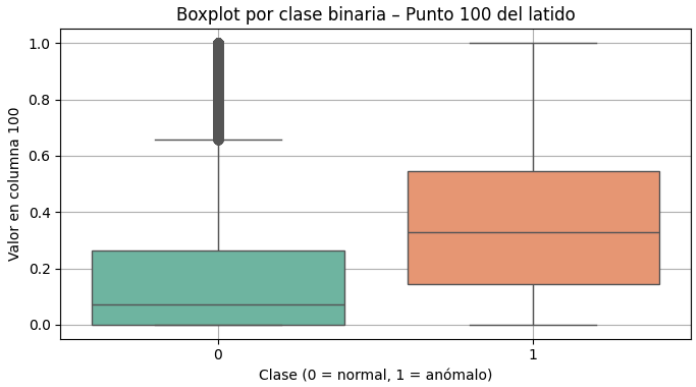
Este patrón se ha escogido concretamente después de estudiar la gráfica anterior en la que se observa que a partir de 0.5 la mayor concentración de latidos son referentes a la clase 1 (anómalos).

	P(clase = 0)	P(clase = 1)	P(patrón | clase = 0)	P(patrón | clase = 1)	P(patrón)	P(clase = 0 | patrón)	P(clase = 1 | patrón)
Valor (%)	82.77	17.23	70.66	29.39	63.55	92.03	7.97

Este patrón puede llegar a ser útil para identificar latidos anómalos. Aunque debemos tener en cuenta algunas ciertas características:

Una probabilidad de 40.99 % de que el patrón ocurra en anomalías, frente a solo 12.89 % en latidos normales hace que este patrón sea interesante y digno de estudio. Los demás resultados no son tan favorables o concluyentes a simple vista pero de nuevo se debe a la fuerte descompensación que existe a favor de latidos normales.

Análisis del Patrón: Valor en columna 100 menor de 0.2 (clasificación binaria)



A continuación, se va a estudiar este patrón dado que observamos en el anterior boxplot que debe haber una gran diferencia entre estas dos clases para el intervalo entre 0 y 0.2 a favor de la clase 0.

	P(clase = 0)	P(clase = 1)	P(patrón | clase = 0)	P(patrón | clase = 1)	P(patrón)	P(clase = 0 | patrón)	P(clase = 1 | patrón)
Valor (%)	82.77	17.23	12.89	40.99	17.73	60.17	39.83

Este patrón es útil para confirmar latidos normales (clase 0). Observamos que:

P(clase = 0 | patrón) = 92.03 %, es decir, si un latido tiene un valor bajo en la columna 100, casi con total certeza es normal. Parece útil para descartar anomalías de forma eficiente aunque como hemos comentado durante el informe tenemos que ser muy cautelosos dado que consideramos el hecho de pasar por alto una anomalía algo grave. En cuanto a lo referente a esto, P(clase = 1 | patrón) = 7.97 % indica que aún hay cierto riesgo de pasar por alto anomalías, por lo que no debe usarse al menos como un único criterio de diagnóstico.