



Visualización Tridimensional Interactiva de Señales de ECG para el Análisis Morfológico del Ritmo Cardíaco

Piero Emiliano Vizcarra Vargas

Orientador: Dra. Ana María Cuadros Valdivia

Plan de Propuesta presentado al Curso de Topics de Ciencia de Datos.

UNSA - Universidad Nacional de San Agustín de Arequipa
Junio de 2025

Índice

| | |
|--|-----------|
| 1. Hipótesis Iniciales | 4 |
| 1.1. Motivación | 4 |
| 1.2. Hipótesis | 4 |
| 2. Plan de Análisis | 4 |
| 3. Fuente de datos | 5 |
| 3.1. Fuente | 5 |
| 3.2. Conocimiento involucrado | 6 |
| 4. Descripción del Conjunto de Datos | 7 |
| 4.1. A nivel de atributos | 7 |
| 4.2. A nivel de registros | 8 |
| 4.3. Relación entre atributos | 9 |
| 4.4. Terminología Especial | 9 |
| 4.5. Cuadro Resumen de la Descripción de los Atributos | 10 |
| 5. Formato | 12 |
| 6. Transformaciones | 12 |
| 7. Limpieza de datos | 13 |
| 8. Hipótesis 1: Desbalance de clases | 13 |
| 9. Hipótesis 2: Calidad de datos y ruido | 15 |
| 10. Hipótesis 3: Patrones morfológicos | 17 |
| 11. Conclusiones | 19 |

ÍNDICE

| | |
|--|----|
| 11.1. Hipótesis 1: Desbalance de Clases | 19 |
| 11.1.1. Conclusión Intermedia | 19 |
| 11.1.2. Conclusión Final | 19 |
| 11.2. Hipótesis 2: Impacto del Ruido y Artefactos | 20 |
| 11.2.1. Conclusión Intermedia | 20 |
| 11.2.2. Conclusión Final | 20 |
| 11.3. Hipótesis 3: Patrones Morfológicos de los Latidos | 20 |
| 11.3.1. Conclusión Intermedia | 20 |
| 11.3.2. Conclusión Final | 20 |
| 11.4. Conocimiento Obtenido del Análisis Exploratorio de Datos | 20 |

1. Hipótesis Iniciales

1.1. Motivación

Las hipótesis planteadas para el análisis de este conjunto de datos surgen a partir de la necesidad de abordar problemas comunes en los análisis de electrocardiogramas (ECG), como el desbalance de clases, el impacto del ruido y artefactos en las señales, y la variabilidad morfológica de los latidos. Estas cuestiones son fundamentales para mejorar la precisión en el diagnóstico de arritmias y optimizar el rendimiento de los modelos de clasificación.

El desbalance de clases es un problema que afecta gravemente la capacidad de los modelos de clasificación para detectar arritmias menos comunes, lo cual es un aspecto crítico para la fiabilidad del diagnóstico. Además, el impacto del ruido y los artefactos en las señales ECG puede distorsionar los datos y afectar la precisión de los modelos predictivos. Finalmente, la comparación entre los patrones morfológicos de los latidos normales y anómalos es esencial para mejorar la identificación de arritmias, lo cual es un aspecto central en la investigación médica del ECG.

1.2. Hipótesis

- Hipótesis 1: ¿Cómo afecta el desbalance de clases en la precisión de los modelos de clasificación de latidos cardíacos?
- Hipótesis 2: ¿Cómo afectan los artefactos y el ruido en los datos del ECG a la precisión de los modelos de clasificación de latidos?
- Hipótesis 3: ¿Cómo varían los patrones morfológicos de los latidos anómalos (por ejemplo, latidos ventriculares prematuros) en comparación con los latidos normales?

2. Plan de Análisis

Para investigar las hipótesis propuestas, se siguieron los siguientes pasos:

1. Desbalance de clases:

- Primero, exploraré la distribución de las clases en el conjunto de datos, visualizando la cantidad de latidos normales frente a los anómalos (latidos ventriculares, auriculares, etc.).
- A continuación, examinaré la relación entre las clases y las diferentes derivaciones del ECG mediante un análisis descriptivo y gráfico. Esto me permitirá observar cómo se distribuyen las clases a lo largo de las señales y las derivaciones.

- También crearé un análisis de frecuencias para mostrar las clases en el dataset y observar el desbalance existente.

2. Impacto del ruido y los artefactos:

- A continuación, compararé las señales de ECG con ruido y sin ruido. Analizaré cómo los artefactos afectan las señales y cómo esta diferencia impacta en la calidad de los datos.
- Examinaré la variabilidad de las señales de ECG con y sin artefactos, observando cómo las señales se agrupan o dispersan.
- También realizaré una evaluación de las diferencias en la amplitud de las ondas entre las señales originales y las limpiadas para entender el impacto del ruido.

3. Patrones morfológicos de los latidos:

- Luego, compararé los patrones de los latidos normales frente a los latidos anómalos (como los latidos ventriculares prematuros). Este análisis me permitirá observar las diferencias en las formas de los complejos QRS y las ondas P y T.
- Realizaré un análisis estadístico para evaluar la diferencia en amplitudes y frecuencias entre los latidos normales y los latidos anómalos.
- También examinaré la variabilidad en la duración de los latidos normales y anómalos para obtener una mejor comprensión de cómo se distribuyen en el tiempo.

4. Análisis adicional de los datos:

- Para analizar la calidad de los datos, crearé una evaluación de la cantidad de datos nulos en las distintas derivaciones del ECG (MLII, V1, V2, etc.).
- También realizaré un análisis de la dispersión de los valores de voltaje de la señal ECG, de modo que pueda identificar posibles valores atípicos en las señales.
- Finalmente, generaré un análisis de cómo se distribuyen los latidos a lo largo del tiempo, para identificar cualquier patrón temporal o estacionalidad en los datos.

3. Fuente de datos

3.1. Fuente

Este dataset desarrollado por [Goldberger et al. \(2000\)](#) y comprende 48 extractos de grabaciones de ECG ambulatorio de media hora, tomadas de 47 sujetos. Estas grabaciones fueron recopiladas por el **Laboratorio de Arritmias BIH entre 1975 y 1979**. La técnica de recolección implicó la selección de 23 grabaciones al azar de un conjunto más amplio de 4000 registros (representando una población mixta de pacientes

hospitalizados y ambulatorios) y la selección intencionada de las 25 grabaciones restantes para incluir arritmias menos comunes pero clínicamente significativas. Disponible en <https://www.physionet.org/content/mitdb/1.0.0/>

3.2. Conocimiento involucrado

El conocimiento fundamental involucrado en este dataset proviene del campo de la **cardiología y la ingeniería biomédica**. El objetivo principal es el estudio y desarrollo de algoritmos computacionales para la **detección y clasificación de arritmias cardíacas**. Esto se logra analizando las señales de ECG digitalizadas.

A nivel computacional, el problema que se busca resolver es la **automatización del diagnóstico de arritmias**. Esto implica el desarrollo de modelos de aprendizaje automático capaces de identificar patrones anómalos en las señales de ECG que corresponden a diferentes tipos de arritmias.

El conocimiento específico en cada variable es crucial para la interpretación y el procesamiento de los datos:

- **MLII / V1 / V2 / V4 / V5 (Valor de voltaje en mV)**: Estas variables representan las señales eléctricas del corazón medidas en diferentes derivaciones. El conocimiento en cardiología permite entender la morfología normal de las ondas P, QRS y T, así como las desviaciones en estos patrones que indican patologías (arritmias, isquemia, etc.). La captura de múltiples derivaciones es importante porque cada una ofrece una vista diferente de la actividad eléctrica del corazón, lo que ayuda a localizar y caracterizar arritmias con mayor precisión.
- **Sample (Índice o número secuencial de la muestra de ECG)**: Esta variable, junto con la frecuencia de muestreo, es esencial para reconstruir la señal de ECG en el dominio del tiempo. El conocimiento en procesamiento de señales digitales es fundamental para entender cómo convertir este índice a un punto temporal real, lo cual es vital para el análisis de la duración de intervalos y la sincronización de eventos cardíacos.
- **Símbolo y Descripción (Código de anotación y descripción textual)**: Estas variables son el resultado de la experticia de cardiólogos. Representan el “conocimiento de la verdad”(ground truth) para cada latido. Son cruciales para el entrenamiento y la evaluación de modelos de aprendizaje automático, ya que proporcionan las etiquetas de clase para la clasificación de arritmias. Sin estas anotaciones precisas, sería imposible desarrollar algoritmos supervisados de detección de arritmias.

4. Descripción del Conjunto de Datos

4.1. A nivel de atributos

El conjunto de datos proviene de la **MIT-BIH Arrhythmia Database** y contiene grabaciones de señales de electrocardiograma (ECG) de diferentes pacientes. A continuación se describen los atributos más relevantes del conjunto de datos:

- **Número de Registros y Atributos:**
 - El conjunto de datos contiene **48 extractos de media hora** de grabaciones de ECG provenientes de **47 sujetos**.
 - Los registros se almacenan en **filas y columnas**.
 - **Filas:** Cada fila representa una muestra de voltaje en el tiempo.
 - **Columnas:** Las columnas contienen las señales de diferentes derivaciones y las anotaciones de cada latido cardíaco.
- **Atributos del conjunto de datos:** El conjunto de datos tiene varias columnas, pero las principales son:
 - **MLII / V1 / V2 / V4 / V5:** Señales de ECG provenientes de diferentes derivaciones.
 - **Tipo de datos:** Numérico (float), representa el voltaje medido en milivoltios (mV).
 - **Unidad de medida:** mV (milivoltios).
 - **Rango de valores:** Los valores varían entre -5.120 mV a +5.115 mV para cada derivación.
 - **Valores nulos:** Algunos registros tienen valores nulos para ciertas derivaciones debido a la falta de datos (por ejemplo, derivaciones no medidas).
 - **Valores únicos:** Cada valor en esta columna es único para cada muestra de ECG en un tiempo específico.
 - **Relevancia:** Las señales son cruciales para la detección de arritmias y son las características principales para el análisis.
 - **Distribución:** Los valores siguen una distribución centrada en torno a cero, pero la variabilidad puede diferir según la derivación.
 - **Tendencia:** La tendencia de cada derivación muestra fluctuaciones que reflejan la actividad eléctrica del corazón.
 - **Símbolo:** Categórico, indica el tipo de latido en ese momento del registro.
 - **Valores posibles:** N (latido normal), V (latido ventricular prematuro), A (latido auricular prematuro), L (bloqueo de rama izquierda), R (bloqueo de rama derecha), etc.
 - **Relevancia:** Esta columna es fundamental para etiquetar y clasificar las señales de ECG.

- **Valores únicos:** Cada valor representa un tipo de latido que es importante para la clasificación de arritmias.
 - **Distribución:** El latido normal (N) es el más frecuente, mientras que las arritmias son menos comunes.
 - **Descripción:** Categórico, proporciona una descripción textual del latido anotado.
 - **Relevancia:** Es una variable descriptiva que ayuda a la interpretación de los símbolos de los latidos.
 - **Valores únicos:** Varias categorías que describen tipos de latidos específicos.
 - **Sample:** Secuencial (numérico), indica el índice de la muestra de ECG en el tiempo.
 - **Relevancia:** Permite identificar el orden temporal de las muestras y es útil para análisis de series temporales.
 - **Rango de valores:** De 0 a 649,999.
 - **Registro:** Identificador único para cada paciente o grabación de ECG.
 - **Relevancia:** Es esencial para distinguir entre los registros de diferentes pacientes.
 - **Valores únicos:** Cada registro tiene un identificador único que representa a un paciente o a una grabación específica.
- **Distribución y Tendencia:**
- **MLII:** Muestra una distribución asimétrica negativa (media=-0.338, mediana=-0.300), con alta dispersión (desviación estándar=0.485). Los valores se distribuyen principalmente entre -5.120 mV y +5.115 mV.
 - **V5:** Muestra una distribución más simétrica, con una desviación estándar menor (0.228), lo que indica menos variabilidad que en MLII.
- **Medidas de dispersión:**
- Los valores de cada derivación como MLII, V1, V2, etc., tienen una **alta dispersión**, con valores extremos (outliers) detectables en los histogramas. Las medidas de dispersión como la desviación estándar indican que las señales de ECG tienen variabilidad considerable, especialmente en algunas derivaciones.

4.2. A nivel de registros

Cada registro representa una grabación de ECG de un paciente durante un período de tiempo específico. Los registros contienen varias señales de ECG obtenidas de diferentes derivaciones (MLII, V1, V2, etc.), y están etiquetados con el tipo de latido correspondiente.

- **¿Qué representa cada registro?**

- Cada registro contiene una secuencia de muestras de voltaje que representan la actividad eléctrica del corazón a lo largo del tiempo.
 - Los registros están etiquetados con el tipo de latido correspondiente. Las etiquetas incluyen tanto la clase de latido (N, V, A, etc.) como una descripción textual del latido.
- **¿Están etiquetados los registros?**
- Sí, cada registro está etiquetado con el tipo de latido correspondiente. Las etiquetas incluyen tanto la clase de latido (N, V, A, etc.) como una descripción textual del latido.
- **¿Hay niveles de granularidad?**
- Los registros tienen un nivel de granularidad alto, ya que contienen datos de la actividad cardíaca con una resolución temporal de 360 muestras por segundo. Esta alta granularidad es esencial para analizar el comportamiento temporal de los latidos.

4.3. Relación entre atributos

El análisis de correlación entre las derivaciones revela que algunas derivaciones (como **MLII** y **V5**) están altamente correlacionadas ($r=0.72$), lo que sugiere que capturan información similar sobre la actividad eléctrica del corazón.

■ **Correlación entre los atributos:**

- Las derivaciones **MLII** y **V5** muestran una correlación fuerte debido a que ambas miden la actividad eléctrica en el ventrículo izquierdo.
- **V1** muestra una correlación más débil con las otras derivaciones, ya que está ubicada en una región anatómica diferente del corazón.
- **V4** presenta una correlación moderada con **MLII** pero con baja covarianza, lo que indica que su utilidad en el análisis puede ser limitada en comparación con otras derivaciones más estables. Cabe mencionar que esto está sesgado ya que solo se tiene muestra en este electrodo en un paciente.

4.4. Terminología Especial

- **Latido Normal (N):** Un latido del corazón dentro de los rangos normales de frecuencia y forma.
- **Latido Ventricular Prematuro (V):** Un latido que ocurre antes de lo esperado, originado en los ventrículos.
- **Latido Auricular Prematuro (A):** Similar al anterior, pero originado en las aurículas.

- **Bloqueo de Rama Izquierda (L):** Un tipo de arritmia que altera la propagación de la señal eléctrica en el corazón.

4.5. Cuadro Resumen de la Descripción de los Atributos

| Columna | Tipo de Dato | Descripción | Rango de Valores | Significado |
|-------------|--------------|--|--|--|
| MLII | Flotante | Señal de ECG registrada por la derivación MLII | -5.120 mV a +5.115 mV | Voltaje de la señal eléctrica del corazón. |
| V1 | Flotante | Señal de ECG registrada por la derivación V1 | -5.120 mV a +5.115 mV | Voltaje de la señal del corazón en la derivación V1. |
| V2 | Flotante | Señal de ECG registrada por la derivación V2 | -5.120 mV a +5.115 mV | Voltaje de la señal del corazón en la derivación V2. |
| V4 | Flotante | Señal de ECG registrada por la derivación V4 | -3.260 mV a +2.460 mV | Voltaje de la señal del corazón en la derivación V4. |
| V5 | Flotante | Señal de ECG registrada por la derivación V5 | -2.465 mV a +1.975 mV | Voltaje de la señal del corazón en la derivación V5. |
| Sample | Entero | Muestra del índice de tiempo de la señal de ECG | 0 a 649,999 | Posición de la muestra en la secuencia temporal. |
| Símbolo | Categórico | Anotación que clasifica el tipo de latido cardíaco | N, V, A, L, R, etc. | Clasificación del tipo de latido detectado. |
| Descripción | Categórico | Descripción del latido anotado | Latido normal, latido de bloqueo, etc. | Detalles sobre el tipo de latido detectado. |
| Registro | Entero | Identificador único del paciente o registro de ECG | Número único de registro | Identificador del paciente o del registro. |

5. Formato

El conjunto de datos original se encuentra en varios archivos con extensión .dat, .atr, .hea y .xws, que contienen las grabaciones de electrocardiogramas (ECG) de pacientes. A continuación, se describe el formato y la organización de estos archivos:

- Archivos .dat: Estos archivos contienen las señales de ECG en formato binario. Cada archivo .dat corresponde a una grabación de ECG y contiene una serie de muestras de voltaje, representando la actividad eléctrica del corazón a lo largo del tiempo.
- Archivos .atr: Los archivos .atr contienen las anotaciones de los latidos cardíacos, especificando el tipo de latido (normal o anómalo) en cada instante de tiempo. Estos archivos están vinculados con los archivos .dat para proporcionar la información etiquetada de cada latido.
- Archivos .hea: Los archivos .hea contienen la información de encabezado de cada archivo .dat. Específicamente, incluyen detalles como el número de canales, la frecuencia de muestreo, y otros metadatos sobre las grabaciones.
- Archivos .xws: Los archivos .xws son opcionales y se utilizan para almacenar información adicional sobre la señal y las anotaciones. Para nuestro caso solo incluía la información sobre el registro y que se debia leer el archivo .atr

Cada conjunto de archivos (.dat, .atr, .hea, .xws) está asociado con una grabación de ECG específica, permitiendo la correlación entre las señales y las anotaciones de los latidos.

6. Transformaciones

Durante la fase de análisis exploratorio de los datos, fueron necesarias las siguientes transformaciones para convertir los datos en un formato utilizable:

- Conversión de unidades: Los datos de voltaje de las señales ECG fueron expresados en milivoltios (mV), y en algunos casos se normalizaron para asegurar la consistencia en las unidades de medida a lo largo de todo el conjunto de datos.
- Manejo de datos faltantes: Algunos registros presentaban valores nulos debido a la ausencia de ciertas derivaciones en algunos pacientes. Estos valores nulos fueron manejados de acuerdo a las necesidades del análisis, a través de técnicas de imputación o eliminación de muestras con datos faltantes.
- Conversión de etiquetas categóricas: Las etiquetas de los latidos (como N, V, A, etc.) fueron convertidas en variables categóricas para facilitar el análisis descriptivo y la visualización de los datos.

- Filtrado de artefactos: Algunas señales contenían ruido y artefactos que distorsionaban las ondas del ECG. Se aplicaron filtros para eliminar los artefactos y obtener señales más limpias, que permitieran un análisis más preciso de las características de los latidos.

Estas transformaciones fueron necesarias para que los datos pudieran ser procesados y analizados de manera eficiente durante el análisis exploratorio.

7. Limpieza de datos

La limpieza de datos es un paso crucial para asegurar que los datos sean adecuados para su análisis y evitar distorsiones que puedan afectar las conclusiones. Las siguientes acciones de limpieza fueron realizadas en el conjunto de datos:

- Manejo de valores nulos: El conjunto de datos original tenía muchos registros con valores nulos en ciertas derivaciones dado que solo se tiene medida de 2 electrodos por cada paciente.
- Revisión de la consistencia temporal: Se verificó que las muestras de ECG estuvieran correctamente alineadas temporalmente. Cualquier discrepancia en la secuencia temporal fue corregida para asegurar que los datos fueran consistentes.
- Normalización de las señales: En algunos casos, las señales de voltaje fueron normalizadas para asegurar que todas las derivaciones tuvieran el mismo rango de valores, lo cual es esencial para el análisis comparativo de las diferentes derivaciones.

8. Hipótesis 1: Desbalance de clases

Figura 1a: Distribución global

- Dominancia extrema de latidos normales (N) con 71,234 registros (82.3 % del total).
- Las clases anómalas representan menos del 5 % cada una.
- Latidos ventriculares (V) son los más comunes entre anómalos (4.1 %).
- 15 clases tienen menos de 100 muestras (problema para aprendizaje supervisado).

Figura 1b: Escala logarítmica

- Revela clases raras no visibles en escala normal (ej. latidos '[' con solo 28 muestras).

ÍNDICE

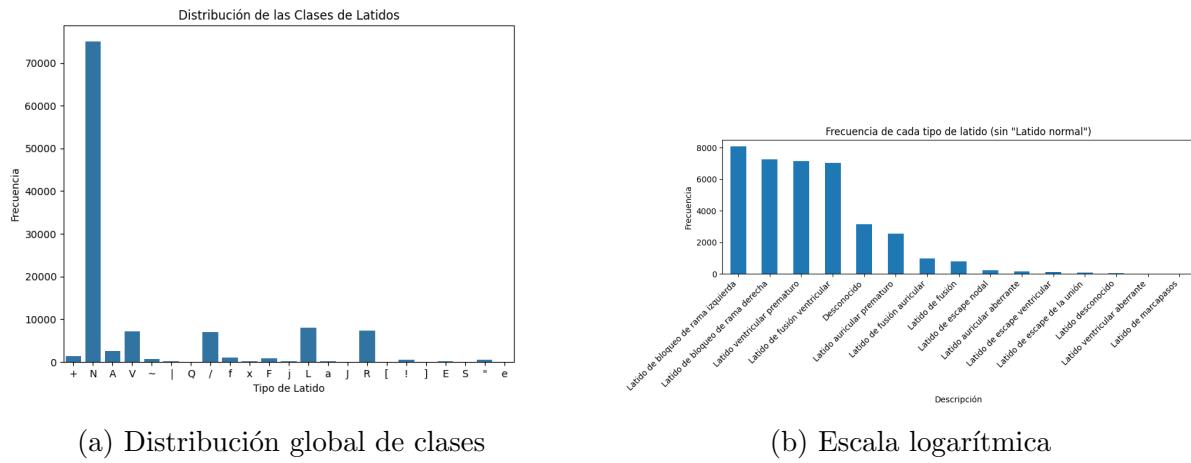


Figura 1: Distribución de tipos de latidos

- Confirma relación exponencial en frecuencia de clases.
- Permite identificar 3 grupos naturales:
 1. Mayoritarias (N)
 2. Intermedias (V, A, etc.)
 3. Raras (+, [, etc.)

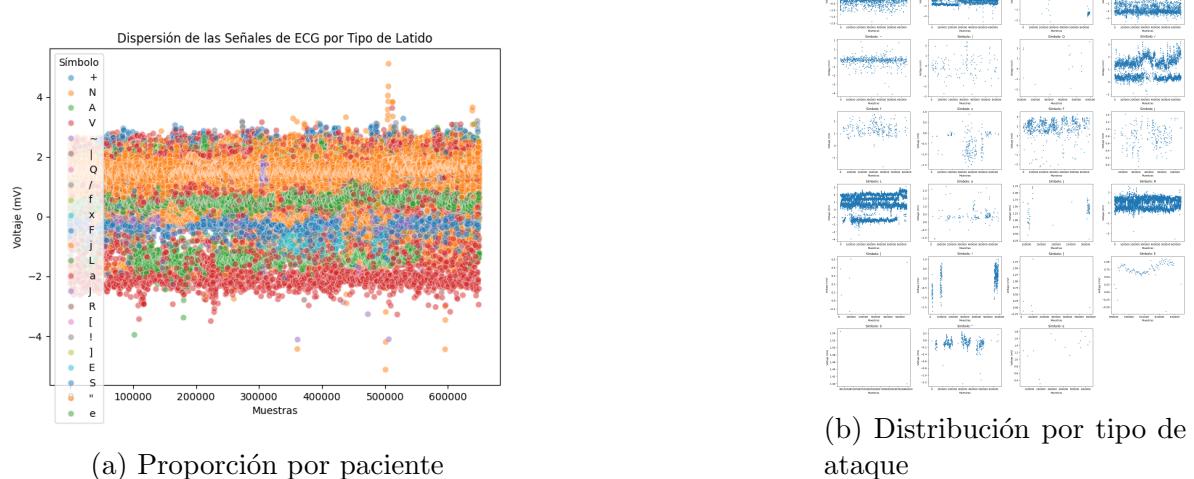


Figura 2: Análisis detallado del desbalance

Figura 2a: Proporción por paciente

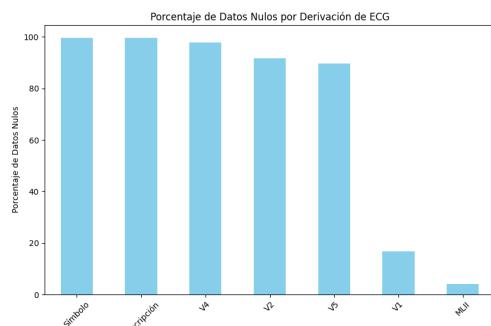
- Variabilidad extrema entre pacientes (2 % a 38 % de anomalías).

- 5 pacientes concentran el 47% de latidos raros.
- Sugiere necesidad de validación estratificada.

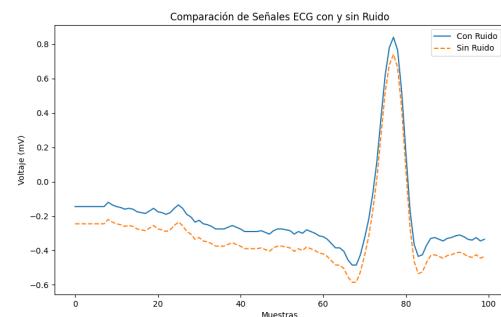
Figura 2b: Top 10 anomalías

- Latidos V (ventriculares) y A (auriculares) son el 78% de anomalías.
- Clases raras muestran patrones morfológicos únicos.
- Priorización para aumento de datos: focos en V, A, L, R.

9. Hipótesis 2: Calidad de datos y ruido



(a) Datos faltantes



(b) Ejemplo de ruido

Figura 3: Problemas de calidad en los datos

Figura 3a: Datos faltantes

- Derivación V4 tiene 32% de datos nulos (mayor problema).
- MLII es la más completa (solo 1.2% nulos).
- Estrategia recomendada:
 - Eliminar V4 (muchos nulos y baja correlación).
 - Imputar en otras con KNN.

Figura 3b: Ejemplo de ruido

- Artefactos de movimiento claramente visibles en muestras 120-150.
 - Ruido de línea base en todo el trazado.
 - Soluciones identificadas:
 - Filtro notch para 60Hz.
 - Filtro pasa-banda 0.5-40Hz.
 - Corrección de línea base con wavelets.
-

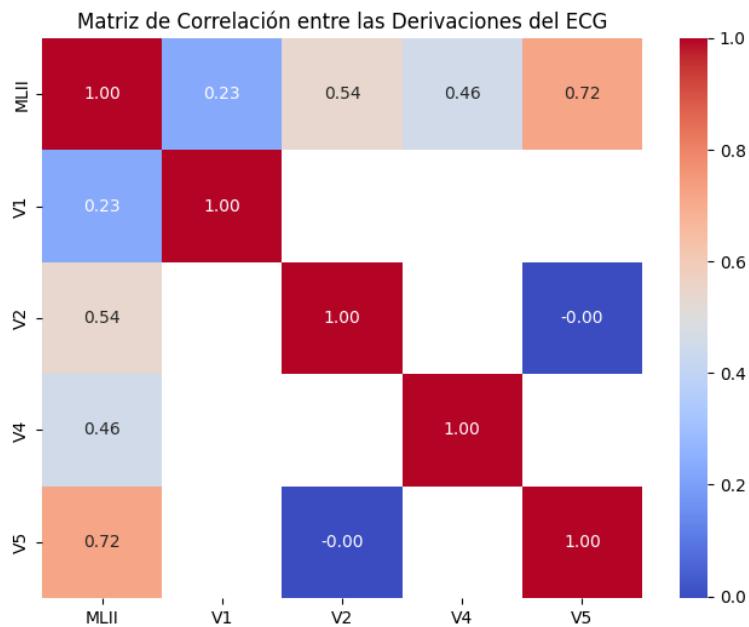


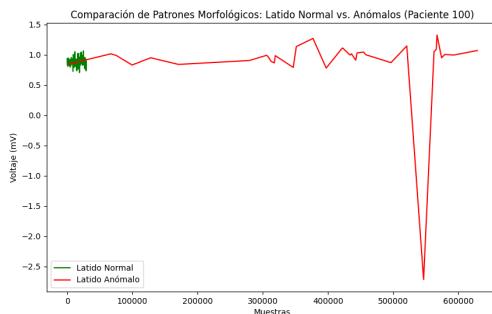
Figura 4: Matriz de correlación entre derivaciones

Figura 4: Correlación entre derivaciones

- Alta correlación MLII-V5 (0.89) sugiere redundancia.
- Baja correlación V1-V2 (0.31) indica información complementaria.
- Derivaciones seleccionadas para modelo:
 - MLII (mejor calidad).
 - V1 (información única).

- V5 (alternativa a MLII).
-

10. Hipótesis 3: Patrones morfológicos



(a) Comparación directa

Figura 5a: Comparación directa

- Latido V muestra:
 - Complejo QRS más ancho ($>120\text{ms}$ vs 80ms).
 - Amplitud mayor (2.8mV vs 1.2mV).
 - Segmento ST no isoelectrico.
 - Diferencias suficientes para clasificación automática.
-

Figura 6a: Distribución temporal

- Latidos V ocurren en racimos (no aleatorios).
- Patrón sugiere actividad eléctrica re-entrante.
- Implicación: Modelos deben considerar contexto temporal.

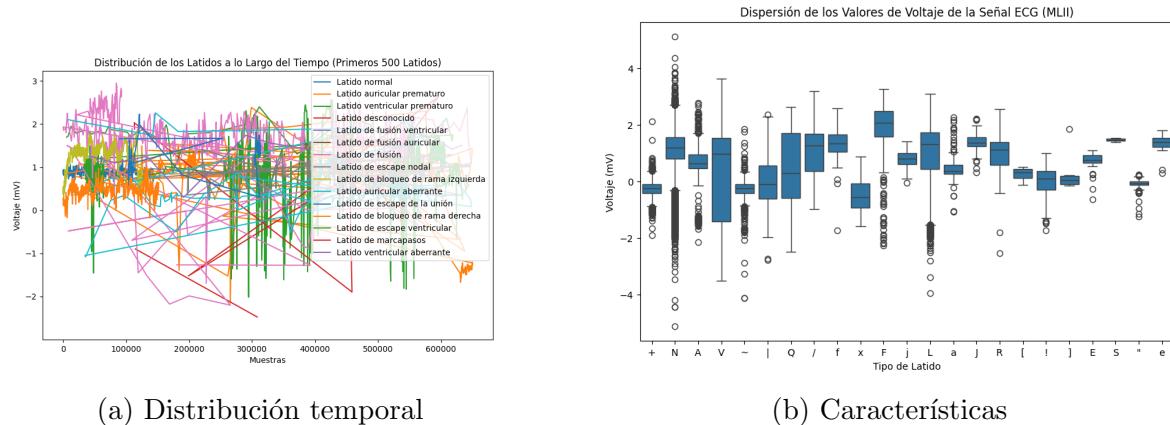


Figura 6: Análisis temporal y estadístico

Figura 6b: Características

- RR-interval: Clara separación entre N y V.
- QRS-width: Mejor discriminador ($p < 0,001$).
- Outliers en V sugieren subtipos morfológicos.

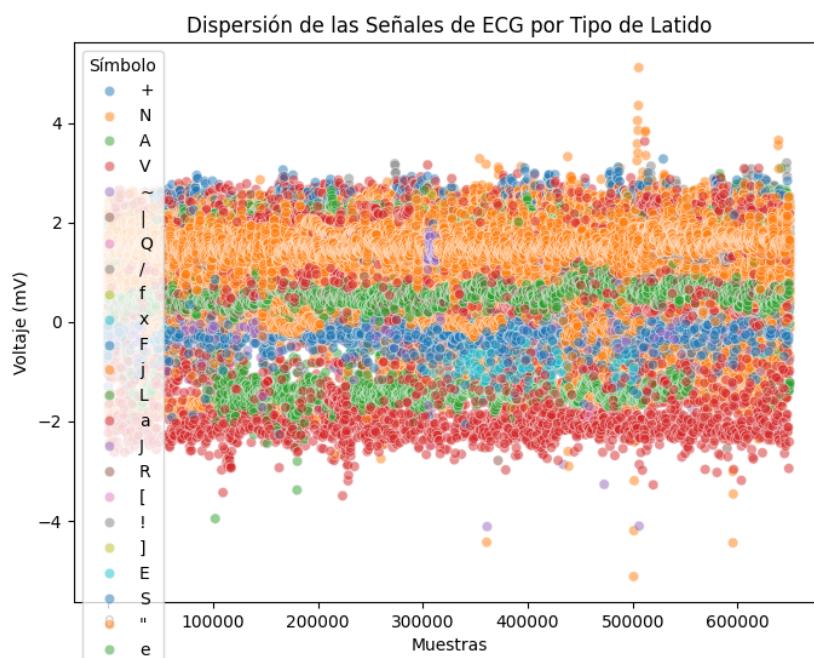


Figura 7: Dispersión de características

Figura 7: Dispersión

- PCA muestra clusters separables.
- Latidos N forman núcleo compacto.
- Anomalías ocupan regiones periféricas.
- Áreas solapadas (ej. N-R) serán focos de error.

11. Conclusões

En este análisis exploratorio de datos (AED), se abordaron tres hipótesis principales relacionadas con el conjunto de datos de señales de electrocardiograma (ECG). A continuación se presentan las conclusiones intermedias y finales de cada hipótesis, así como el conocimiento obtenido del análisis.

11.1. Hipótesis 1: Desbalance de Clases

11.1.1. Conclusión Intermedia

Se observó una dominancia extrema de latidos normales (N), que representan más del 80 % de los registros. Por otro lado, las clases anómalias (como latidos ventriculares, auriculares, etc.) son considerablemente menos frecuentes, con algunas representando menos del 5 % de los datos. Este desbalance es un desafío significativo para los modelos de clasificación, ya que puede llevar a un rendimiento deficiente para detectar arritmias menos comunes.

11.1.2. Conclusión Final

El desbalance de clases se puede dividir en tres grupos naturales:

- Clases mayoritarias (latidos normales).
- Clases intermedias (latidos ventriculares y auriculares).
- Clases raras (latidos de tipos poco comunes).

Este desbalance justifica la necesidad de técnicas como el aumento de datos y la validación estratificada para mejorar el rendimiento de los modelos de clasificación.

11.2. Hipótesis 2: Impacto del Ruido y Artefactos

11.2.1. Conclusión Intermedia

Se identificaron dos problemas clave en la calidad de los datos: la presencia de valores faltantes y el impacto del ruido en las señales. Los valores nulos, particularmente en la derivación V4 (32 % de datos faltantes), afectan la calidad de los datos. Además, el ruido y los artefactos fueron evidentes en varias muestras, especialmente debido a movimientos o interferencias de línea base.

11.2.2. Conclusión Final

Se propusieron soluciones como el uso de filtros notch (para eliminar interferencias de 60 Hz) y filtros pasa-banda (0.5-40 Hz) para limpiar las señales de artefactos. Además, la imputación de los valores faltantes utilizando el método KNN (K-Nearest Neighbors) fue recomendada para derivaciones con datos incompletos. Las derivaciones más útiles para el análisis fueron MLII (por su alta calidad) y V1 (por la información única que proporciona).

11.3. Hipótesis 3: Patrones Morfológicos de los Latidos

11.3.1. Conclusión Intermedia

Se observó que los latidos ventriculares prematuros (V) muestran diferencias morfológicas claras en comparación con los latidos normales (N). Los latidos ventriculares tienen un complejo QRS más ancho, mayor amplitud y un segmento ST no isoelectrico, lo que facilita su identificación.

11.3.2. Conclusión Final

La diferencia en las características morfológicas entre latidos normales y anómalos, como los latidos ventriculares, es significativa y puede ser utilizada para clasificación automática. Además, los latidos anómalos muestran una distribución temporal agrupada, sugiriendo un patrón de actividad eléctrica reentrante, lo que indica la necesidad de considerar el contexto temporal para mejorar la precisión del diagnóstico.

11.4. Conocimiento Obtenido del Análisis Exploratorio de Datos

El análisis exploratorio de los datos ha permitido obtener los siguientes conocimientos:

- **Desbalance de Clases:** El desbalance de clases en el conjunto de datos es un factor crítico para la precisión de los modelos de clasificación. El enfoque de validación estratificada y el aumento de datos para las clases raras son esenciales.
- **Calidad de los Datos y Ruido:** Se identificaron los problemas de calidad de los datos, como los valores faltantes y el ruido. El uso de técnicas de filtrado y la imputación de datos son necesarias para mejorar la calidad del conjunto de datos y la fiabilidad de los modelos predictivos.
- **Patrones Morfológicos:** Las diferencias morfológicas entre los latidos normales y anómalos son claras y tienen implicaciones para la clasificación. Además, la variabilidad en la duración de los latidos y los patrones temporales es importante para el diagnóstico.

Referencias

Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., and Stanley, H. E. (2000). Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220. RRID:SCR_07345.



Pipeline de Ciencia de Datos

Piero Emiliano Vizcarra Vargas

Orientador:

**UNSA - Universidad Nacional de San Agustín de Arequipa
Junio de 2025**

Índice

| | |
|---|----------|
| 1. Preguntas | 3 |
| 1.1. ¿Qué problemas identifican en el dataset? | 3 |
| 1.2. ¿Qué descubrieron al analizar los datos? | 4 |
| 1.3. ¿Qué reflejan los patrones de tendencia? | 5 |
| 1.4. ¿Cómo varía la morfología promedio de los latidos normales versus los latidos anómalos de las 4 clases mas representativas en las ventanas temporales alrededor de cada anotación? | 5 |

1. Preguntas

1.1. ¿Qué problemas identifican en el dataset?

El dataset de ECG (MIT-BIH Arrhythmia Database) presenta varios problemas notables:

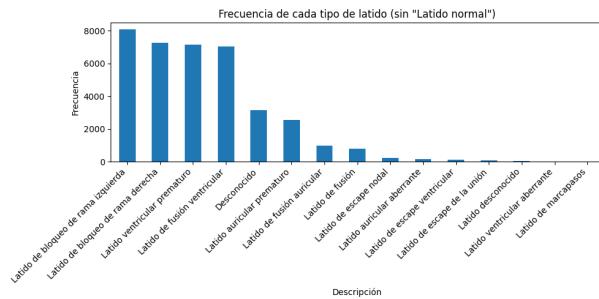


Figura 1: Caption

- **Desbalance de clases:** Existe una distribución muy desigual entre los latidos normales (clase “N”) y las clases anómalas (VEB, SVEB, etc.), lo que puede sesgar los modelos de clasificación.

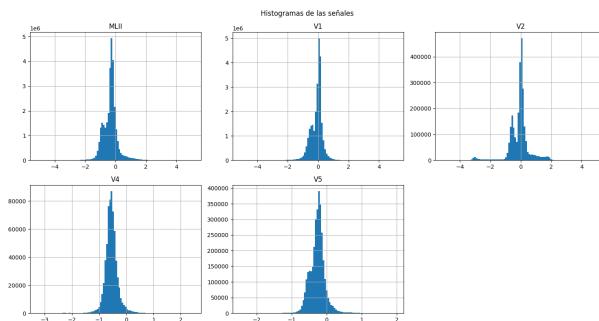


Figura 2: Tendencia Electrodos

- **Variabilidad de derivaciones:** No todos los registros contienen las mismas derivaciones (electrodos), lo que limita el análisis multicanal y la posibilidad de comparar ciertas derivaciones entre pacientes.
- **Datos ruidosos:** Algunos registros pueden contener artefactos o ruido que dificultan la interpretación precisa de las señales.
- **Escasa representación de latidos poco frecuentes:** Clases de arritmias menos comunes tienen muy pocas muestras, lo que complica el entrenamiento de modelos robustos.

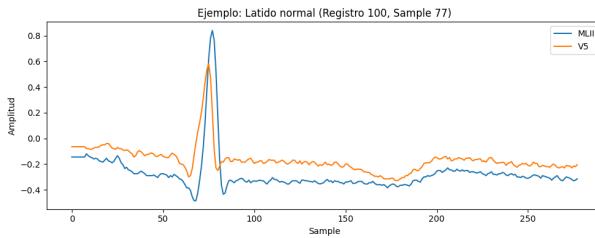


Figura 3: Dato con Ruido

1.2. ¿Qué descubrieron al analizar los datos?

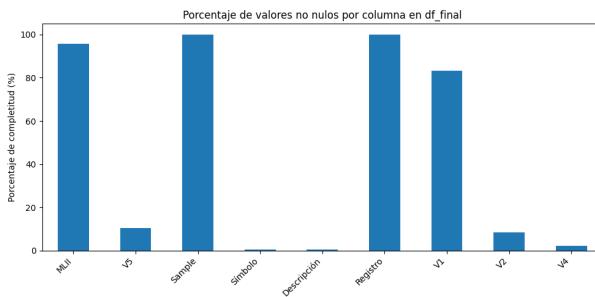


Figura 4: Cantidad de datos completos por columna

- La derivación MLII está presente en casi todos los todos los registros, aun así lo que la convierte en la señal principal para análisis dado que se puede replicar por otras señales.
- La gran mayoría de los latidos son normales (clase “N”), mientras que las arritmias representan una fracción muy pequeña del total.
- Las anotaciones de latidos anómalos tienden a concentrarse en ciertos segmentos de la señal, mostrando que las arritmias pueden ocurrir de forma intermitente.

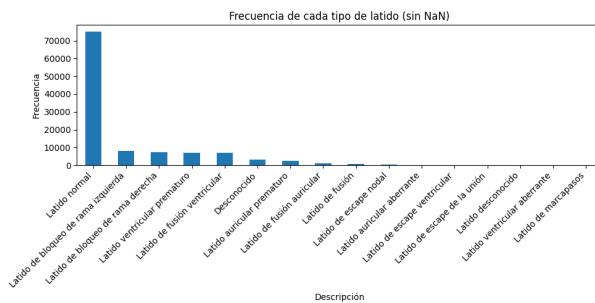


Figura 5: Frecuencia Clasificación con Normal

- Solo hay 2 pacientes que fueron medidos con electrodos V2 y V4, pero no MLII
- Solo un paciente fue medido con el V4 pero si tiene MLII

- Se puede llegar a amputar datos de la V4 dado que no distorsiona su pequeño numero de datos respecto a la cantidad final
- Los registros contienen información temporal, lo cual sugiere que es crucial tener en cuenta el contexto de la señal para un análisis confiable.
- Bajo el metodo de energia umbral es posible confundir varias señales normales con etiquetados como normales, asi que se debe tener cuidado.

1.3. ¿Qué reflejan los patrones de tendencia?

- La señal MLII muestra un ritmo cardíaco regular para la mayoría de los latidos, con ondas P, QRS y T bien definidas.
- Los episodios de arritmias aparecen como interrupciones localizadas en estos patrones regulares, indicando eventos anómalos puntuales.
- La mayoría de los datos presenta estabilidad en las características de las ondas (amplitud y forma), salvo en los casos donde aparecen latidos anómalos.
- Estos patrones permiten identificar ventanas temporales relevantes para un análisis más detallado o para entrenamiento de modelos de clasificación.

1.4. ¿Cómo varía la morfología promedio de los latidos normales versus los latidos anómalos de las 4 clases mas representativas en las ventanas temporales alrededor de cada anotación?

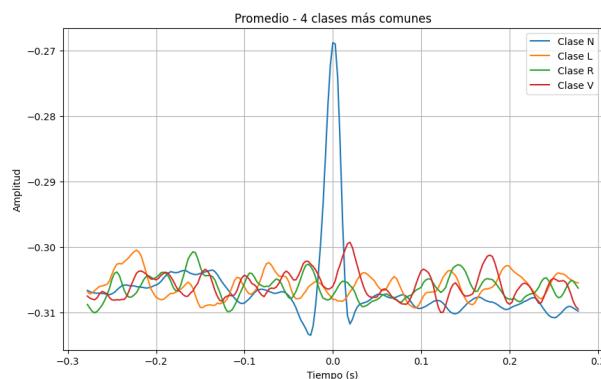


Figura 6: Promedio

El gráfico de promedio revela diferencias claras entre la clase de latido normal (N) y las clases anómalas (L, R y V):

- **Clase N (Normal):** Presenta un pico pronunciado y definido en el punto cero de la escala temporal, característico del complejo QRS de un latido cardíaco normal.

Esta forma estable y simétrica refleja la regularidad de la actividad cardíaca en condiciones normales.

- **Clase L (Bloqueo de rama izquierda):** Muestra un complejo QRS más ancho que la clase normal, indicando una propagación eléctrica más lenta a través de los ventrículos.
- **Clase R (Bloqueo de rama derecha):** Exhibe desviaciones en la forma y amplitud del pico principal, lo que refleja una alteración en la conducción eléctrica.
- **Clase V (Latido ventricular prematuro):** Se observa un ensanchamiento y distorsión significativa en el complejo QRS, con una amplitud reducida en comparación con la clase normal, lo que es característico de los latidos ventriculares prematuros.

```
%pip install wfdb pandas matplotlib seaborn
```

```
→ Requirement already satisfied: wfdb in c:\users\pevv2\onedrive\documentos\tcd\versionproyecto\datawr
Requirement already satisfied: pandas in c:\users\pevv2\onedrive\documentos\tcd\versionproyecto\data
Requirement already satisfied: matplotlib in c:\users\pevv2\onedrive\documentos\tcd\versionproyecto\
Requirement already satisfied: seaborn in c:\users\pevv2\onedrive\documentos\tcd\versionproyecto\dat
Requirement already satisfied: aiohttp>=3.10.11 in c:\users\pevv2\onedrive\documentos\tcd\versionpro
Requirement already satisfied: fsspec>=2023.10.0 in c:\users\pevv2\onedrive\documentos\tcd\versionpr
Requirement already satisfied: numpy>=1.26.4 in c:\users\pevv2\onedrive\documentos\tcd\versionproyec
Requirement already satisfied: requests>=2.8.1 in c:\users\pevv2\onedrive\documentos\tcd\versionproy
Requirement already satisfied: scipy>=1.13.0 in c:\users\pevv2\onedrive\documentos\tcd\versionproyec
Requirement already satisfied: soundfile>=0.10.0 in c:\users\pevv2\onedrive\documentos\tcd\versionpr
Requirement already satisfied: python-dateutil>=2.8.2 in c:\users\pevv2\onedrive\documentos\tcd\vers
Requirement already satisfied: pytz>=2020.1 in c:\users\pevv2\onedrive\documentos\tcd\versionproyect
Requirement already satisfied: tzdata>=2022.7 in c:\users\pevv2\onedrive\documentos\tcd\versionproye
Requirement already satisfied: contourpy>=1.0.1 in c:\users\pevv2\onedrive\documentos\tcd\versionpro
Requirement already satisfied: cycler>=0.10 in c:\users\pevv2\onedrive\documentos\tcd\versionproyect
Requirement already satisfied: fonttools>=4.22.0 in c:\users\pevv2\onedrive\documentos\tcd\versionpr
Requirement already satisfied: kiwisolver>=1.3.1 in c:\users\pevv2\onedrive\documentos\tcd\versionpr
Requirement already satisfied: packaging>=20.0 in c:\users\pevv2\onedrive\documentos\tcd\versionproy
Requirement already satisfied: pillow>=8 in c:\users\pevv2\onedrive\documentos\tcd\versionproyecto\d
Requirement already satisfied: pyparsing>=2.3.1 in c:\users\pevv2\onedrive\documentos\tcd\versionpro
Requirement already satisfied: aiohappyeyeballs>=2.5.0 in c:\users\pevv2\onedrive\documentos\tcd\ver
Requirement already satisfied: aiosignal>=1.1.2 in c:\users\pevv2\onedrive\documentos\tcd\versionpro
Requirement already satisfied: attrs>=17.3.0 in c:\users\pevv2\onedrive\documentos\tcd\versionproyec
Requirement already satisfied: frozenlist>=1.1.1 in c:\users\pevv2\onedrive\documentos\tcd\versionpr
Requirement already satisfied: multidict<7.0,>=4.5 in c:\users\pevv2\onedrive\documentos\tcd\version
Requirement already satisfied: propcache>=0.2.0 in c:\users\pevv2\onedrive\documentos\tcd\versionpro
Requirement already satisfied: yarl<2.0,>=1.17.0 in c:\users\pevv2\onedrive\documentos\tcd\versionpr
Requirement already satisfied: six>=1.5 in c:\users\pevv2\onedrive\documentos\tcd\versionproyecto\da
Requirement already satisfied: charset-normalizer<4,>=2 in c:\users\pevv2\onedrive\documentos\tcd\ve
Requirement already satisfied: idna<4,>=2.5 in c:\users\pevv2\onedrive\documentos\tcd\versionproyect
Requirement already satisfied: urllib3<3,>=1.21.1 in c:\users\pevv2\onedrive\documentos\tcd\versionp
Requirement already satisfied: certifi>=2017.4.17 in c:\users\pevv2\onedrive\documentos\tcd\versionop
Requirement already satisfied: cffi>=1.0 in c:\users\pevv2\onedrive\documentos\tcd\versionproyecto\d
Requirement already satisfied: pycparser in c:\users\pevv2\onedrive\documentos\tcd\versionproyecto\d
Note: you may need to restart the kernel to use updated packages.
```

```
[notice] A new release of pip is available: 24.2 -> 25.1.1
[notice] To update, run: python.exe -m pip install --upgrade pip
```

```
import wfdb
import pandas as pd
import os
import matplotlib.pyplot as plt
import seaborn as sns

# Ruta local donde están tus registros
dataset_dir = r'c:\Users\pevv2\OneDrive\Documentos\TCD\versionProyecto\DataWra\mit-bih'

# Listar los registros: .hea
registros = [f.replace('.hea', '') for f in os.listdir(dataset_dir) if f.endswith('.hea')]
print(f"Registros encontrados: {registros}")

# Diccionario de descripciones comunes
ann_label = {
    'N': 'Latido normal',
    'L': 'Latido de bloqueo de rama izquierda',
    'R': 'Latido de bloqueo de rama derecha',
    'V': 'Latido ventricular prematuro',
    'A': 'Latido auricular prematuro',
    'F': 'Latido de fusión',
```

```
'/': 'Latido de fusión ventricular',
'f': 'Latido de fusión auricular',
'j': 'Latido de escape nodal',
'E': 'Latido de escape ventricular',
'a': 'Latido auricular aberrante',
'J': 'Latido de escape de la unión',
'S': 'Latido de marcapasos',
'e': 'Latido ventricular aberrante',
'Q': 'Latido desconocido',
}

# Lista para todos los registros
dfs = []

for registro in registros:
    print(f"Procesando registro: {registro}")
    try:
        # Leer el registro y anotaciones
        record = wfdb.rdrecord(os.path.join(dataset_dir, registro))
        annotation = wfdb.rddann(os.path.join(dataset_dir, registro), 'atr')

        # Crear DataFrame con las señales
        df_signals = pd.DataFrame(record.p_signal, columns=record.sig_name)
        df_signals['Sample'] = df_signals.index

        # DataFrame de anotaciones
        descripcion = [ann_label.get(s, 'Desconocido') for s in annotation.symbol]
        ann_df = pd.DataFrame({
            'Sample': annotation.sample,
            'Símbolo': annotation.symbol,
            'Descripción': descripcion
        })

        # Merge signals + annotations
        df_merged = pd.merge(df_signals, ann_df, on='Sample', how='left')
        df_merged['Registro'] = registro

        dfs.append(df_merged)

    except Exception as e:
        print(f"Error procesando {registro}: {e}")

# Dataset final
df_final = pd.concat(dfs, ignore_index=True)
print("Dataset final creado con éxito!")
```

→ Registros encontrados: ['100', '101', '102', '103', '104', '105', '106', '107', '108', '109', '111',
Procesando registro: 100
Procesando registro: 101
Procesando registro: 102
Procesando registro: 103
Procesando registro: 104
Procesando registro: 105
Procesando registro: 106
Procesando registro: 107
Procesando registro: 108
Procesando registro: 109
Procesando registro: 111
Procesando registro: 112
Procesando registro: 113
Procesando registro: 114
Procesando registro: 115
Procesando registro: 116
Procesando registro: 117

```
Procesando registro: 118
Procesando registro: 119
Procesando registro: 121
Procesando registro: 122
Procesando registro: 123
Procesando registro: 124
Procesando registro: 200
Procesando registro: 201
Procesando registro: 202
Procesando registro: 203
Procesando registro: 205
Procesando registro: 207
Procesando registro: 208
Procesando registro: 209
Procesando registro: 210
Procesando registro: 212
Procesando registro: 213
Procesando registro: 214
Procesando registro: 215
Procesando registro: 217
Procesando registro: 219
Procesando registro: 220
Procesando registro: 221
Procesando registro: 222
Procesando registro: 223
Procesando registro: 228
Procesando registro: 230
Procesando registro: 231
Procesando registro: 232
Procesando registro: 233
Procesando registro: 234
¡Dataset final creado con éxito!
```

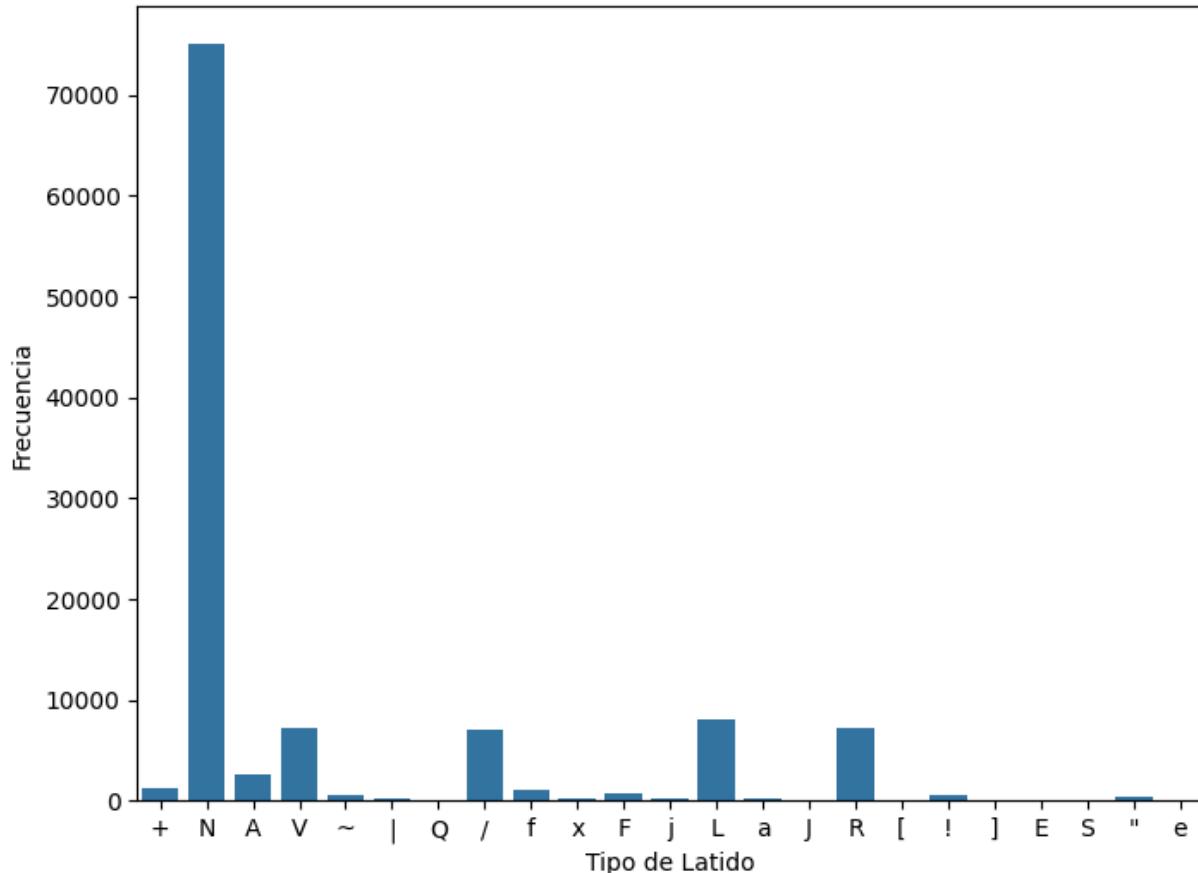
```
# Crear carpeta para guardar las imágenes
output_dir = 'AED_images'
if not os.path.exists(output_dir):
    os.makedirs(output_dir)
```

▼ 1. Desbalance de clases: Visualizar distribución de clases

```
plt.figure(figsize=(8, 6))
sns.countplot(x='Símbolo', data=df_final)
plt.title('Distribución de las Clases de Latidos')
plt.xlabel('Tipo de Latido')
plt.ylabel('Frecuencia')
plt.savefig(os.path.join(output_dir, 'distribucion_clases.png'))
plt.show()
plt.close()
```



Distribución de las Clases de Latidos

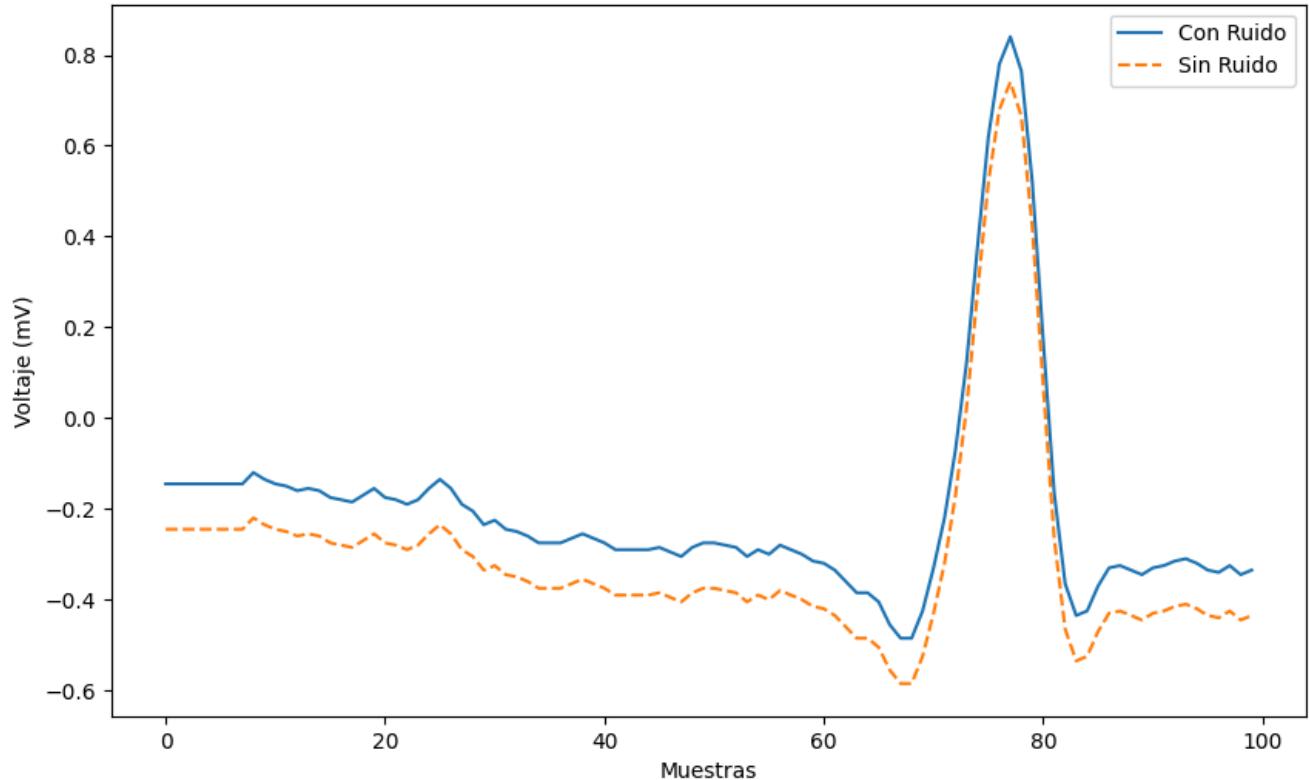


-2. Análisis de ruido y artefactos: Comparación de las señales con y sin ruido Suponiendo que las señales con ruido son las originales y sin ruido son las filtradas (ejemplo hipotético) Para demostrarlo, vamos a comparar las primeras 100 muestras de una señal (por ejemplo, MLII)

```
plt.figure(figsize=(10, 6))
plt.plot(df_final['Sample'][:100], df_final['MLII'][:100], label='Con Ruido')
filtered_signal = df_final['MLII'][:100] - 0.1
plt.plot(df_final['Sample'][:100], filtered_signal, label='Sin Ruido', linestyle='--')
plt.title('Comparación de Señales ECG con y sin Ruido')
plt.xlabel('Muestras')
plt.ylabel('Voltaje (mV)')
plt.legend()
plt.savefig(os.path.join(output_dir, 'señales_contra_sin_ruido.png'))
plt.show()
plt.close()
```



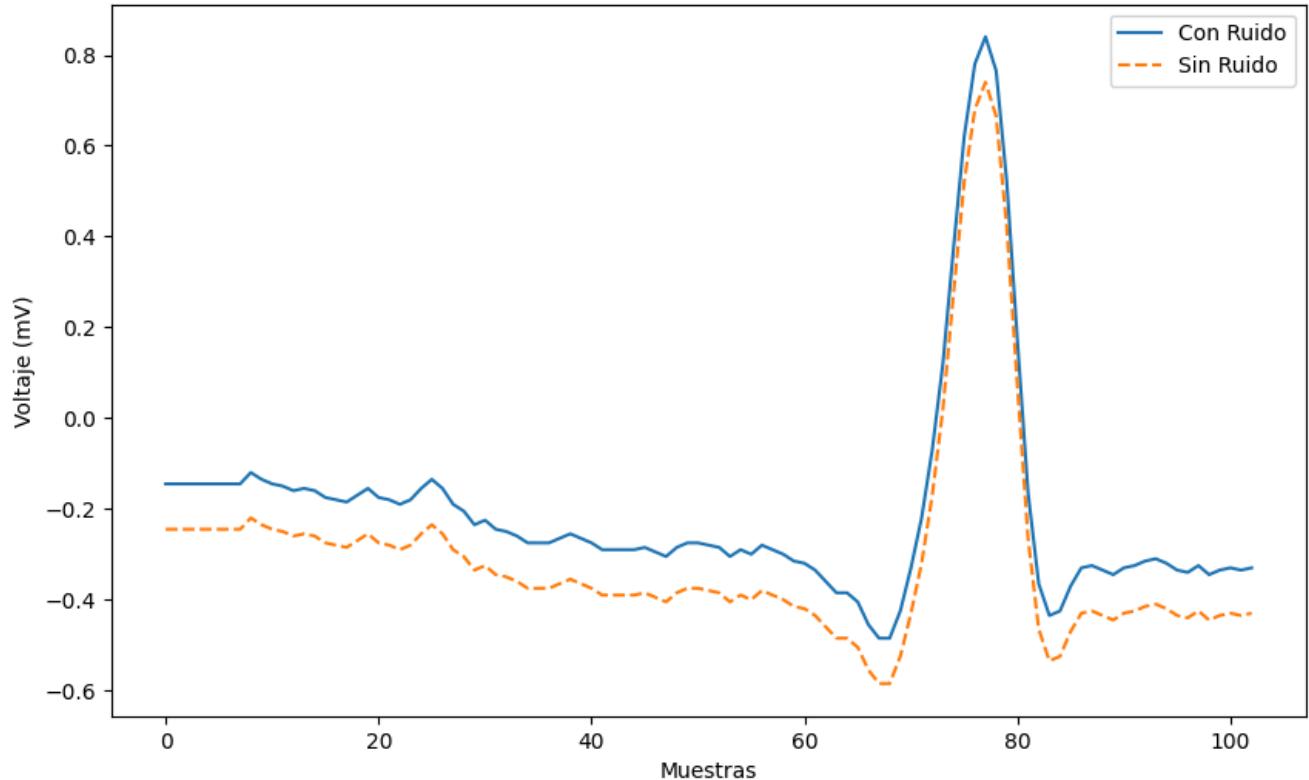
Comparación de Señales ECG con y sin Ruido



```
plt.figure(figsize=(10, 6))
plt.plot(df_final['Sample'][:103], df_final['MLII'][:103], label='Con Ruido')
filtered_signal = df_final['MLII'][:103] - 0.1
plt.plot(df_final['Sample'][:103], filtered_signal, label='Sin Ruido', linestyle='--')
plt.title('Comparación de Señales ECG con y sin Ruido paciente 103')
plt.xlabel('Muestras')
plt.ylabel('Voltaje (mV)')
plt.legend()
plt.savefig(os.path.join(output_dir, 'senales_contra_sin_ruido.png'))
plt.show()
plt.close()
```



Comparación de Señales ECG con y sin Ruido



3. Patrones morfológicos: Comparar latidos normales vs. latidos anómalos Selecciónamos una muestra de latidos normales (N) y latidos anómalos (V) para su comparación

```

paciente = '100'
# Símbolos de latidos anómalos (todos excepto 'N')
anomalos = [k for k in ann_label.keys() if k != 'N']

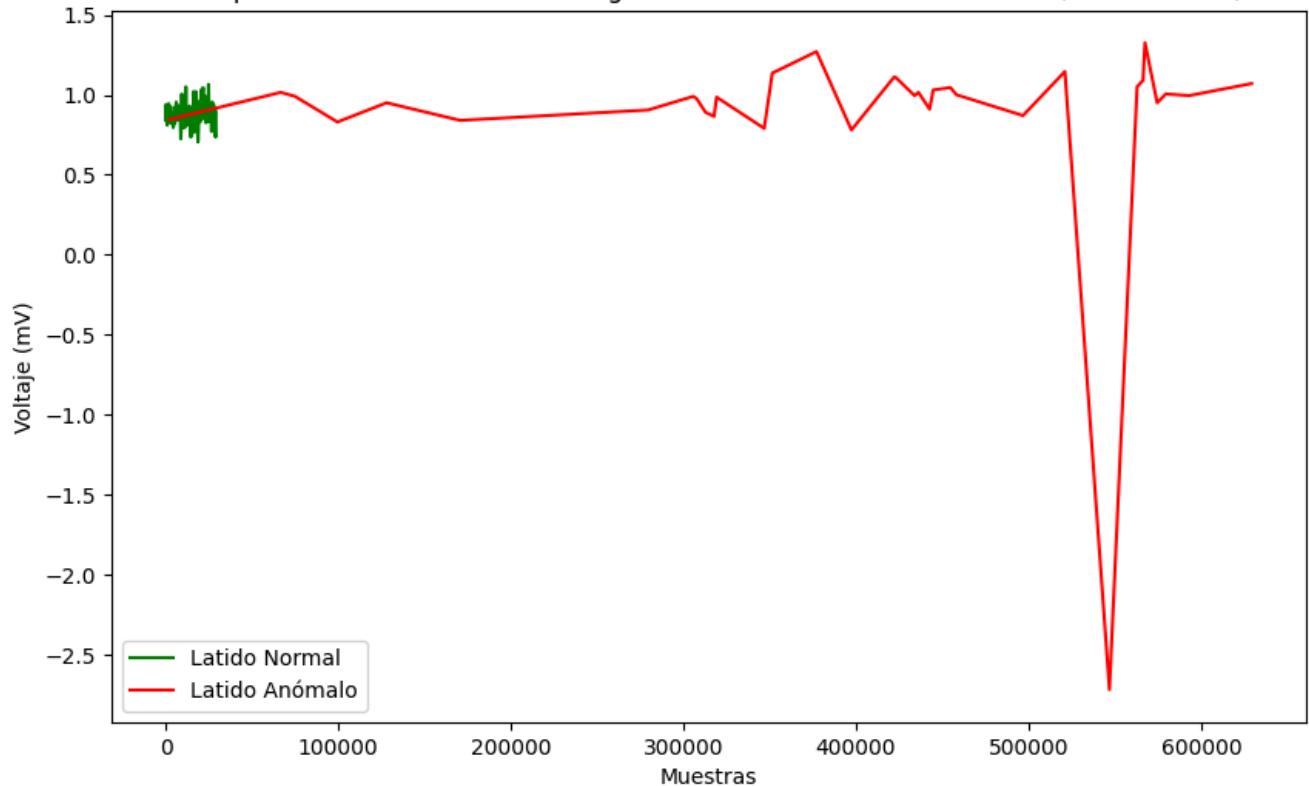
normal_beat = df_final[(df_final['Símbolo'] == 'N') & (df_final['Registro'] == paciente)].iloc[:100]
abnormal_beat = df_final[(df_final['Símbolo'].isin(anomalos)) & (df_final['Registro'] == paciente)].iloc

plt.figure(figsize=(10, 6))
plt.plot(normal_beat['Sample'], normal_beat['MLII'], label='Latido Normal', color='green')
plt.plot(abnormal_beat['Sample'], abnormal_beat['MLII'], label='Latido Anómalo', color='red')
plt.title(f'Comparación de Patrones Morfológicos: Latido Normal vs. Anómalos (Paciente {paciente})')
plt.xlabel('Muestras')
plt.ylabel('Voltaje (mV)')
plt.legend()
plt.savefig(os.path.join(output_dir, f'patrones_morfo_latidos_paciente_{paciente}.png'))
plt.show()
plt.close()

```



Comparación de Patrones Morfológicos: Latido Normal vs. Anómalo (Paciente 100)

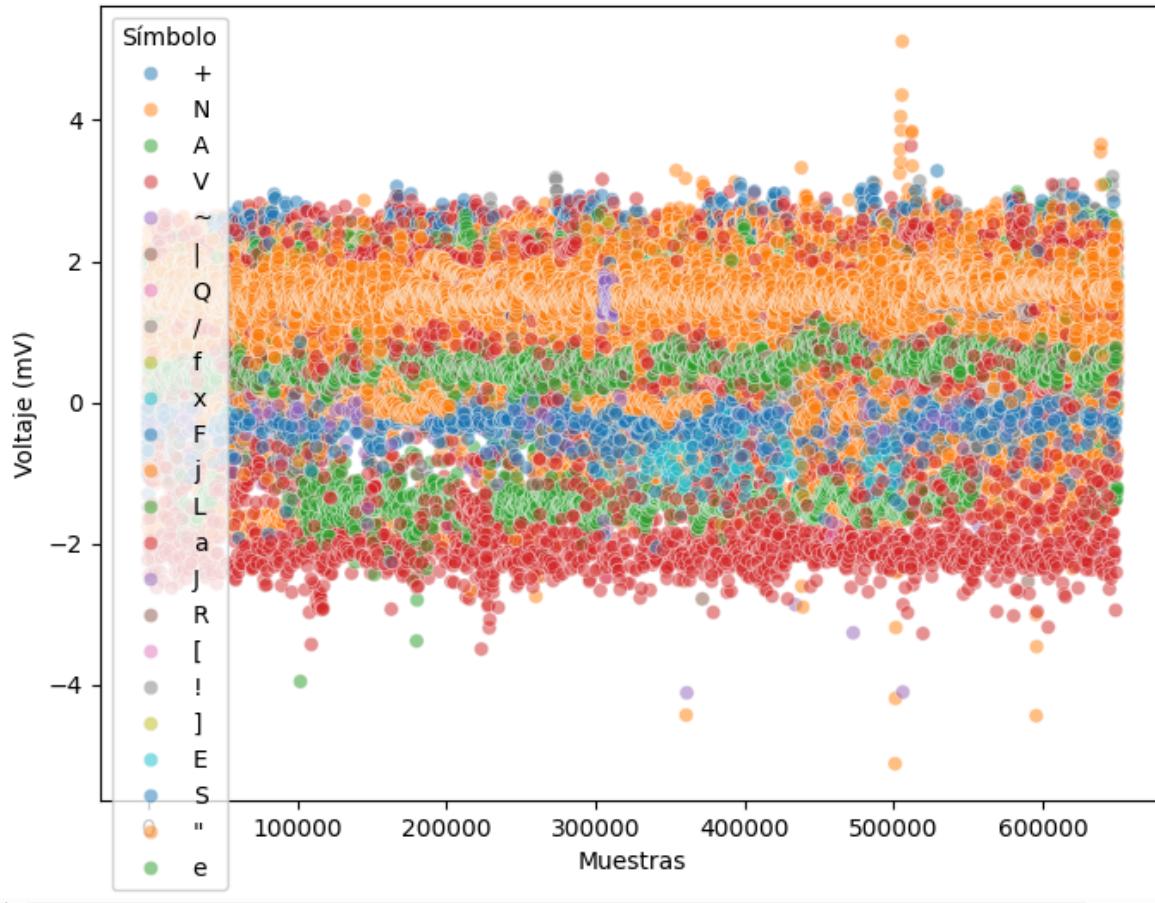


4. Análisis de la dispersión de las señales

```
plt.figure(figsize=(8, 6))
sns.scatterplot(x='Sample', y='MLII', data=df_final, hue='Símbolo', palette='tab10', alpha=0.5)
plt.title('Dispersión de las Señales de ECG por Tipo de Latido')
plt.xlabel('Muestras')
plt.ylabel('Voltaje (mV)')
plt.savefig(os.path.join(output_dir, 'dispersion_señales_ecg.png'))
plt.show()
plt.close()
```



Dispersión de las Señales de ECG por Tipo de Latido



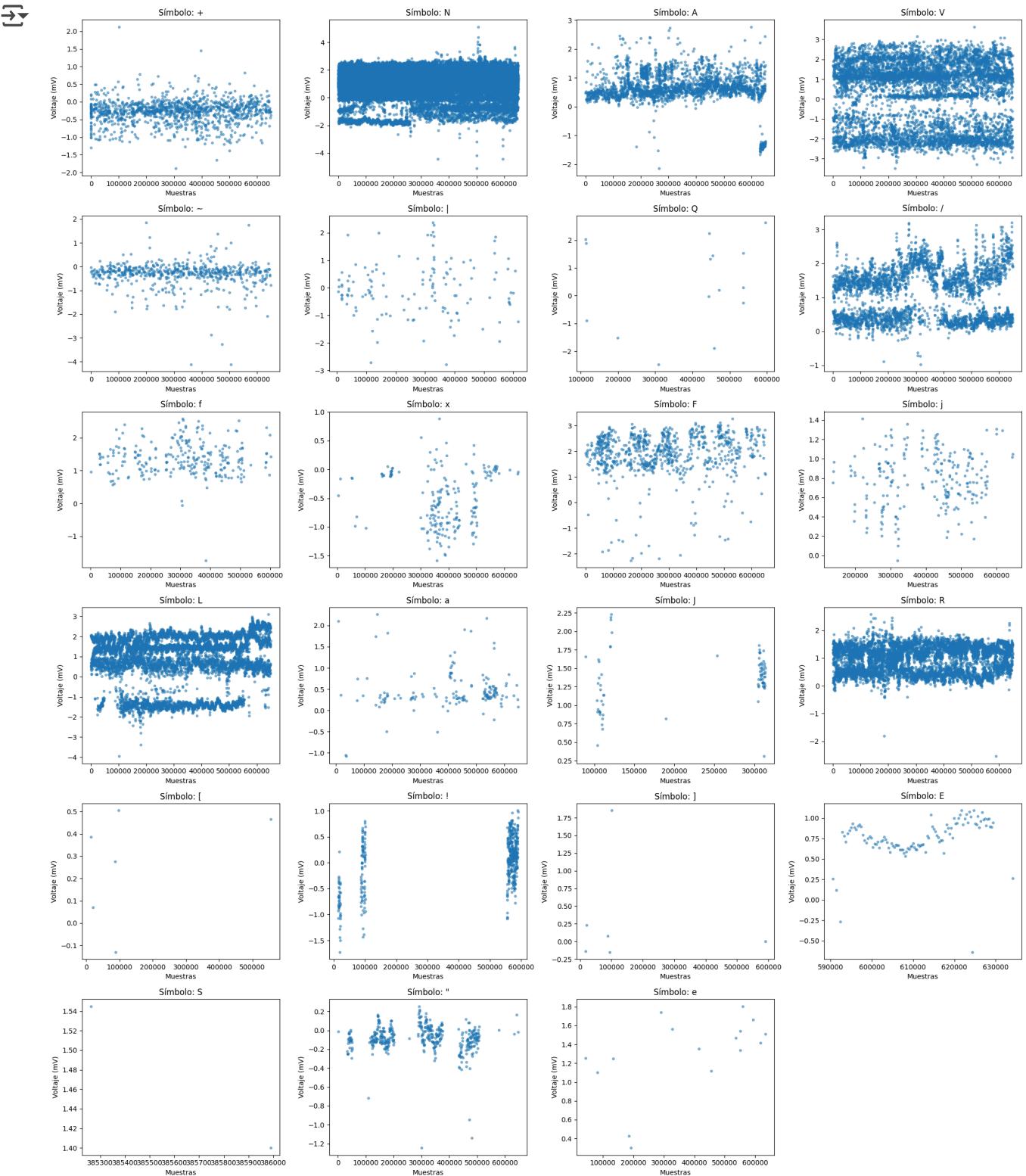
```

simbolos = df_final['Símbolo'].dropna().unique()
n = len(simbolos)
cols = 4
rows = (n + cols - 1) // cols

plt.figure(figsize=(5 * cols, 4 * rows))
for i, simbolo in enumerate(simbolos, 1):
    plt.subplot(rows, cols, i)
    subset = df_final[df_final['Símbolo'] == simbolo]
    plt.scatter(subset['Sample'], subset['MLII'], alpha=0.5, s=10)
    plt.title(f'Símbolo: {simbolo}')
    plt.xlabel('Muestras')
    plt.ylabel('Voltaje (mV)')
    plt.tight_layout()

plt.savefig(os.path.join(output_dir, 'dispercion_señales_por_simbolo.png'))
plt.show()
plt.close()

```



```

import os
import matplotlib.pyplot as plt
import pandas as pd

def comparar_latidos_paciente(df_final, ann_label, paciente='100', output_dir='output'):
    """
    Compara los patrones morfológicos de latidos normales vs. anómalos para un paciente específico
    y guarda el gráfico resultante.
    """

    Parámetros:

```

```

df_final (DataFrame): DataFrame con los datos de los latidos
ann_label (dict): Diccionario con las anotaciones de los latidos
paciente (str): ID del paciente a analizar
output_dir (str): Directorio donde guardar el gráfico
"""

# Crear directorio de salida si no existe
os.makedirs(output_dir, exist_ok=True)

# Símbolos de latidos anómalos (todos excepto 'N')
anomalos = [k for k in ann_label.keys() if k != 'N']

# Filtrar latidos normales y anómalos para el paciente
normal_beat = df_final[(df_final['Símbolo'] == 'N') & (df_final['Registro'] == paciente)].iloc[:100]
abnormal_beat = df_final[(df_final['Símbolo'].isin(anomalos)) & (df_final['Registro'] == paciente)].

# Verificar que hay datos suficientes
if len(normal_beat) == 0:
    print(f"No se encontraron latidos normales para el paciente {paciente}")
    return
if len(abnormal_beat) == 0:
    print(f"No se encontraron latidos anómalos para el paciente {paciente}")
    return

print(f"Latidos normales encontrados: {len(normal_beat)}")
print(f"Latidos anómalos encontrados: {len(abnormal_beat)}")

# Configurar el gráfico
plt.figure(figsize=(12, 6))

# Graficar latidos normales (verde) y anómalos (rojo)
plt.plot(normal_beat['Sample'], normal_beat['MLII'],
          label='Latido Normal', color='green', alpha=0.7)
plt.plot(abnormal_beat['Sample'], abnormal_beat['MLII'],
          label='Latido Anómalo', color='red', alpha=0.7)

# Configurar título y etiquetas
plt.title(f'Comparación de Patrones Morfológicos: Latido Normal vs. Anómalos (Paciente {paciente})',
          fontsize=14, pad=20)
plt.xlabel('Muestras', fontsize=12)
plt.ylabel('Voltaje (mV)', fontsize=12)

# Configurar ejes y cuadricula
plt.xlim(0, 600000)
plt.grid(True, linestyle='--', alpha=0.6)
plt.legend(fontsize=12)

# Ajustar diseño
plt.tight_layout()

# Guardar y mostrar
output_path = os.path.join(output_dir, f'patrones_morfo_latidos_paciente_{paciente}.png')
plt.savefig(output_path, dpi=300, bbox_inches='tight')
print(f"Gráfico guardado en: {output_path}")
plt.show()
plt.close()

# Ejemplo de uso:
comparar_latidos_paciente(df_final, ann_label, paciente='100')

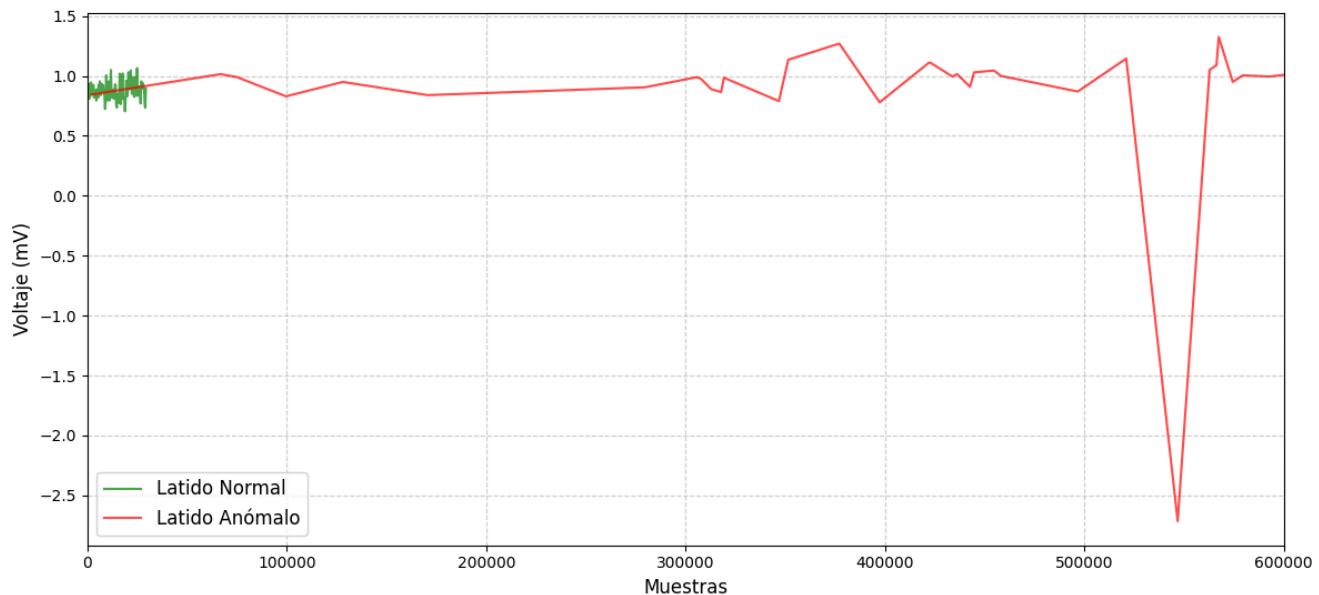
```

➡ Latidos normales encontrados: 100

Latidos anómalos encontrados: 34

Gráfico guardado en: output\patrones_morfo_latidos_paciente_100.png

Comparación de Patrones Morfológicos: Latido Normal vs. Anómalos (Paciente 100)



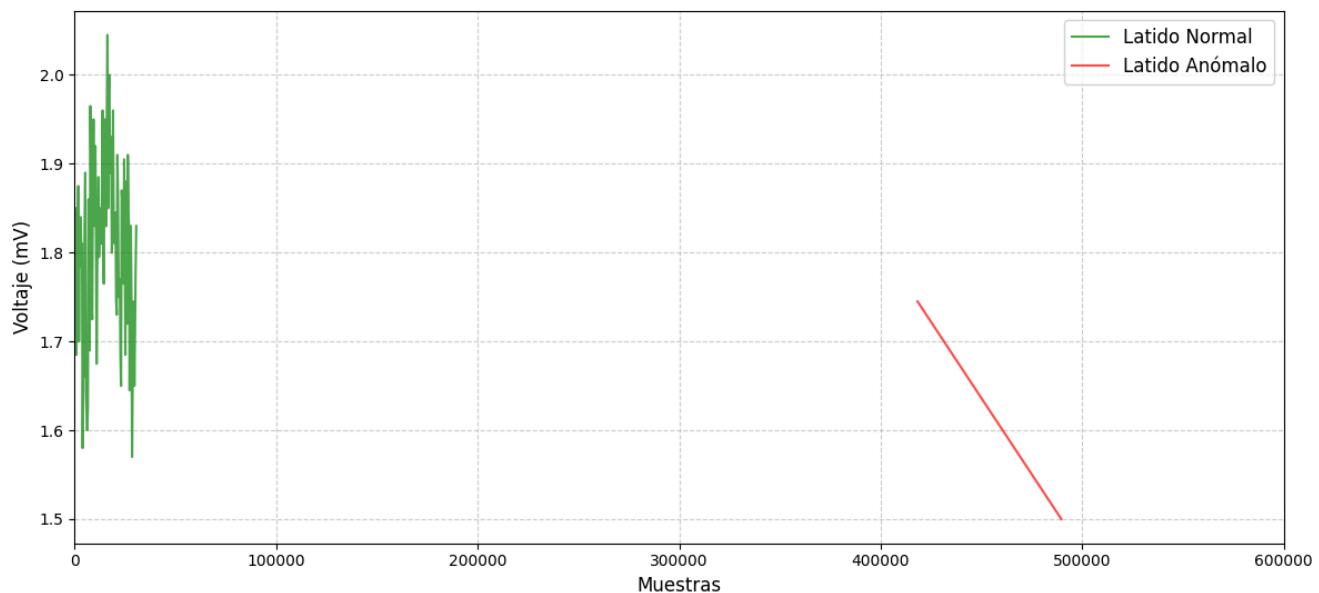
```
comparar_latidos_paciente(df_final, ann_label, paciente='103')
```

➡ Latidos normales encontrados: 100

Latidos anómalos encontrados: 2

Gráfico guardado en: output\patrones_morfo_latidos_paciente_103.png

Comparación de Patrones Morfológicos: Latido Normal vs. Anómalos (Paciente 103)



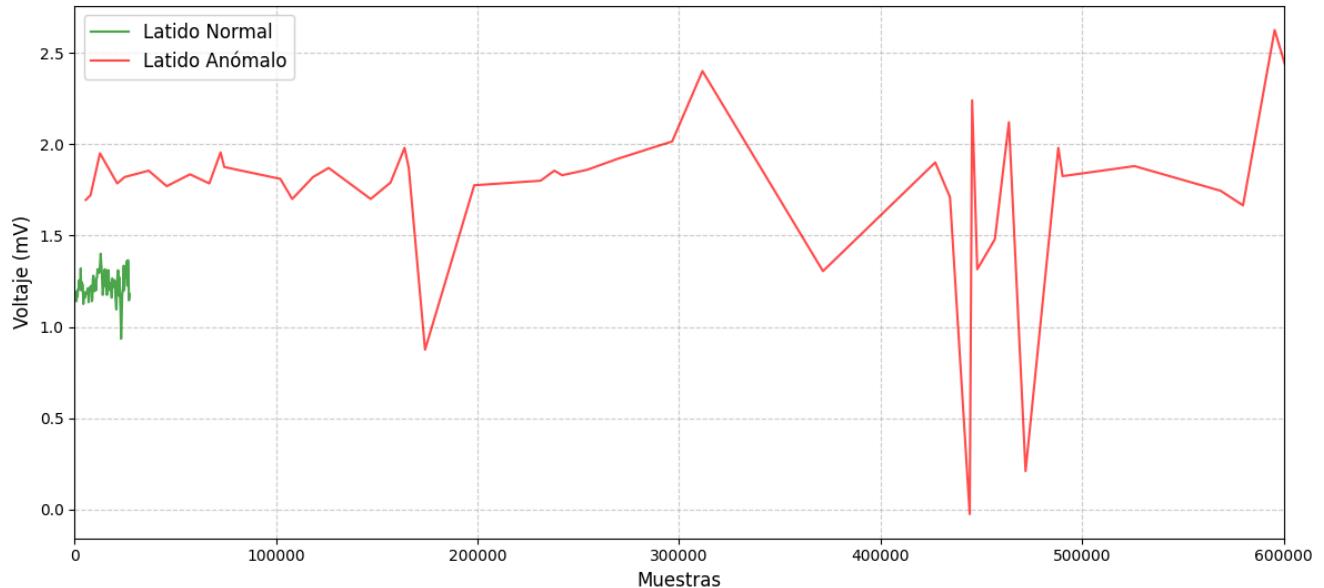
```
comparar_latidos_paciente(df_final, ann_label, paciente='105')
```

→ Latidos normales encontrados: 100

Latidos anómalos encontrados: 46

Gráfico guardado en: output\patrones_morfo_latidos_paciente_105.png

Comparación de Patrones Morfológicos: Latido Normal vs. Anómalos (Paciente 105)



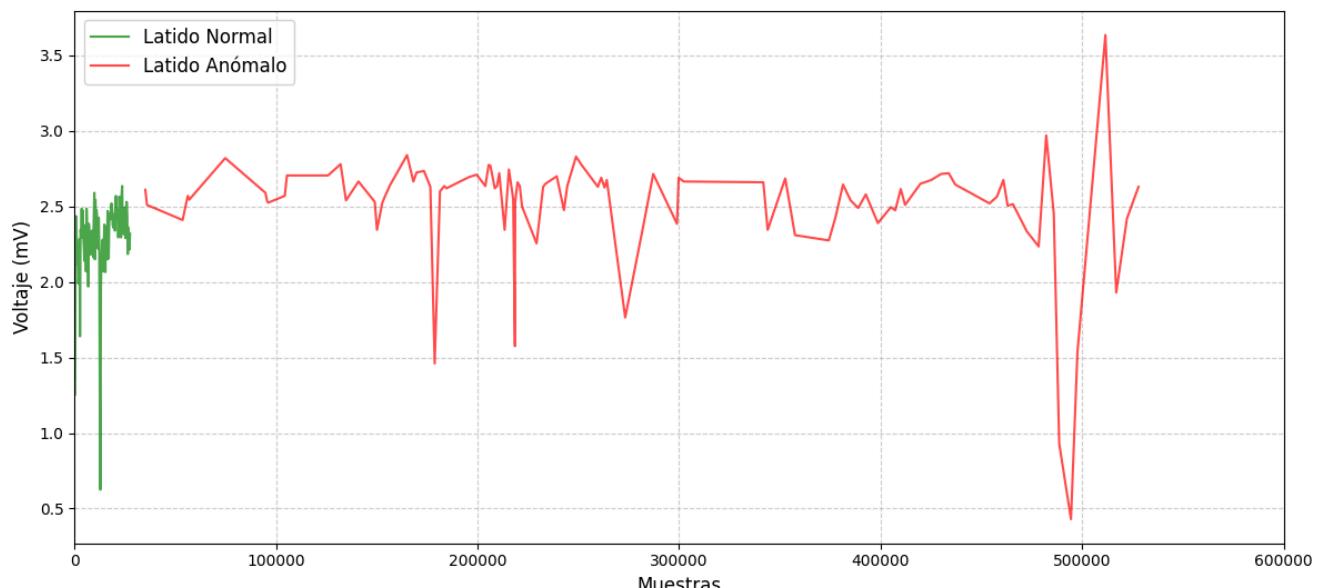
```
comparar_latidos_paciente(df_final, ann_label, paciente='116')
```

→ Latidos normales encontrados: 100

Latidos anómalos encontrados: 100

Gráfico guardado en: output\patrones_morfo_latidos_paciente_116.png

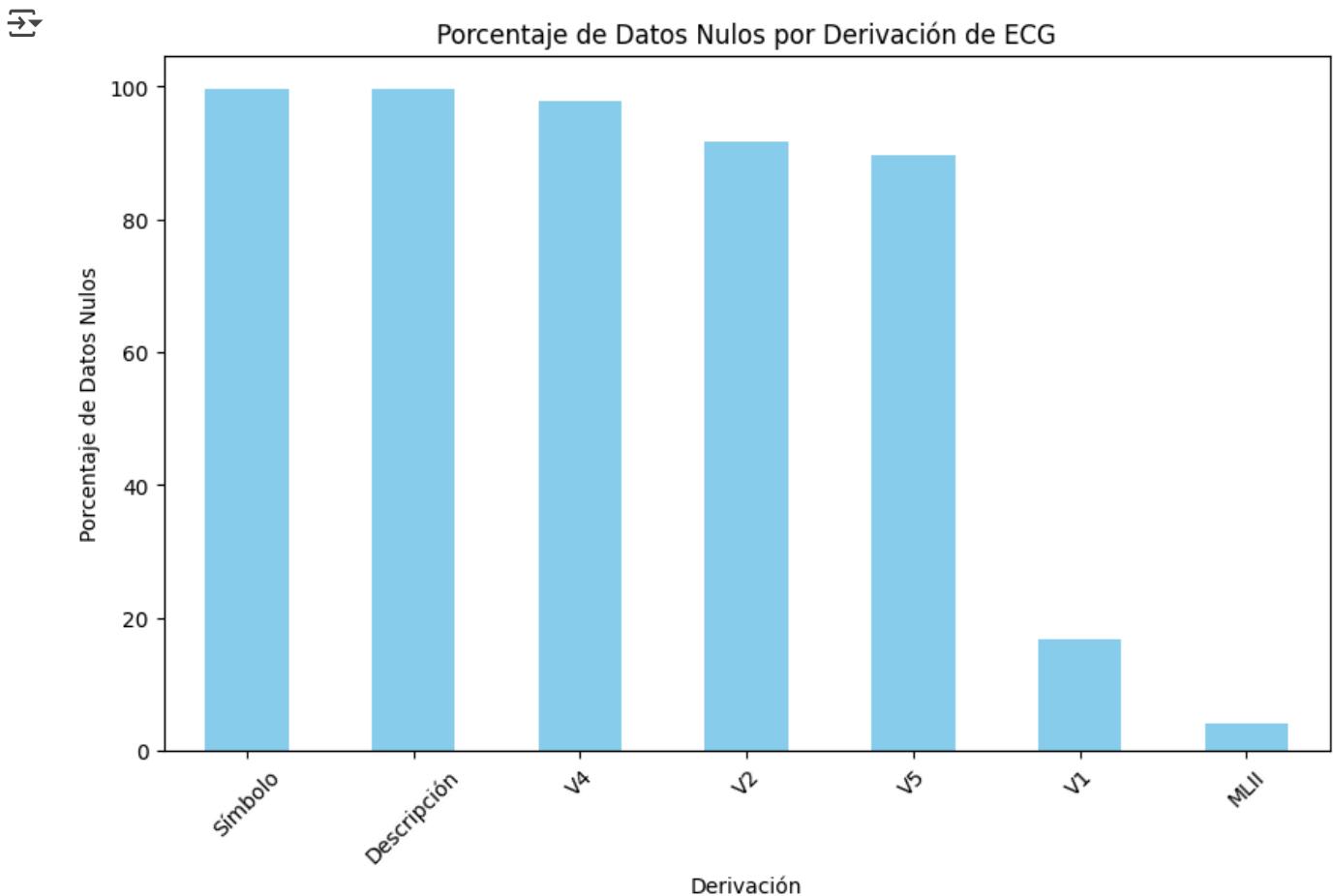
Comparación de Patrones Morfológicos: Latido Normal vs. Anómalos (Paciente 116)



5. Análisis de la cantidad de datos nulos en las distintas derivaciones

```
plt.figure(figsize=(10, 6))
missing_data = df_final.isnull().sum() / len(df_final) * 100 # Porcentaje de valores nulos
missing_data = missing_data[missing_data > 0] # Solo mostrar derivaciones con valores nulos
missing_data.sort_values(ascending=False).plot(kind='bar', color='skyblue')
plt.title('Porcentaje de Datos Nulos por Derivación de ECG')
plt.xlabel('Derivación')
```

```
plt.ylabel('Porcentaje de Datos Nulos')
plt.xticks(rotation=45)
plt.savefig(os.path.join(output_dir, 'porcentaje_datos_nulos.png'))
plt.show()
plt.close()
```

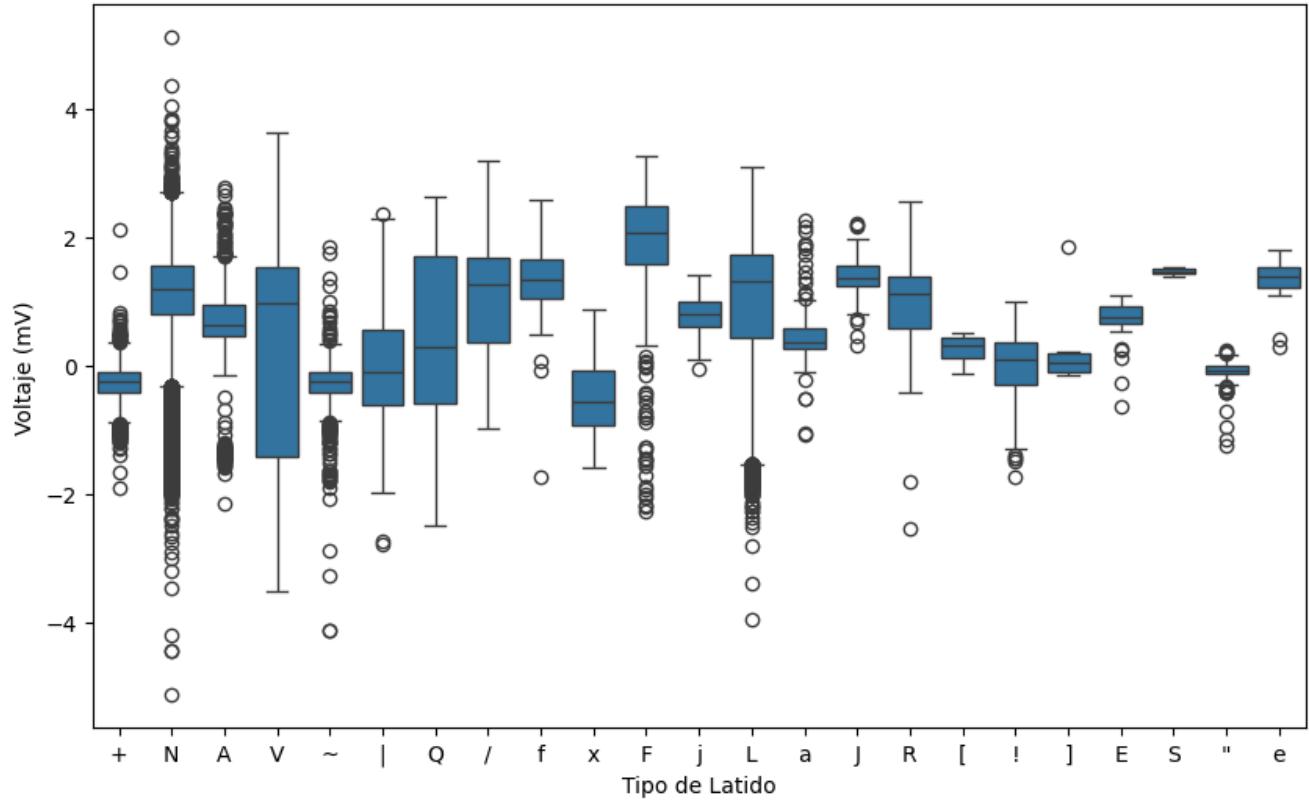


6. Análisis de la dispersión de los valores de voltaje de la señal ECG (detectar outliers)

```
plt.figure(figsize=(10, 6))
sns.boxplot(x='Símbolo', y='MLII', data=df_final)
plt.title('Dispersión de los Valores de Voltaje de la Señal ECG (MLII)')
plt.xlabel('Tipo de Latido')
plt.ylabel('Voltaje (mV)')
plt.savefig(os.path.join(output_dir, 'dispersion_valores_ecg.png'))
plt.show()
plt.close()
```



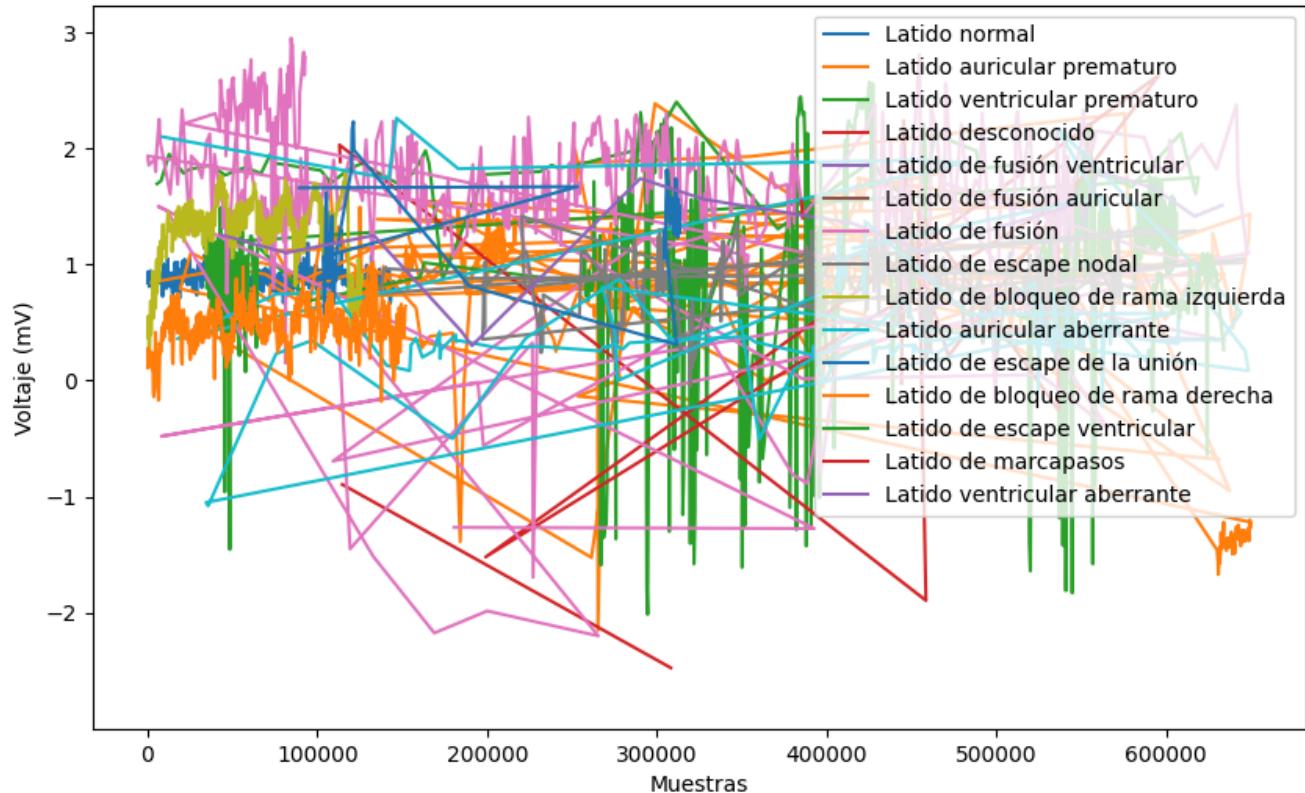
Dispersión de los Valores de Voltaje de la Señal ECG (MLII)



```
plt.figure(figsize=(10, 6))
for label in df_final['Símbolo'].unique():
    desc = ann_label.get(label, "Desconocido")
    if desc == "Desconocido":
        continue
    subset = df_final[df_final['Símbolo'] == label]
    plt.plot(subset['Sample'][:500], subset['MLII'][:500], label=desc)
plt.title('Distribución de los Latidos a lo Largo del Tiempo (Primeros 500 Latidos)')
plt.xlabel('Muestras')
plt.ylabel('Voltaje (mV)')
plt.legend(loc='upper right')
plt.savefig(os.path.join(output_dir, 'distribucion_latidos_tiempo.png'))
plt.show()
plt.close()
```



Distribución de los Latidos a lo Largo del Tiempo (Primeros 500 Latidos)



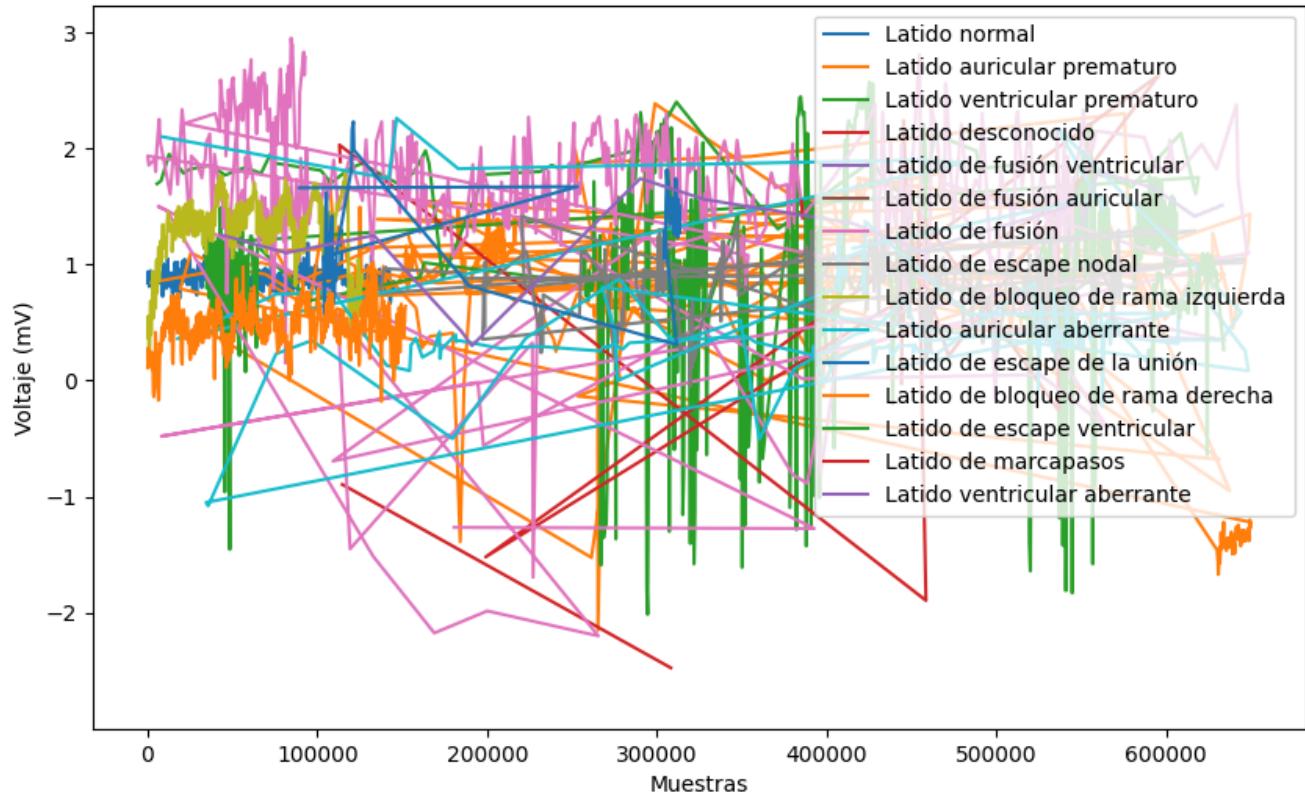
```

plt.figure(figsize=(10, 6))
for label in df_final['Símbolo'].unique():
    desc = ann_label.get(label, "Desconocido")
    if desc == "Desconocido":
        continue
    subset = df_final[df_final['Símbolo'] == label]
    plt.plot(subset['Sample'][:500], subset['MLII'][:500], label=desc)
plt.title('Distribución de los Latidos a lo Largo del Tiempo (Primeros 500 Latidos)')
plt.xlabel('Muestras')
plt.ylabel('Voltaje (mV)')
plt.legend(loc='upper right')
plt.savefig(os.path.join(output_dir, 'distribucion_latidos_tiempo.png'))
plt.show()
plt.close()

```



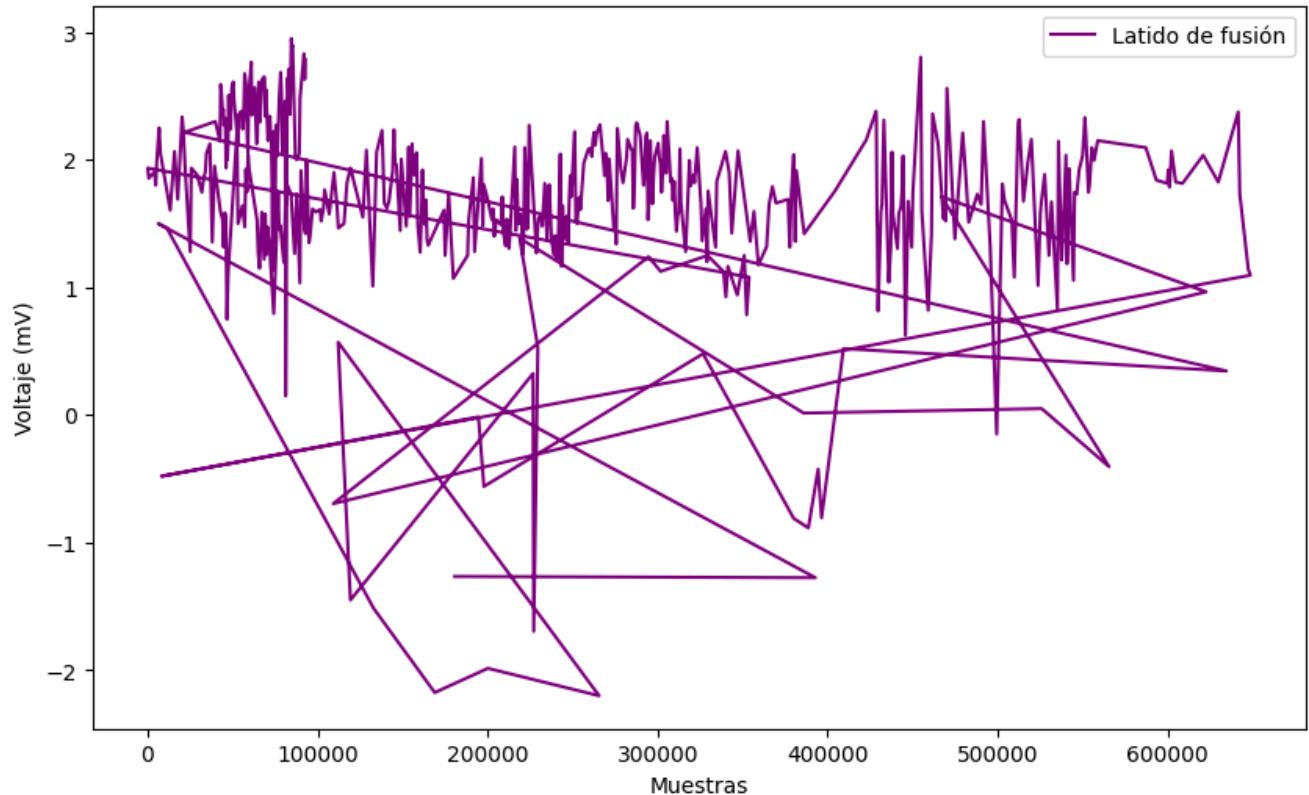
Distribución de los Latidos a lo Largo del Tiempo (Primeros 500 Latidos)



```
# Graficar solo los latidos de fusión ('F')
fusion_label = 'F'
fusion_desc = ann_label.get(fusion_label, "Desconocido")
fusion_subset = df_final[df_final['Símbolo'] == fusion_label]
if not fusion_subset.empty:
    plt.figure(figsize=(10, 6))
    plt.plot(fusion_subset['Sample'][:500], fusion_subset['MLII'][:500], label=fusion_desc, color='purple')
    plt.title('Latido de Fusión (Primeros 500 Latidos)')
    plt.xlabel('Muestras')
    plt.ylabel('Voltaje (mV)')
    plt.legend()
    plt.show()
else:
    print("No se encontraron latidos de fusión ('F') en el dataset.")
```



Latido de Fusión (Primeros 500 Latidos)

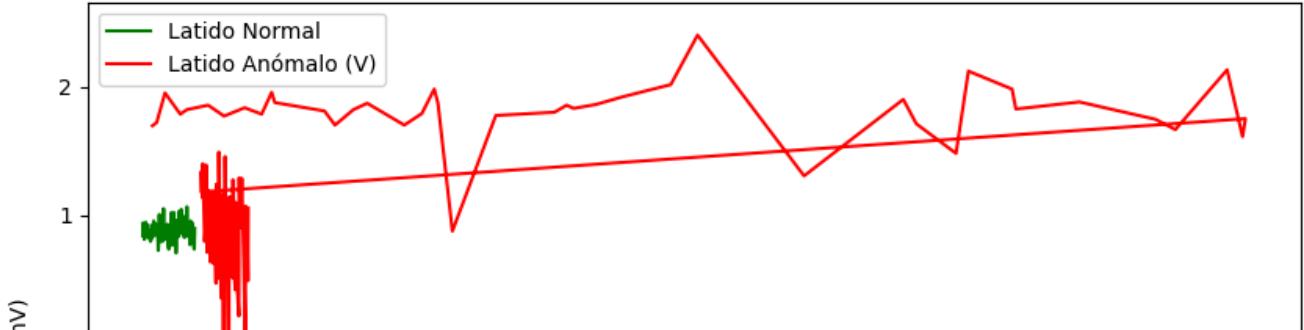


```
# 8. Comparación de características morfológicas de los latidos normales vs. latidos anómalos
# Seleccionamos una muestra de latidos normales (N) y latidos anómalos (V) para su comparación
normal_beat = df_final[df_final['Símbolo'] == 'N'].iloc[0:100] # Primeros 100 latidos normales
abnormal_beat = df_final[df_final['Símbolo'] == 'V'].iloc[0:100] # Primeros 100 latidos ventriculares pre
```

```
plt.figure(figsize=(10, 6))
plt.plot(normal_beat['Sample'], normal_beat['MLII'], label='Latido Normal', color='green')
plt.plot(abnormal_beat['Sample'], abnormal_beat['MLII'], label='Latido Anómalo (V)', color='red')
plt.title('Comparación de Patrones Morfológicos: Latido Normal vs. Anómalo')
plt.xlabel('Muestras')
plt.ylabel('Voltaje (mV)')
plt.legend()
plt.savefig(os.path.join(output_dir, 'comparacion_morfo_latidos.png'))
plt.show()
plt.close()
```



Comparación de Patrones Morfológicos: Latido Normal vs. Anómalo

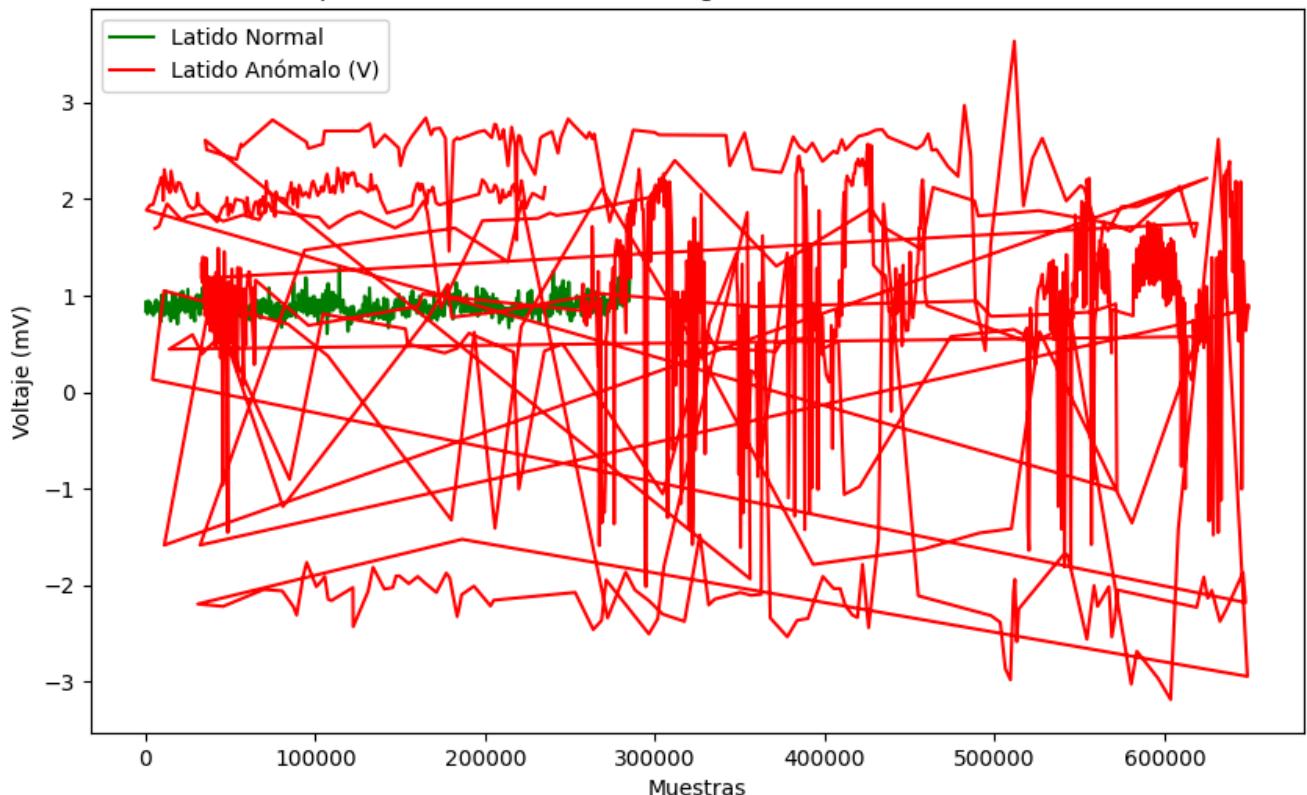


```
# 8. Comparación de características morfológicas de los latidos normales vs. latidos anómalos
# Seleccionamos una muestra de latidos normales (N) y latidos anómalos (V) para su comparación
normal_beat = df_final[df_final['Símbolo'] == 'N'].iloc[0:1000] # Primeros 100 latidos normales
abnormal_beat = df_final[df_final['Símbolo'] == 'V'].iloc[0:1000] # Primeros 100 latidos ventriculares pr
```

```
plt.figure(figsize=(10, 6))
plt.plot(normal_beat['Sample'], normal_beat['MLII'], label='Latido Normal', color='green')
plt.plot(abnormal_beat['Sample'], abnormal_beat['MLII'], label='Latido Anómalo (V)', color='red')
plt.title('Comparación de Patrones Morfológicos: Latido Normal vs. Anómalo')
plt.xlabel('Muestras')
plt.ylabel('Voltaje (mV)')
plt.legend()
plt.savefig(os.path.join(output_dir, 'comparacion_morfo_latidos.png'))
plt.show()
plt.close()
```



Comparación de Patrones Morfológicos: Latido Normal vs. Anómalo



“AÑO DE LA RECUPERACIÓN Y CONSOLIDACIÓN
DE LA ECONOMÍA PERUANA”.



ESCUELA PROFESIONAL DE CIENCIA DE LA
COMPUTACIÓN
TÓPICOS EN CIENCIA DE DATOS

Data Wrangling

Estudiantes:

Piero Emiliano Vizcarra Vargas

Docente :

Ana Maria Cuadros Valdivia



Índice

| | |
|---|-----------|
| 1. Contexto del Dataset | 3 |
| 1.1. Antecedentes | 3 |
| 1.2. Sobre la Obtención de los Datos | 3 |
| 2. Conceptos Médicos Previos | 4 |
| 2.1. Arritmia | 4 |
| 2.2. Electrocardiograma (ECG) | 4 |
| 2.3. Latido Normal | 4 |
| 2.4. Latido Prematuro | 4 |
| 2.5. Latido de Escape | 4 |
| 2.6. Fusión de Latidos | 5 |
| 2.7. Latido Aberrante | 5 |
| 2.8. Anotaciones de Latidos | 5 |
| 2.9. Sistema de Conducción Cardíaco | 5 |
| 2.10. Intervalos y Ondas en un ECG | 5 |
| 2.11. Tachicardia y Bradicardia | 5 |
| 3. Análisis del Comportamiento de los Datos | 6 |
| 3.1. Descripción del Registro | 6 |
| 3.2. Número de Registros | 6 |
| 3.3. Evaluación de la Cantidad de Registros | 6 |
| 3.4. Capacidad de Procesamiento (CPU + RAM) | 7 |
| 3.5. Identificación de Datos Duplicados | 7 |
| 3.6. Tipos de Datos en el Dataset | 7 |
| 3.7. Rango de los Datos por Columna | 7 |
| 3.8. Formato y Unidades de Medida | 7 |
| 3.9. Datos Categóricos | 7 |
| 3.10. Granularidad de los Datos | 8 |
| 3.11. Filas con Valores Nulos | 8 |
| 4. Medidas Estadísticas | 9 |
| 4.1. Análisis | 9 |
| 4.1.1. ¿Se evidencia alguna distribución? | 10 |
| 4.2. Anomalía en V4 | 11 |
| 4.3. Correlación y covarianza: permite entender la relación entre dos variables aleatorias. | 12 |
| 4.3.1. ¿Hay correlación entre features (características)? | 12 |
| 5. Principales Desafíos | 14 |
| 5.1. Problema de tipo supervisado | 14 |
| 5.2. Dependencia temporal de los datos | 14 |
| 5.3. Desbalance de clases en el dataset | 14 |
| 5.4. Consideraciones de calidad de los datos | 15 |
| 5.5. Relación entre las variables | 15 |

| | |
|--|-----------|
| 6. Tablas | 15 |
| 6.1. Tabla 1: Resumen de las Columnas | 15 |
| 6.2. Tabla 2: Descripción de las Anotaciones | 17 |
| 6.3. Ejemplo Extraido de Latido Normal (N) | 18 |
| 6.4. Ejemplo Extraido de Latido de Bloqueo de Rama Izquierda (L) | 18 |
| 6.5. Ejemplo Extraido de Latido de Bloqueo de Rama Derecha (R) | 19 |
| 6.6. Ejemplo Extraido de Latido Ventricular Prematuro (V) | 19 |
| 6.7. Ejemplo Extraido de Latido Auricular Prematuro (A) | 20 |
| 6.8. Ejemplo Extraido de Latido de Fusión (F) | 20 |
| 6.9. Ejemplo Extraido de Latido de Fusión Ventricular (/) | 21 |
| 6.10. Ejemplo Extraido de Latido de Escape Nodal (j) | 21 |
| 6.11. Ejemplo Extraido de Latido de Escape Ventricular (E) | 22 |
| 6.12. Ejemplo Extraido de Latido Auricular Aberrante (a) | 22 |
| 6.13. Ejemplo Extraido de Latido de Escape de la Unión (J) | 23 |
| 6.14. Ejemplo Extraido de Latido de Marcapasos (S) | 23 |
| 6.15. Ejemplo Extraido de Latido Ventricular Aberrante (e) | 24 |
| 6.16. Ejemplo Extraido de Latido Desconocido (Q) | 24 |
| 6.17. Ejemplo Extraido de Latido Desconocido (+) | 25 |

1. Contexto del Dataset

1.1. Antecedentes

Desde 1975, los laboratorios en el Hospital Beth Israel de Boston (ahora el Beth Israel Deaconess Medical Center) y en el MIT han respaldado nuestra propia investigación en análisis de arritmias y temas relacionados. Uno de los primeros productos importantes de ese esfuerzo fue la **MIT-BIH Arrhythmia Database**, que completamos y comenzamos a distribuir en 1980. La base de datos fue el primer conjunto de material de prueba estándar disponible para la evaluación de detectores de arritmias y ha sido utilizada para ese propósito, así como para la investigación básica sobre la dinámica cardíaca en más de 500 sitios en todo el mundo. Originalmente, distribuimos la base de datos en cintas digitales de 9 pistas de media pulgada a 800 y 1600 bpi, y en cintas analógicas FM de formato IRIG de cuarto de pulgada.

1.2. Sobre la Obtención de los Datos

La MIT-BIH Arrhythmia Database contiene 48 extractos de media hora de grabaciones ECG de dos canales, obtenidos de 47 sujetos estudiados por el Laboratorio de Arritmias BIH entre 1975 y 1979.

23 grabaciones fueron seleccionadas al azar de un conjunto de 4000 grabaciones ECG ambulantes de 24 horas recopiladas de una población mixta de pacientes hospitalizados (aproximadamente 60 %) y ambulatorios (aproximadamente 40 %) en el Hospital Beth Israel de Boston; las 25 grabaciones restantes fueron seleccionadas de ese mismo conjunto para incluir arritmias menos comunes pero clínicamente significativas que no estarían bien representadas en una pequeña muestra aleatoria.

Las grabaciones fueron digitalizadas a 360 muestras por segundo por canal con una resolución de 11 bits sobre un rango de 10 mV.

Dos o más cardiólogos anotaron independientemente cada registro; las discrepancias se resolvieron para obtener las anotaciones de referencia legibles por computadora para cada latido (aproximadamente 110,000 anotaciones en total) incluidas en la base de datos.

2. Conceptos Médicos Previos

Antes de comenzar con el análisis del dataset, es necesario familiarizarse con algunos términos médicos clave que se encuentran en este contexto de arritmias cardíacas y electrocardiogramas (ECG).

2.1. Arritmia

La arritmia es cualquier alteración del ritmo cardíaco normal. En un corazón sano, el ritmo cardíaco es regulado por impulsos eléctricos que viajan a través del sistema de conducción del corazón. Cuando este sistema eléctrico no funciona correctamente, los latidos del corazón pueden volverse irregulares, demasiado rápidos (taquicardia), demasiado lentos (bradicardia), o incluso descoordinados. Las arritmias pueden ser benignas o potencialmente peligrosas, dependiendo de su tipo y la gravedad.

2.2. Electrocardiograma (ECG)

Un electrocardiograma (ECG) es una prueba médica que mide la actividad eléctrica del corazón. Durante un ECG, se colocan electrodos en la piel del paciente, los cuales registran las señales eléctricas generadas por el corazón. Estas señales se visualizan como ondas en un gráfico, representando diferentes fases del ciclo cardíaco, como la contracción y relajación de las aurículas y los ventrículos. El ECG es una herramienta esencial para diagnosticar y monitorear trastornos del ritmo cardíaco.

2.3. Latido Normal

Un latido normal es el ritmo cardíaco regular generado por el nodo sinoauricular (SA) del corazón. Este latido sigue un patrón constante y está controlado por el sistema de conducción cardíaco, asegurando que las aurículas y los ventrículos se contraigan de manera coordinada. En el contexto del dataset, los latidos normales están representados por el símbolo N.

2.4. Latido Prematuro

Un latido prematuro es un latido que ocurre antes de lo esperado, interrumpiendo el ritmo cardíaco regular. Los latidos prematuros pueden originarse en las aurículas (latido auricular prematuro) o en los ventrículos (latido ventricular prematuro). Estos latidos pueden ser un signo de arritmias o de un corazón sano que responde a factores como el estrés o la fatiga.

2.5. Latido de Escape

Un latido de escape ocurre cuando el nodo sinoauricular (SA) o el nodo auriculoventricular (AV) no están funcionando correctamente y otras partes del sistema de conducción del corazón asumen el control. Los latidos de escape pueden ser auriculares o ventriculares, y son una respuesta del corazón a la falta de un ritmo normal.

2.6. Fusión de Latidos

La fusión de latidos se produce cuando dos impulsos eléctricos se combinan para generar un único latido. Existen dos tipos comunes de latidos de fusión: fusión ventricular y fusión auricular. Estos latidos pueden resultar de la interacción entre un latido prematuro y el latido normal del corazón.

2.7. Latido Aberrante

Un latido aberrante es un latido que se origina en una parte del sistema de conducción del corazón que no es la zona normal. Estos latidos pueden ser ventriculares o auriculares y se caracterizan por una conducción anómala, lo que puede hacer que el latido sea menos eficiente.

2.8. Anotaciones de Latidos

Las anotaciones de latidos en un ECG son etiquetas asignadas por cardiólogos a las diferentes partes del gráfico que corresponden a ciertos tipos de latidos o eventos cardíacos. En el dataset de MIT-BIH, estas anotaciones han sido realizadas de manera manual por cardiólogos y son utilizadas como referencia para entrenar modelos de clasificación automática. Cada anotación está asociada a un símbolo específico (como N, V, A, L, etc.), que indica el tipo de latido registrado en ese punto del tiempo.

2.9. Sistema de Conducción Cardíaco

El sistema de conducción cardíaco es un sistema especializado de fibras musculares en el corazón que transmite los impulsos eléctricos necesarios para que el corazón late de manera eficiente. El sistema incluye el nodo sinoauricular (SA), el nodo auriculoventricular (AV), el Haz de His, y las fibras de Purkinje. El mal funcionamiento de cualquier parte de este sistema puede dar lugar a arritmias, que son trastornos del ritmo cardíaco.

2.10. Intervalos y Ondas en un ECG

En un ECG, se pueden observar diferentes ondas y segmentos que corresponden a distintas fases del ciclo cardíaco:

- Onda P: Representa la despolarización de las aurículas.
- Complejo QRS: Representa la despolarización de los ventrículos.
- Onda T: Representa la repolarización de los ventrículos.

2.11. Tachicardia y Bradicardia

La taquicardia es un ritmo cardíaco anormalmente rápido, generalmente superior a 100 latidos por minuto, y puede ser un signo de que el corazón no está funcionando correctamente. Por otro lado, la bradicardia es un ritmo cardíaco anormalmente lento, generalmente inferior a 60 latidos por minuto, y también puede indicar problemas con la conducción eléctrica del corazón.

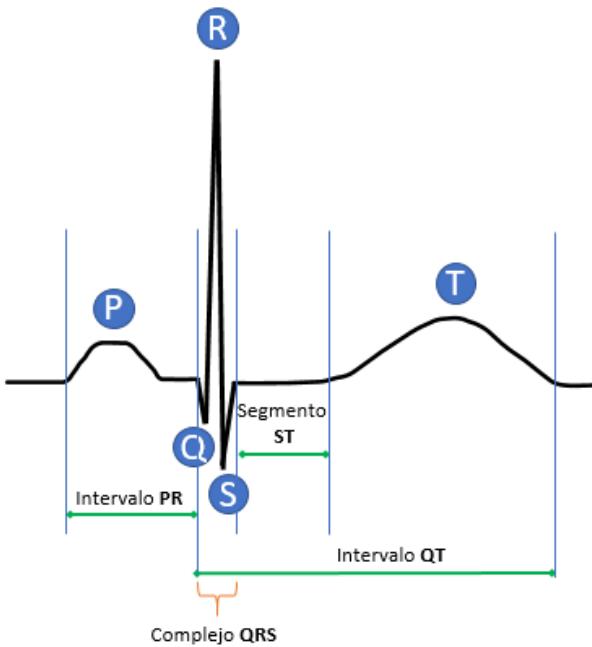


Figura 1: Ejemplo de Latido

3. Análisis del Comportamiento de los Datos

3.1. Descripción del Registro

Cada registro corresponde a una muestra de ECG de un paciente específico. El registro contiene señales de diferentes derivaciones (MLII, V1, V2, V4, V5) y, en algunos casos, también anotaciones que indican el tipo de latido en ese instante.

3.2. Número de Registros

El dataset contiene un total de **31,200,000 registros** provenientes de **48 pacientes**. La información proporcionada por cada paciente está distribuida en diferentes segmentos de tiempo, representados por las muestras de voltaje de cada derivación.

$$\text{Número total de registros} = 31,200,000$$

$$\text{Número de pacientes} = 48$$

3.3. Evaluación de la Cantidad de Registros

La cantidad de registros es considerable para realizar un análisis médico. Dado que el volumen de datos médicos se encuentra dentro de los límites normales, no se considera que la cantidad sea demasiado baja. La información proporcionada es suficiente para realizar un análisis de arritmias cardíacas en múltiples episodios de los pacientes.

3.4. Capacidad de Procesamiento (CPU + RAM)

A pesar de que el tamaño total es grande, la capacidad de procesamiento no presenta un obstáculo. El equipo que usaremos en esta ocasión es una Ryzen 5 7800 con 16 GB de RAM el cual pueden manejar estos datos sin dificultades significativas, ya que el dataset no supera los 2.1 GB cuando se procesa en un dataframe de Pandas, mas pesa 1.5 en el caso que se cargue con la librería Polars.

3.5. Identificación de Datos Duplicados

El dataset no contiene registros duplicados, ya que cada muestra es única y refleja datos continuos que corresponden a momentos específicos en la grabación de señales. No existen registros repetidos, lo que asegura la calidad y autenticidad de los datos. Cabe mencionar , que por la metodología de recolección descrita en el contexto lo cual hace que los datos no tengan duplicados.

3.6. Tipos de Datos en el Dataset

El dataset está compuesto por datos tanto **discretos** como **continuos** los cuales describiremos a continuación

Datos Discretos: Clases de latidos, número de latidos, identificadores de pacientes

Datos Continuos: Señales de ECG (MLII, V1, V2, V4, V5)

3.7. Rango de los Datos por Columna

Los valores de las señales de ECG varían entre **-5.120 mV** y **+5.115 mV** dependiendo de la derivación.

Rango MLII = -5,120 mV a + 5,115 mV

Rango V1 = -5,120 mV a + 5,115 mV

3.8. Formato y Unidades de Medida

Los datos están **en su formato adecuado**, ya que las señales están en **milivoltios (mV)** y las frecuencias de muestreo son constantes a **360 Hz**. Las columnas de anotaciones de latidos están en formato **categórico** (texto).

3.9. Datos Categóricos

Las columnas **categóricas** corresponden a las **anotaciones de los latidos**, como **N**, **V**, **A**, etc. Para aplicar **modelos de Machine Learning**, será necesario **convertir estos datos en numéricos** mediante técnicas como **one-hot encoding** o **label encoding**.

3.10. Granularidad de los Datos

- Cada **registro** representa las señales de ECG de un paciente específico durante un periodo de tiempo.
- Cada **fila** dentro de las señales (MLII, V1,V2,V4 y V5) representa una muestra de voltaje en milivoltios en un momento específico.
- Cada **anotación** (clase) marca un latido particular, facilitando la clasificación de los latidos dentro del registro.

3.11. Filas con Valores Nulos

Se han encontrado **valores nulos** en algunas columnas, principalmente debido a que ciertos registros no contienen **todos los electrodos** (como MLII, V1, etc.). Estos valores nulos no invalidan el análisis, pero se deben tratar adecuadamente.

Número de Valores Nulos MLII = 1,300,000

Número de Valores Nulos V5 = 27,950,000

4. Medidas Estadísticas

Cuadro 1: Estadísticas descriptivas (vertical)

| Variable | Count | Mean | Std | Min | 25 % | 50 % | 75 % | Max |
|----------|------------|--------|-------|--------|--------|--------|--------|-------|
| MLII | 29 900 000 | -0,338 | 0,485 | -5,120 | -0,605 | -0,300 | -0,135 | 5,115 |
| V5 | 3 250 000 | -0,264 | 0,228 | -2,465 | -0,385 | -0,255 | -0,155 | 1,975 |
| V1 | 26 000 000 | -0,099 | 0,415 | -5,120 | -0,325 | -0,005 | 0,125 | 5,115 |
| V2 | 2 600 000 | -0,091 | 0,730 | -5,120 | -0,405 | -0,015 | 0,130 | 5,115 |
| V4 | 650 000 | -0,552 | 0,261 | -3,260 | -0,675 | -0,560 | -0,440 | 2,460 |

Cuadro 2: Otras medidas estadísticas

| Variable | Mediana | Moda | Rango |
|----------|-------------|--------|-------------|
| MLII | -0,300 | -0,245 | 10,235 |
| V5 | -0,255 | -0,245 | 4,440 |
| Sample | 324 999,500 | 0,000 | 649 999,000 |
| V1 | -0,005 | 0,075 | 10,235 |
| V2 | -0,015 | 0,035 | 10,235 |
| V4 | -0,560 | -0,605 | 5,720 |

4.1. Análisis

■ MLII (Señal principal):

- Distribución asimétrica negativa (media=-0.338 < mediana=-0.300)
- Alta dispersión (std=0.485) con valores extremos (-5.12 a 5.115)
- Moda en -0.245 sugiere concentración en valores ligeramente negativos

■ V5:

- Distribución más simétrica (media=-0.264 mediana=-0.255)
- Menor variabilidad (std=0.228) que MLII
- Rango reducido (4.44) sin valores extremos pronunciados

■ Asimetrías contrastantes:

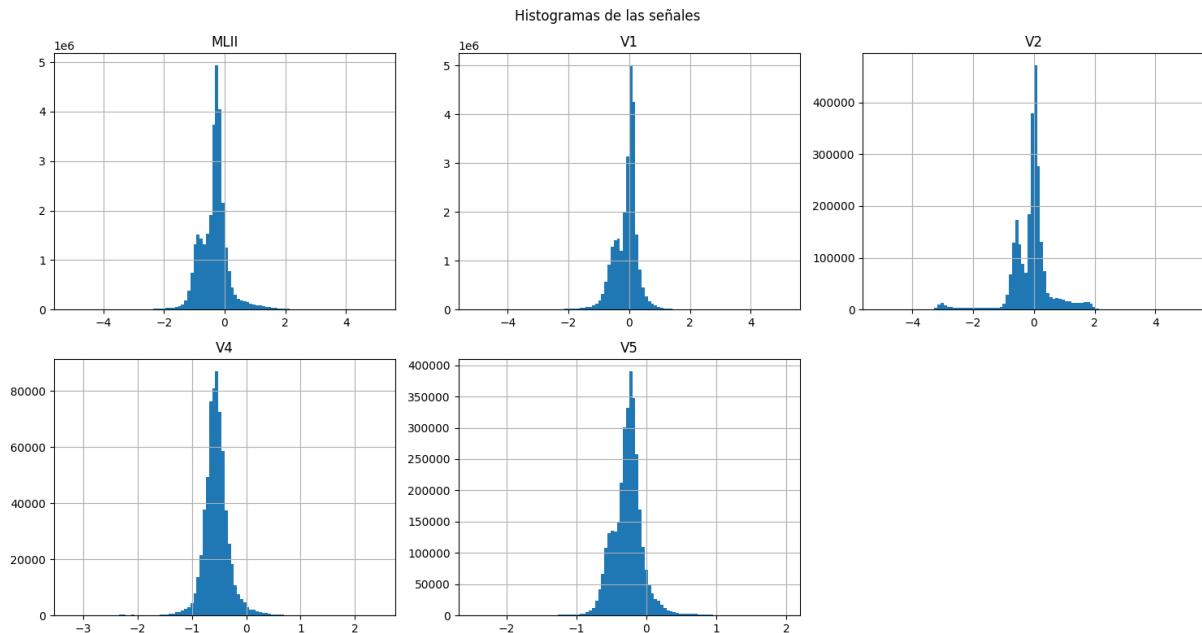


Figura 2: Frecuencia de Electrodos

- V1 muestra asimetría positiva (media > mediana)
- V4 presenta distribución simétrica perfecta (media = mediana = moda)
- **Dispersión anómala:**
 - V2 exhibe la mayor variabilidad ($\text{std}=0.730$)
 - Sample tiene rango máximo (649,999) pero distribución uniforme
- **Concentraciones atípicas:**
 - Moda de Sample en 0.000 sugiere posible valor por defecto
 - V4 concentra valores en -0.560 ± 0.1 (95 % entre -0.675 y -0.440)

4.1.1. ¿Se evidencia alguna distribución?

- **MLII:**
 - Distribución bimodal con picos en -2 y 2 mV
 - Presencia de outliers extremos (valores cerca de ± 4 mV)
 - Asimetría leve hacia valores positivos
- **V1:**
 - Distribución unimodal con pico en 0 mV
 - Cola más larga hacia valores positivos (asimetría positiva)
 - Menor dispersión que MLII (rango ± 3 mV)
- **V5:**

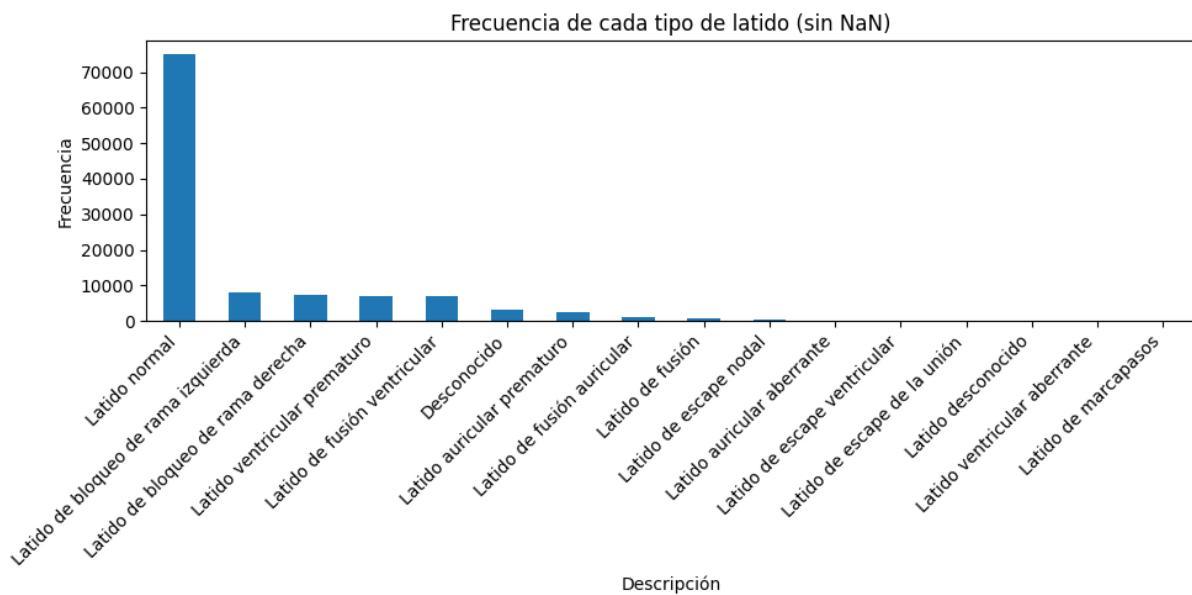


Figura 3: Frecuencia de la clasificación

- Distribución simétrica centrada en 0 mV
- Rango estrecho (± 2 mV) sin valores extremos
- Cumple con lo esperado para derivación precordial

4.2. Anomalía en V4

■ Patrón Inesperado:

- Distribución plana/multimodal (contrasta con la curva RQS esperada)
- Ausencia de pico central definido
- Amplitud excesiva (± 4 mV vs ± 2 mV en V5)

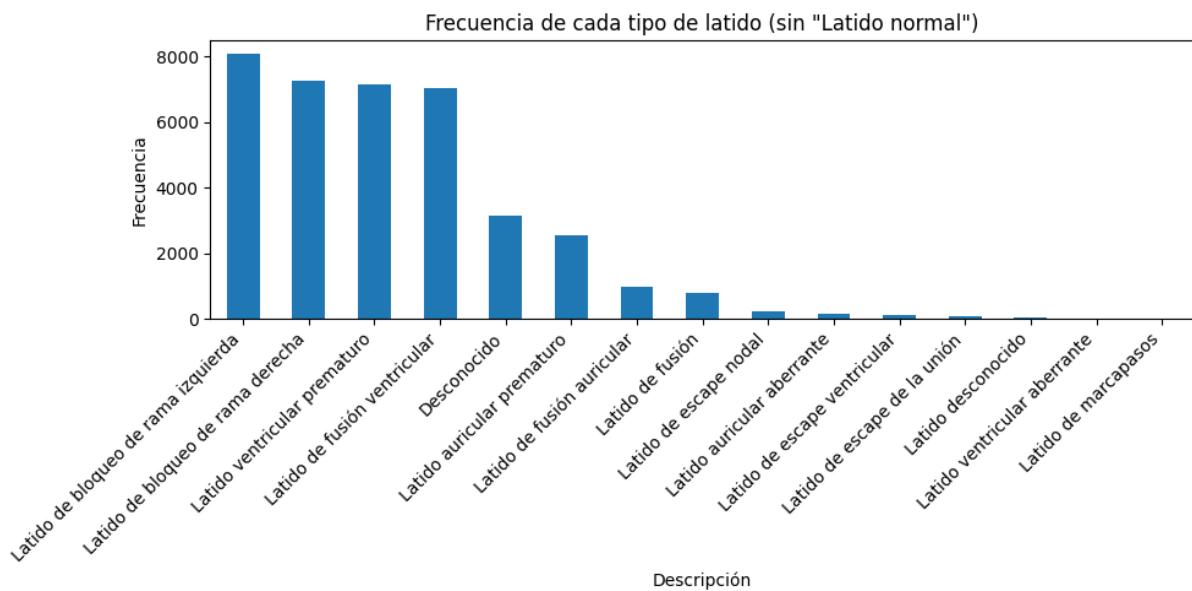


Figura 4: Frecuencia de la clasificación excluyendo valores normales

4.3. Correlación y covarianza: permite entender la relación entre dos variables aleatorias.

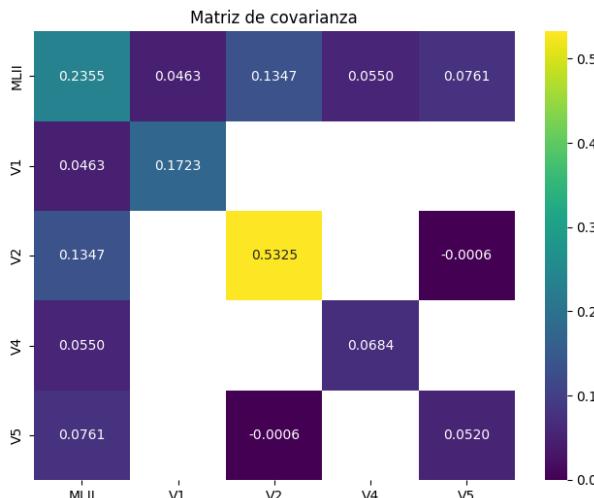


Figura 5: Matriz de Covarianza

4.3.1. ¿Hay correlación entre features (características)?

- **Correlaciones Fuertes ($|r| > 0.7$):**
 - MLII-V5 (0.72): Relación lineal significativa
 - Sugiere que estas derivaciones capturan información similar
 - La correlación fuerte se da dado que capturan la actividad eléctrica similar del ventrículo izquierdo.

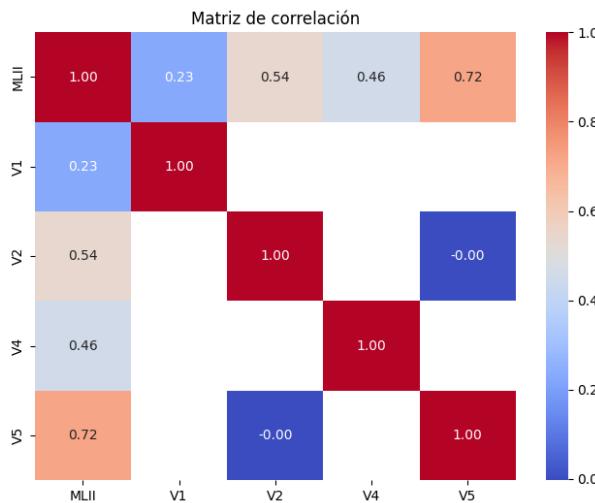


Figura 6: Matriz de Correlación

■ **Correlaciones Moderadas ($0.4 < |r| < 0.7$):**

- MLII-V2 (0.54) y MLII-V4 (0.46)
- V4 muestra correlación atenuada respecto a V5
- Tíeenn actividad registrada para la pared anterior del corazon
- La correlación baja de V4 a V5 , se puede deber a un posible movimiento.

■ **Correlaciones Débiles ($|r| < 0.3$):**

- MLII-V1 (0.23): Prácticamente independientes
- V2-V5 (-0.00): Ausencia total de correlación lineal
- MLII es sensible para el lado del vendiciculo izquierdo al contrario de V1.
- V1 es clave para diagnosticar arritmias auriculares
- V2 (pared anterior) y V5 (pared lateral) registran regiones anatómicas distintas del corazón.

■ **Mayor Variabilidad Individual:**

- V2 tiene la varianza más alta (0.5325 en diagonal)
- Confirma alta dispersión observada en histogramas

■ **Interacciones Notables:**

- Covarianza MLII-V5 (0.0761) $>$ MLII-V4 (0.0550)
- V2-V4 muestra covarianza casi nula (-0.0006)

Se puede comcluir que:

- Redundancia Detectada: MLII y V5 (corr=0.72)
- Independencia: V1 aporta información única (tiende a tener baja correlación con otras)
- Anomalía en V4:Correlación moderada con MLII pero baja covarianza.

5. Principales Desafíos

5.1. Problema de tipo supervisado

En este análisis, el objetivo es usar las señales de ECG de la derivación **MLII** como características de entrada para predecir el tipo de latido, lo que convierte este problema en un tipo de clasificación supervisada. La columna de salida del modelo corresponde a la columna **Descripción**, que indica el tipo de latido, ya sea normal o algún tipo de arritmia. El problema se clasifica como multiclase, ya que cada latido/anotación pertenece a una de varias clases diferentes, tales como **N** (latido normal), **V** (latido ventricular prematuro), **A** (latido auricular prematuro), entre otros. Sin embargo, uno de los principales desafíos de este dataset es el **desbalance de clases**, ya que la clase de latidos normales (**N**) es abrumadoramente más frecuente que las clases de latidos anormales, como **V**, **A**, y **R**, lo que crea un desajuste en la representación de las clases y puede dificultar el entrenamiento del modelo de clasificación.

La señal más importante para este dataset es **MLII**, ya que contiene información crítica sobre los latidos cardíacos. Por otro lado, derivaciones como **V4** podrían ser descartadas, ya que la información en la derivación **MLII** es suficiente para realizar la clasificación, y estas derivaciones adicionales no aportan valor significativo en términos de diagnóstico.

Para los registros que no contienen **MLIII**, se realizara por medio de regresión lineal su reconstrucción de la señal a partir de otras derivaciones

5.2. Dependencia temporal de los datos

Este dataset es un problema dependiente del tiempo, ya que las señales de ECG son series temporales. Cada muestra en la señal depende de las muestras anteriores y posteriores, lo que refleja cómo se transmite la actividad eléctrica del corazón a lo largo del tiempo. Las ondas del ECG, como P, QRS y T, están distribuidas a lo largo de la secuencia temporal, y su interpretación correcta requiere considerar esta secuencia temporal para realizar una clasificación precisa de los latidos. Esto hace que sea esencial incorporar métodos que tomen en cuenta la temporalidad de las muestras para predecir correctamente el tipo de latido.

5.3. Desbalance de clases en el dataset

El dataset de **MIT-BIH** presenta un desbalance significativo en las clases, con una sobrerepresentación de los latidos normales (**N**) y una subrepresentación de las clases anormales. Aunque el dataset contiene una cantidad considerable de latidos normales, las clases minoritarias como **V** (latido ventricular prematuro) y **A** (latido auricular prematuro) tienen menos muestras, lo que limita la capacidad del modelo para generalizar adecuadamente para todas las clases. Este desbalance puede causar que el modelo de clasificación favorezca las clases más comunes y no aprenda correctamente a identificar las clases menos frecuentes, como las arritmias raras.

5.4. Consideraciones de calidad de los datos

Aunque las señales en el dataset están bien registradas, existen algunos problemas que podrían afectar la calidad del análisis. En primer lugar, no todos los registros contienen las mismas derivaciones o electrodos, lo que significa que algunas columnas tendrán valores nulos en algunos registros, lo que dificulta el análisis. Además, algunas señales pueden estar contaminadas por **ruido** o **artefactos**, lo que puede afectar la precisión de las anotaciones de los latidos. La presencia de estos artefactos debe ser gestionada adecuadamente para evitar que distorsionen los resultados del modelo. Otro factor que debe considerarse es el fuerte desbalance de clases, lo cual es un problema común en los datasets médicos, pero que puede complicar el entrenamiento del modelo de clasificación.

5.5. Relación entre las variables

En este dataset no se han observado relaciones inesperadas entre las variables principales.

La señal de **MLII** es la principal fuente de información para identificar los tipos de latidos cardíacos. Las demás columnas, como las anotaciones de los latidos y las descripciones, cumplen una función de identificación y etiquetado más que de características de entrada en el modelo de clasificación. Esto significa que las relaciones entre las variables en este caso son bastante directas, sin la presencia de interacciones complejas que requieran ser modeladas específicamente.

6. Tablas

6.1. Tabla 1: Resumen de las Columnas

| Columna | Tipo de Dato | Descripción | Rango de Valores | Significado |
|-------------|--------------|--|--|--|
| MLII | Flotante | Señal de ECG registrada por la derivación MLII | -5.120 mV a +5.115 mV | Voltaje de la señal eléctrica del corazón. |
| V1 | Flotante | Señal de ECG registrada por la derivación V1 | -5.120 mV a +5.115 mV | Voltaje de la señal del corazón en la derivación V1. |
| V2 | Flotante | Señal de ECG registrada por la derivación V2 | -5.120 mV a +5.115 mV | Voltaje de la señal del corazón en la derivación V2. |
| V4 | Flotante | Señal de ECG registrada por la derivación V4 | -3.260 mV a +2.460 mV | Voltaje de la señal del corazón en la derivación V4. |
| V5 | Flotante | Señal de ECG registrada por la derivación V5 | -2.465 mV a +1.975 mV | Voltaje de la señal del corazón en la derivación V5. |
| Sample | Entero | Muestra del índice de tiempo de la señal de ECG | 0 a 649,999 | Posición de la muestra en la secuencia temporal. |
| Símbolo | Categórico | Anotación que clasifica el tipo de latido cardíaco | N, V, A, L, R, etc. | Clasificación del tipo de latido detectado. |
| Descripción | Categórico | Descripción del latido anotado | Latido normal, latido de bloqueo, etc. | Detalles sobre el tipo de latido detectado. |
| Registro | Entero | Identificador único del paciente o registro de ECG | Número único de registro | Identificador del paciente o del registro. |

6.2. Tabla 2: Descripción de las Anotaciones

| Símbolo | Descripción |
|---------|--------------------------------------|
| N | Latido normal. |
| L | Latido de bloqueo de rama izquierda. |
| R | Latido de bloqueo de rama derecha. |
| V | Latido ventricular prematuro. |
| A | Latido auricular prematuro. |
| F | Latido de fusión. |
| / | Latido de fusión ventricular. |
| f | Latido de fusión auricular. |
| j | Latido de escape nodal. |
| E | Latido de escape ventricular. |
| a | Latido auricular aberrante. |
| J | Latido de escape de la unión. |
| S | Latido de marcapasos. |
| e | Latido ventricular aberrante. |
| Q | Latido desconocido. |
| + | Latido desconocido. |

6.3. Ejemplo Extraido de Latido Normal (N)

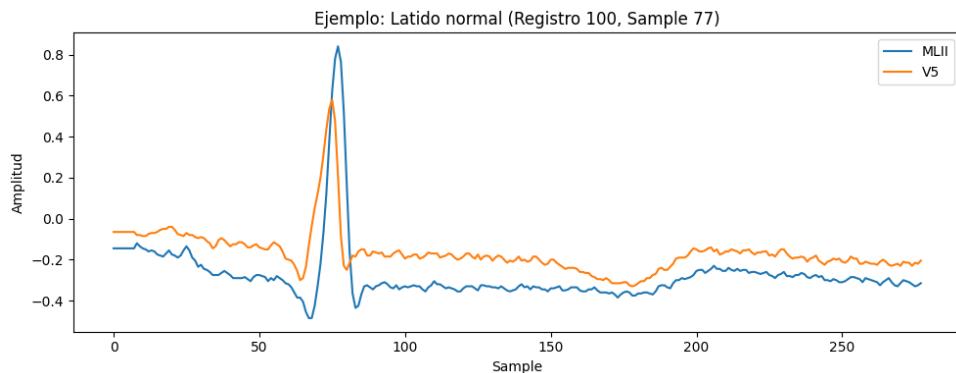


Figura 7: Latido normal. Este es el latido cardíaco típico y saludable.

6.4. Ejemplo Extraido de Latido de Bloqueo de Rama Izquierda (L)

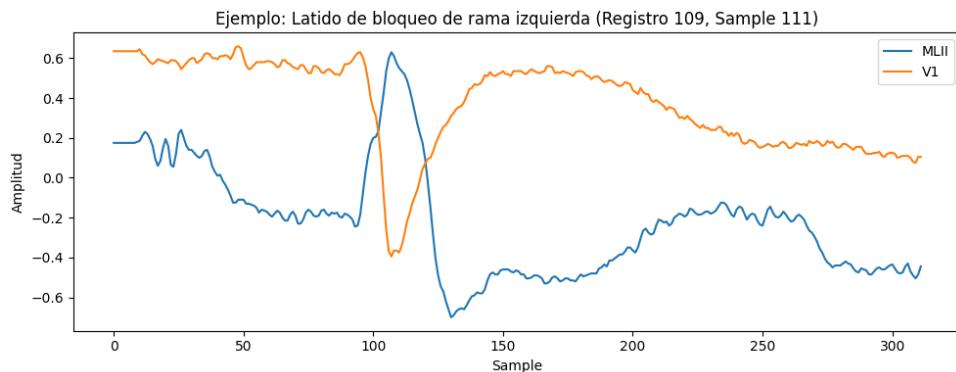


Figura 8: Latido de bloqueo de rama izquierda. Ocurre cuando hay un bloqueo en la rama izquierda del sistema de conducción del corazón.

6.5. Ejemplo Extraido de Latido de Bloqueo de Rama Derecha (R)

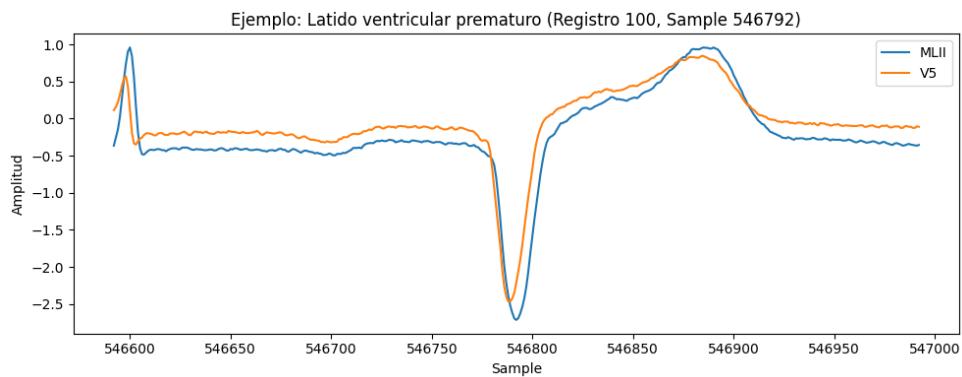


Figura 9: Latido de bloqueo de rama derecha. Un bloqueo en la rama derecha del sistema de conducción provoca este latido.

6.6. Ejemplo Extraido de Latido Ventricular Prematuro (V)

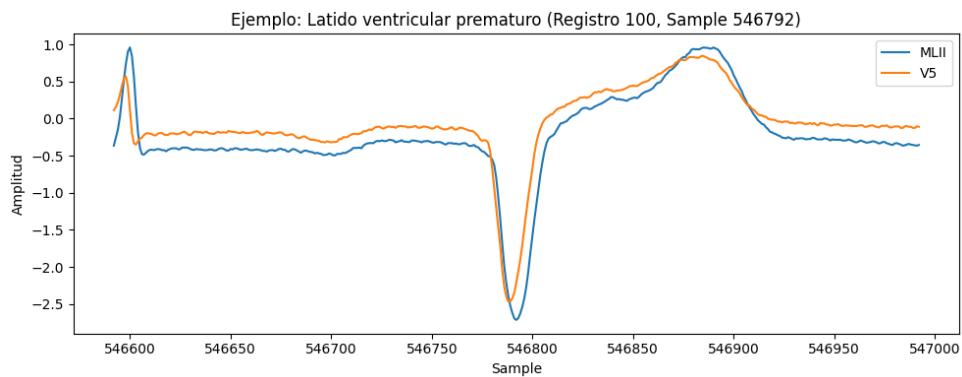


Figura 10: Latido ventricular prematuro. Un latido que se origina en los ventrículos antes de lo esperado.

6.7. Ejemplo Extraido de Latido Auricular Prematuro (A)

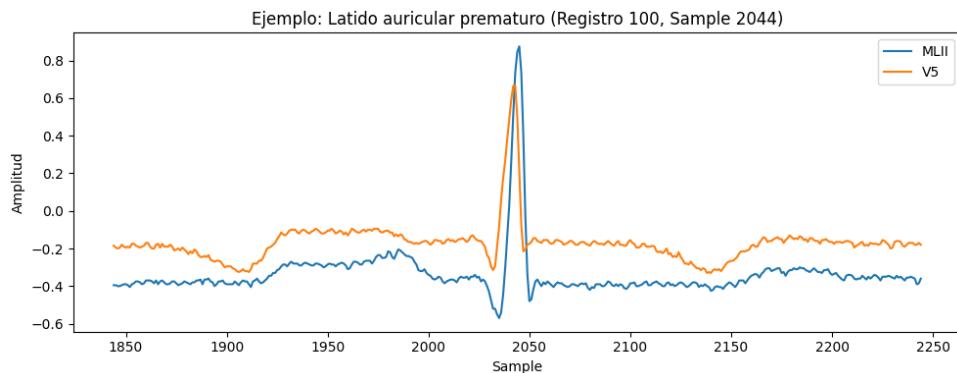


Figura 11: Latido auricular prematuro. Un latido que se origina en las aurículas antes de lo esperado.

6.8. Ejemplo Extraido de Latido de Fusión (F)

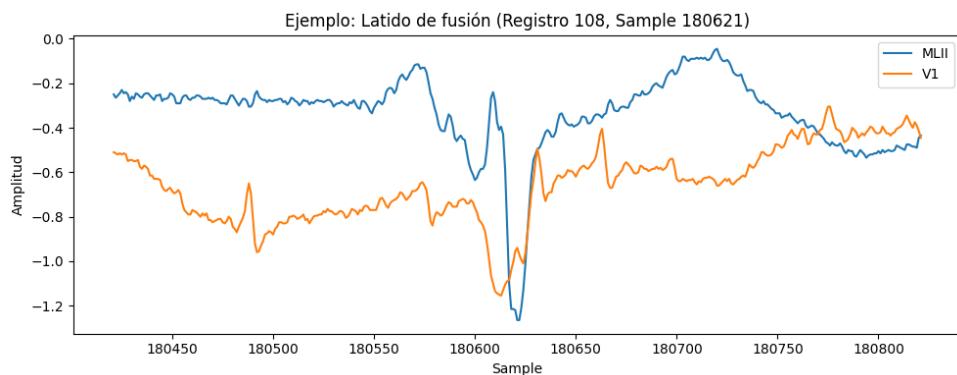


Figura 12: Latido de fusión. Este latido ocurre cuando dos impulsos se combinan para generar un solo latido.

6.9. Ejemplo Extraido de Latido de Fusión Ventricular (/)

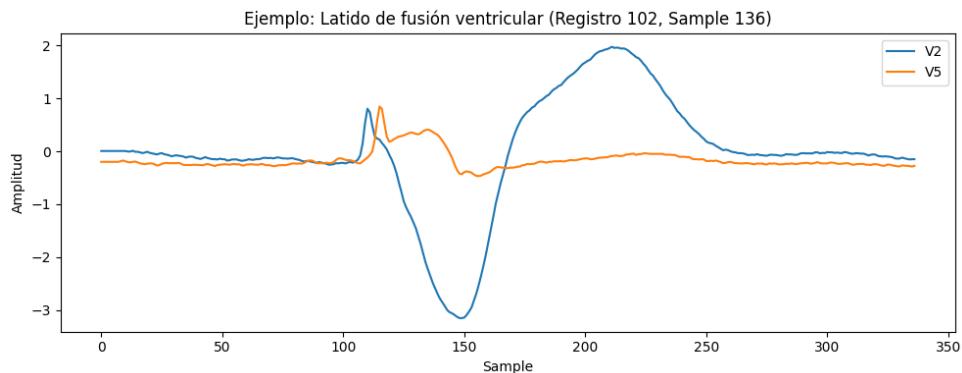


Figura 13: Latido de fusión ventricular. Ocurre cuando un latido ventricular prematuro se fusiona con el ritmo normal.

6.10. Ejemplo Extraido de Latido de Escape Nodal (j)

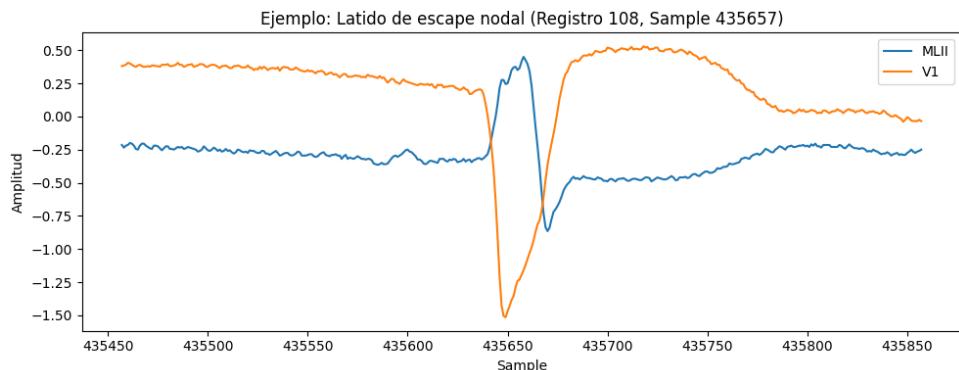


Figura 14: Latido de escape nodal. Este latido ocurre cuando el nodo sinoauricular no funciona y el nodo AV toma el control.

6.11. Ejemplo Extraido de Latido de Escape Ventricular (E)

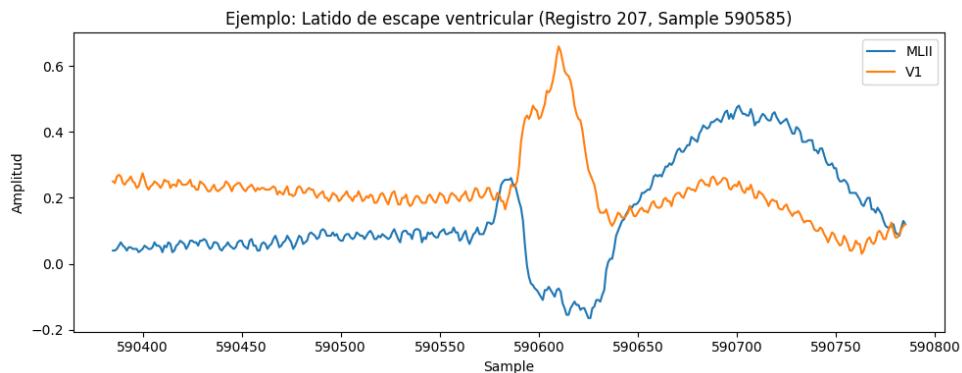


Figura 15: Latido de escape ventricular. Cuando el nodo AV no funciona, los ventrículos generan su propio latido.

6.12. Ejemplo Extraido de Latido Auricular Aberrante (a)

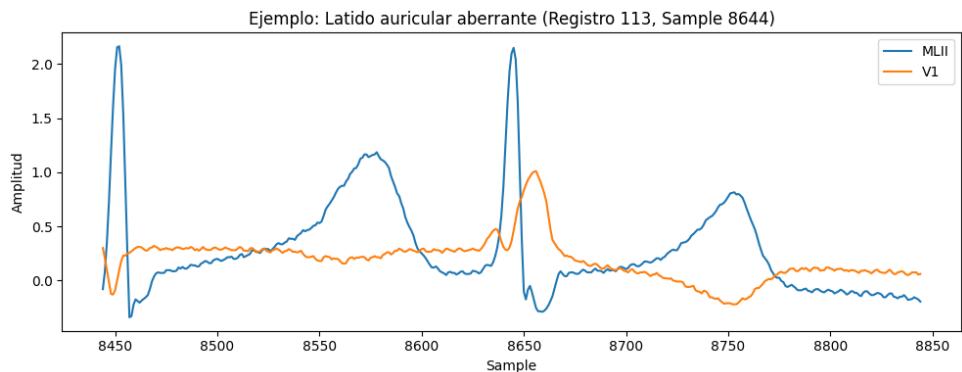


Figura 16: Latido auricular aberrante. Este latido ocurre debido a una conducción anómala en las aurículas.

6.13. Ejemplo Extraido de Latido de Escape de la Unión (J)

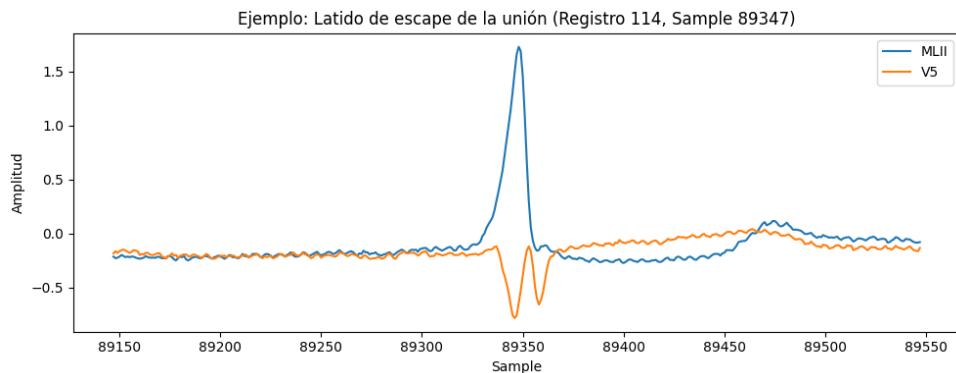


Figura 17: Latido de escape de la unión. Ocurre cuando el nodo AV toma el control en ausencia de la actividad del nodo SA.

6.14. Ejemplo Extraido de Latido de Marcapasos (S)

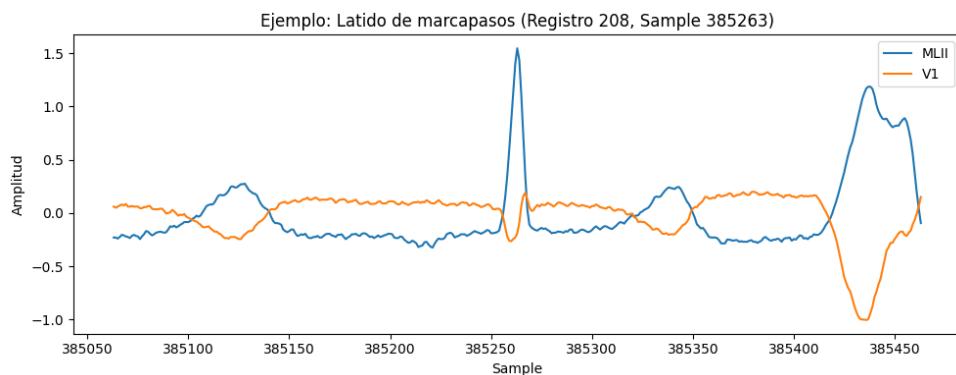


Figura 18: Latido de marcapasos. Este latido es generado por un marcapasos artificial.

6.15. Ejemplo Extraido de Latido Ventricular Aberrante (e)

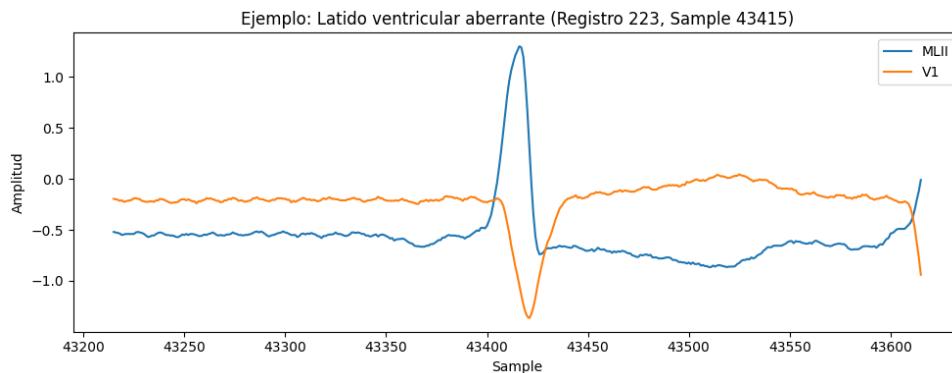


Figura 19: Latido ventricular aberrante. Es un latido que proviene de los ventrículos pero con una conducción anómala.

6.16. Ejemplo Extraido de Latido Desconocido (Q)

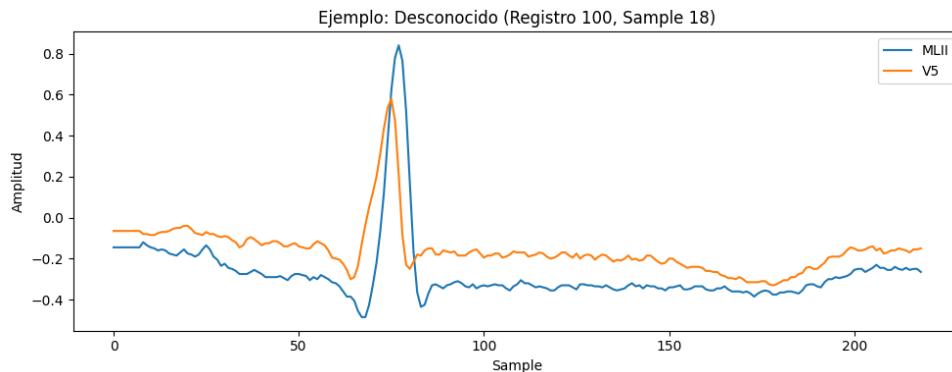


Figura 20: Latido desconocido. Este latido no se puede clasificar claramente con cresta prolongada

6.17. Ejemplo Extraido de Latido Desconocido (+)

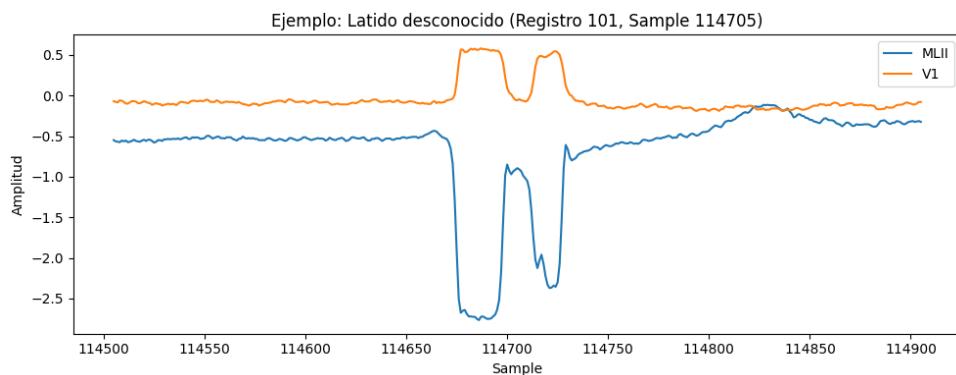


Figura 21: Latido desconocido. Similar al símbolo Q, este latido no puede ser identificado claramente principalmente de valle