Pipeline Ciencia de Datos

Piero Vizcarra

2025 Edition

pvizcarra.org.pe

¿Qué problemas identifican en el dataset?

```
Descripción
NaN                                    31087353
Latido normal                             75052
Latido de bloqueo de rama izquierda        8075
Latido de bloqueo de rama derecha          7259
Latido ventricular prematuro               7130
Latido de fusión ventricular               7028
Desconocido                                3153
Latido auricular prematuro                 2546
Latido de fusión auricular                  982
Latido de fusión                            803
Latido de escape nodal                      229
Latido auricular aberrante                  150
Latido de escape ventricular                106
Latido de escape de la unión                 83
Latido desconocido                           33
Latido ventricular aberrante                 16
Latido de marcapasos                          2
Name: count, dtype: int64
```

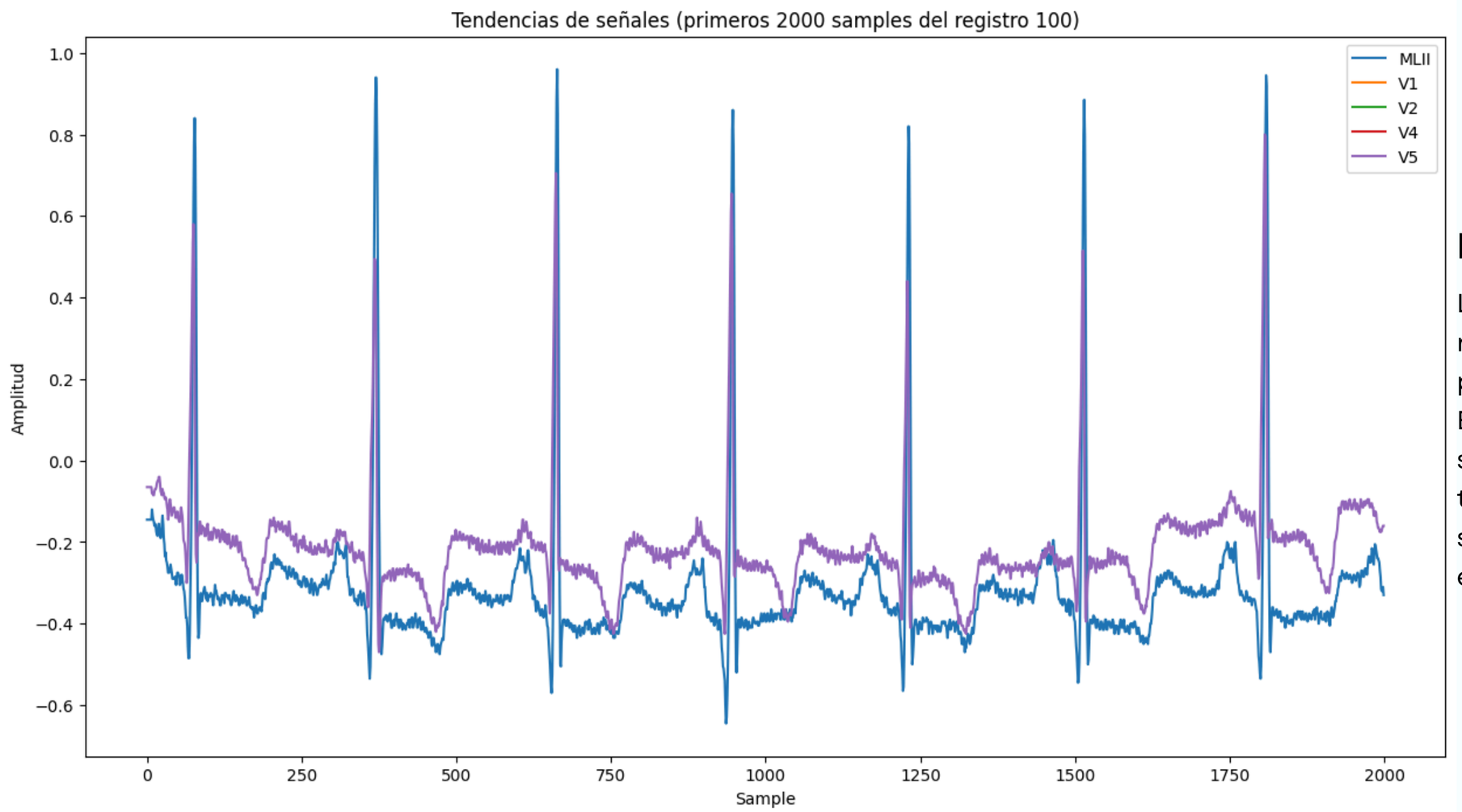
Desbalance de clases

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.

```
MLII              1300000
V5               27950000
Sample                 0
Símbolo          31087353
Descripción      31087353
Registro               0
V1                5200000
V2               28600000
V4               30550000
dtype: int64
```

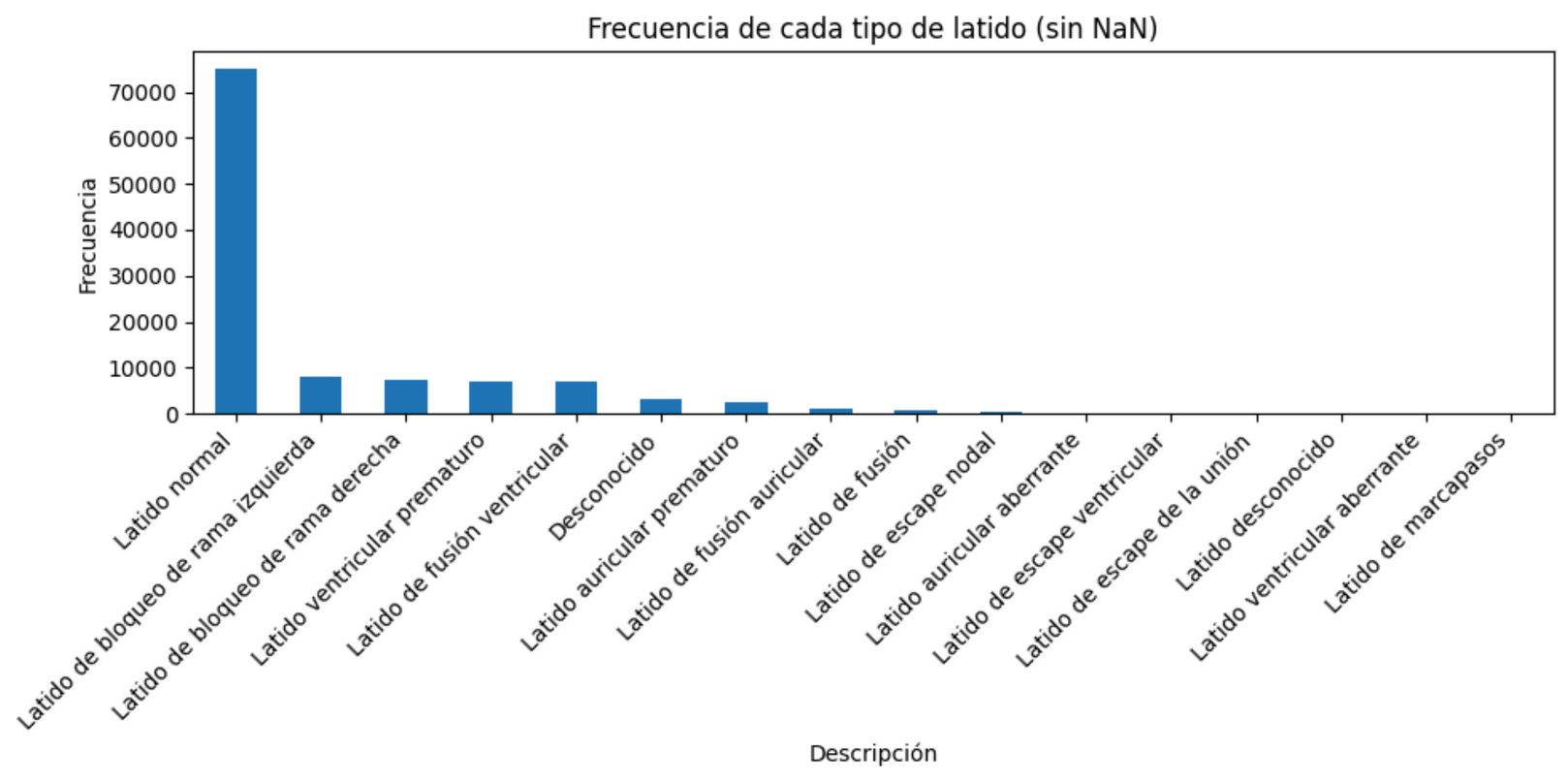
Variabilidad de derivaciones

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

Los datos contienen artefactos o ruido que dificultan la interpretación precisa de las señales.
Esto se debe debido a que son sensores y por si mismos los sensores tienen un grado de ruido esperable y se debe de normalizar los datos para ello

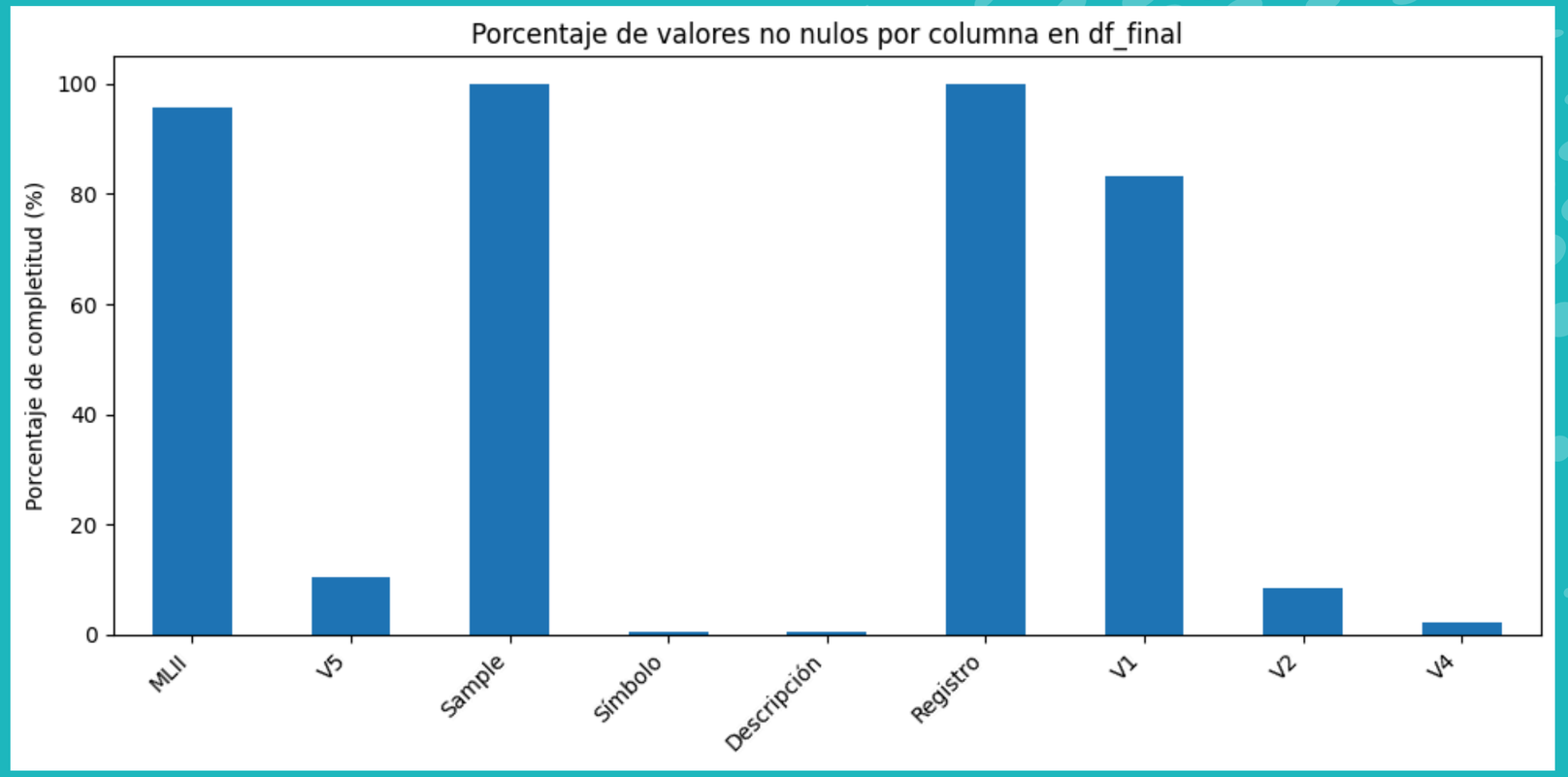


Escasa representación de latidos poco frecuentes

Clases de arritmias menos comunes tienen muy pocas muestras, lo que complica el entrenamiento de modelos robustos.

En el grafico podemos ver como al retirar los Nan aun asi tenemos un desbalance del Latido Normal siendo la clase principal pero que no es para nuestro objetivo, dado que buscamos representar las arritmias

¿Qué descubrieron al analizar los datos?



La derivación MLII está presente en casi todos los todos los registros, aun asi 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.
- 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 distorcia 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.

```python
# Mostrar los registros únicos donde MLII es NaN
registros_con_nan_mlii = df_final[df_final['MLII'].isnull()]['Registro'].unique()
print("Registros con valores NaN en MLII:")
print(registros_con_nan_mlii)
```

```
Registros con valores NaN en MLII:
['102' '104']
```

```python
# Mostrar los registros únicos donde MLII es NaN
registros_con_nan_mlii = df_final[df_final['V4'].notnull()]['Registro'].unique()
print("Registros con valores en MLII:")
print(registros_con_nan_mlii)
```

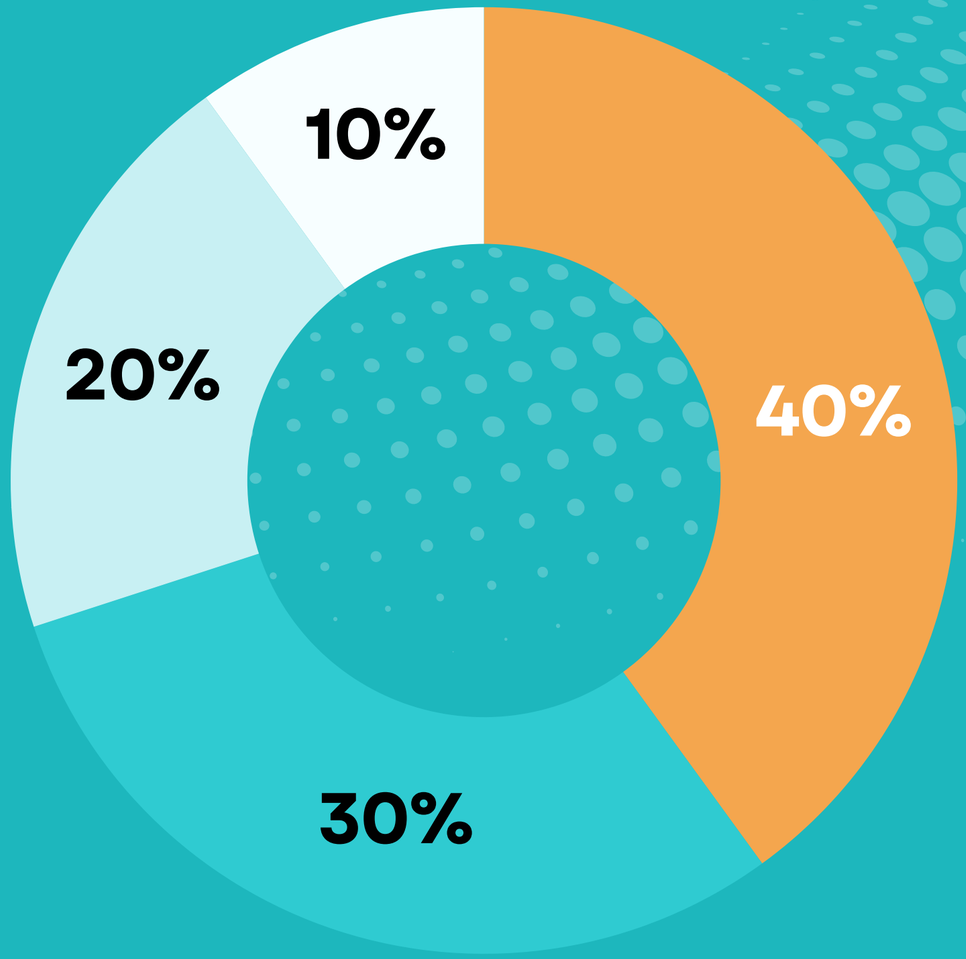
```
Registros con valores en MLII:
['124']
```

```python
df_final[df_final['Registro'] == '124']
```

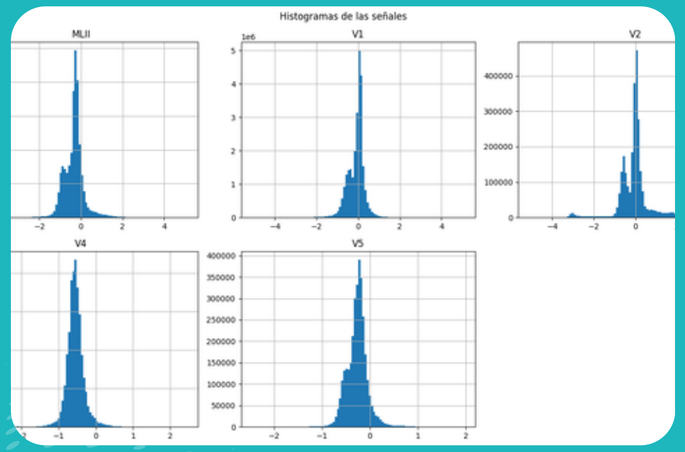
	MLII	V5	Sample	Símbolo	Descripción	Registro	V1	V2	V4
14300000	-1.000	NaN	0	NaN	NaN	124	NaN	NaN	-0.685
14300001	-1.000	NaN	1	NaN	NaN	124	NaN	NaN	-0.685
14300002	-1.000	NaN	2	NaN	NaN	124	NaN	NaN	-0.685
14300003	-1.000	NaN	3	NaN	NaN	124	NaN	NaN	-0.685
14300004	-1.000	NaN	4	NaN	NaN	124	NaN	NaN	-0.685
...	...	...	...	...	...	...	...	...	...
14949995	-0.745	NaN	649995	NaN	NaN	124	NaN	NaN	-0.595
14949996	-0.735	NaN	649996	NaN	NaN	124	NaN	NaN	-0.585
14949997	-0.720	NaN	649997	NaN	NaN	124	NaN	NaN	-0.585
14949998	-0.705	NaN	649998	NaN	NaN	124	NaN	NaN	-0.570
14949999	-1.280	NaN	649999	NaN	NaN	124	NaN	NaN	0.000

550000 rows × 9 columns

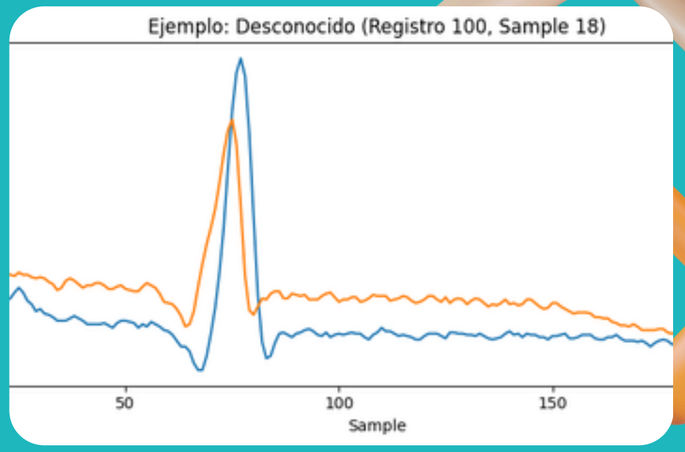
¿Qué reflejan los patrones de tendencia?



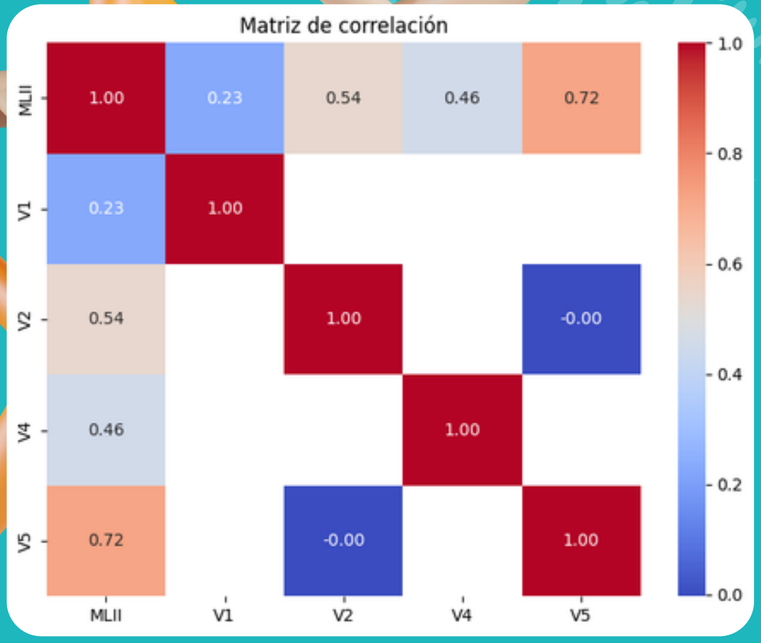
Tendencias



La señal MLII muestra un ritmo cardíaco regular para la mayoría de los latidos, con ondas P, QRS y T bien definidas.

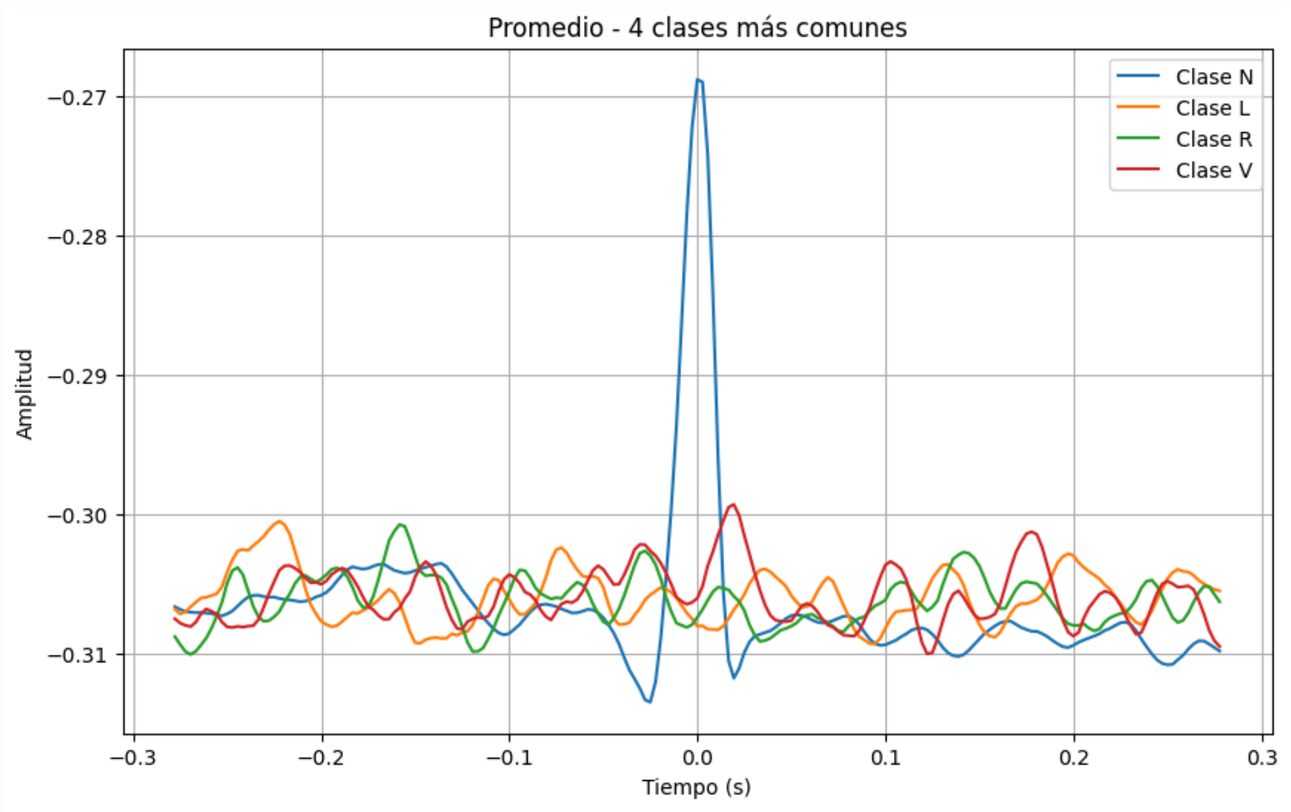


Los episodios de arritmias aparecen como interrupciones localizadas en estos patrones regulares, indicando eventos anómalos puntuales.

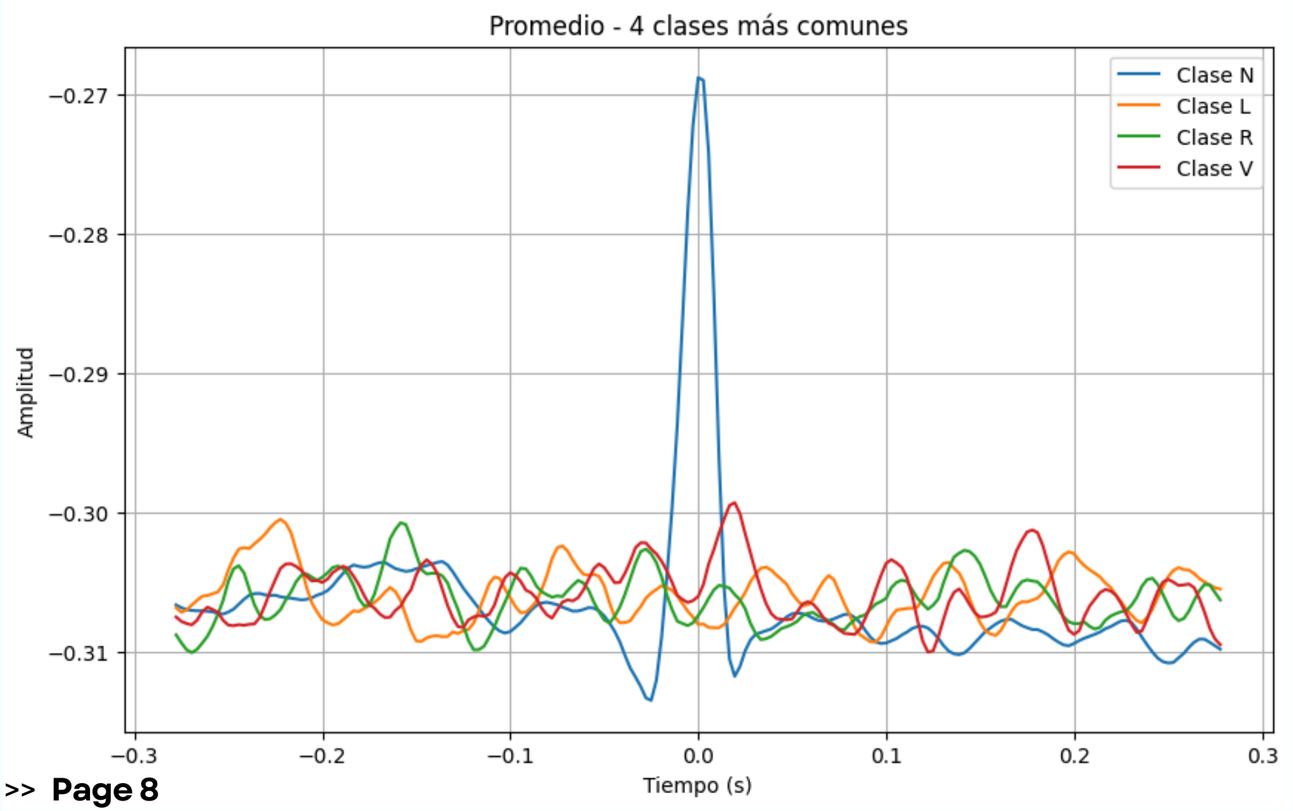


Estos patrones permiten identificar ventanas temporales relevantes para un análisis más detallado o para entrenamiento de modelos de clasificación.

¿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?



¿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?



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

>> **Page 8**

Gracias