# Práctica Análisis Exploratorio con Visualizaciones y Librería Personalizada

## 1. Descripción de Funciones Principales

### 1.1 remove_high_nulls(df)

Elimina automáticamente todas las columnas que contienen más del 20% de valores nulos dentro del DataFrame. Esto permite conservar únicamente variables con suficiente información útil y evitar sesgos generados por columnas muy incompletas.

Después del análisis correspondiente se detecto que el dataset tenía una mínima cantidad de datos nulos.

### 1.2 impute_missing_values(df)

Realiza la imputación de valores faltantes usando:

- **Media o mediana** para variables numéricas (dependiendo del comportamiento de la distribución).
- **Moda** para variables categóricas.

Garantiza que el dataset quede sin valores nulos y mantenga coherencia estadística.

### 1.3 detect_outliers_iqr(df)

Detecta valores atípicos mediante el método **IQR (Interquartile Range)**. Permite identificar puntos extremos que pueden distorsionar modelos o análisis estadísticos.

### 1.4 check_data_completeness(df)

Genera un resumen con:

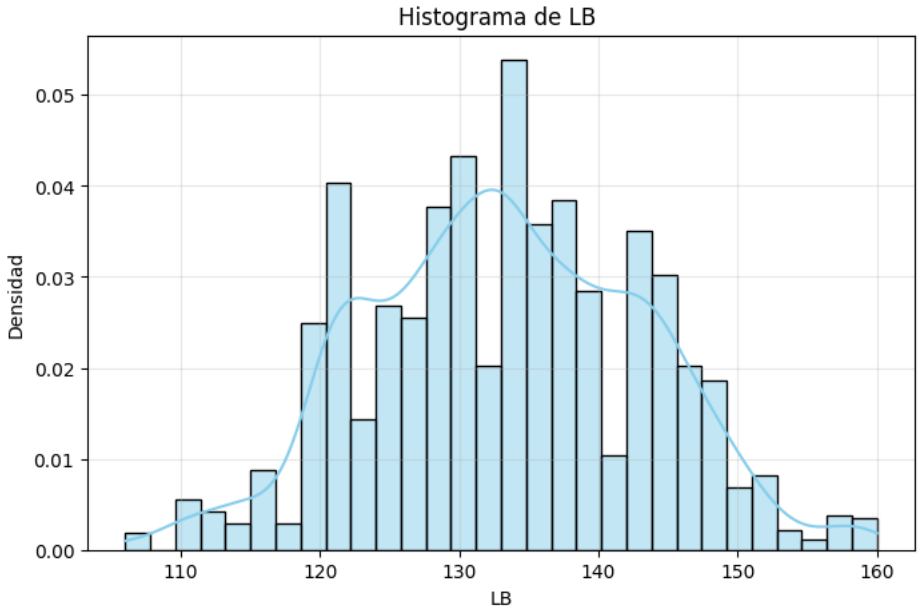
- Conteo de nulos
- Porcentaje de completitud
- Estadísticos clave (min, max, std, var, IQR)
- Clasificación automática como variable **continua** o **discreta**

---

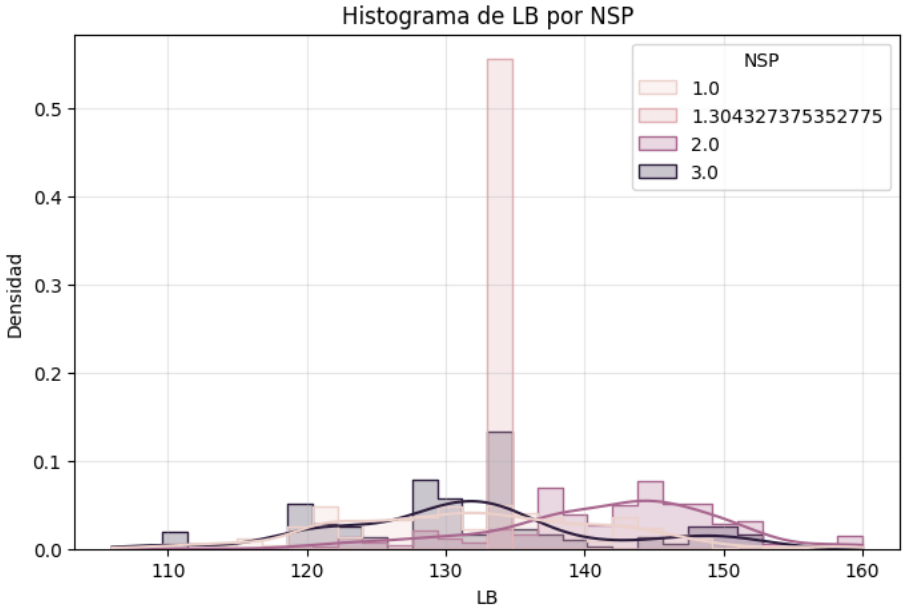
## 2. Visualizaciones Generadas

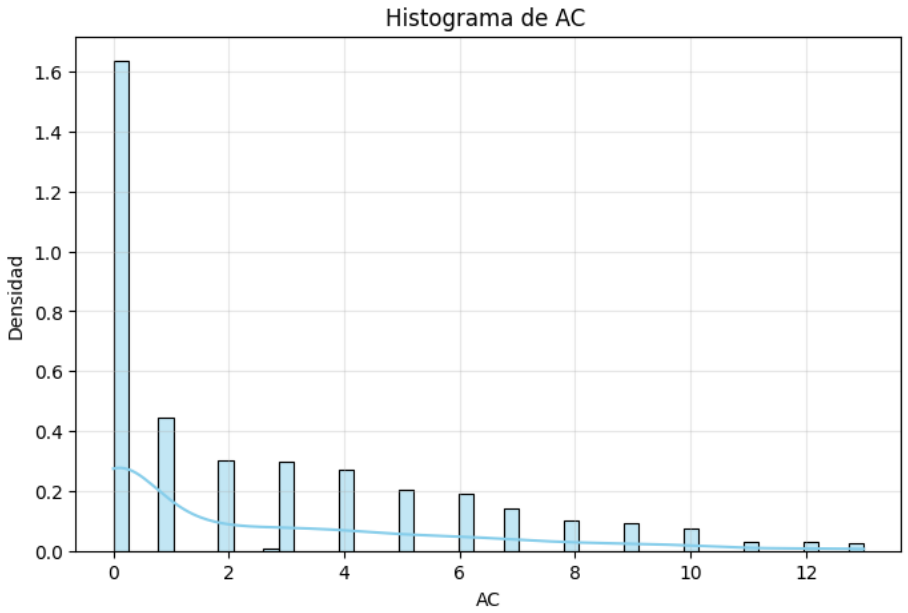
### 2.1 Histograma de la variable LB

Los siguientes histogramas nos permiten visualizar la forma de la distribución de cada variable, detectar posibles outliers, comparar comportamiento entre clases y decidir si se requieren transformaciones o técnicas de preprocesamiento adicionales.




### 2.1.1 Histograma por grupo segmentado

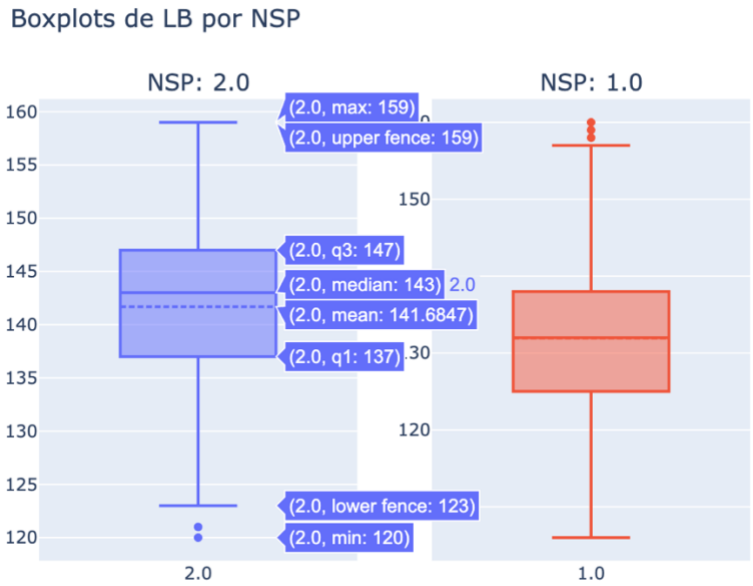


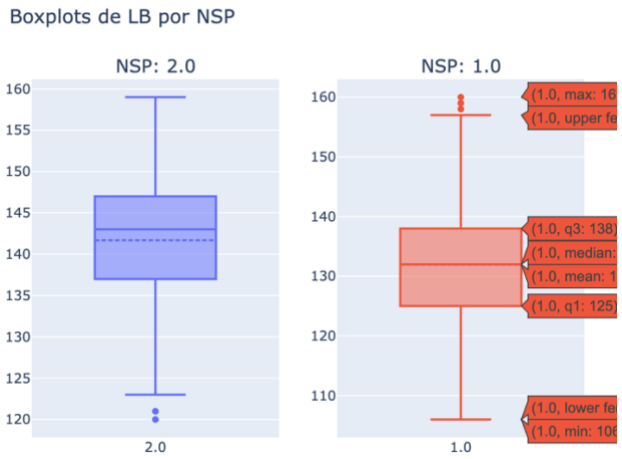



---

## 2.2 Boxplots por clase (NSP)

Los boxplots permiten identificar rápidamente la mediana, la variabilidad, y los valores atípicos de una variable, así como comparar su distribución entre las diferentes clases fetales (NSP). Esto ayuda a entender qué tan distinta es la respuesta fisiológica entre fetos normales, sospechosos y patológicos y facilita detectar patrones relevantes para diagnóstico y modelado.






---

## 2.3 Gráfica de Barras Horizontales

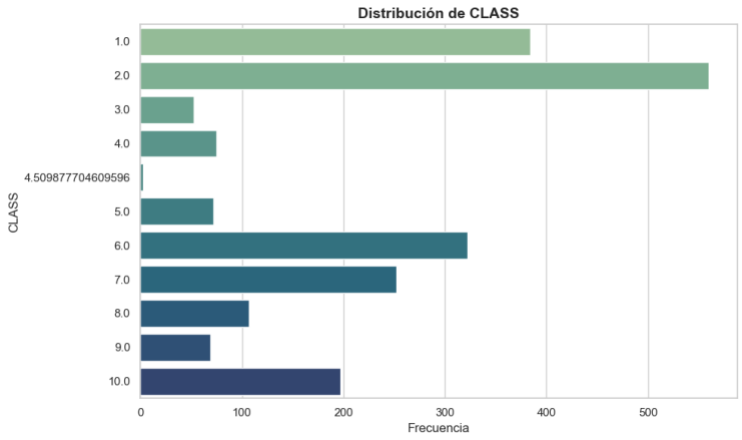
Las gráficas de barras horizontales permiten visualizar la frecuencia de categorías de forma clara y ordenada.

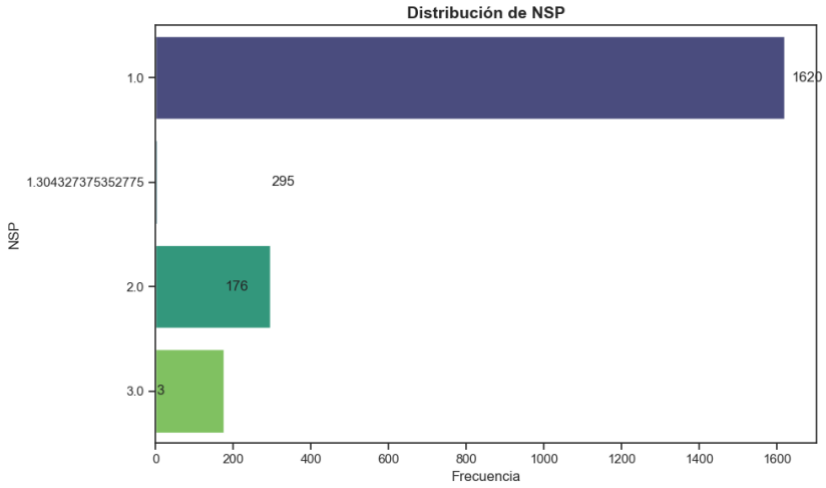
En el caso del dataset CTG:

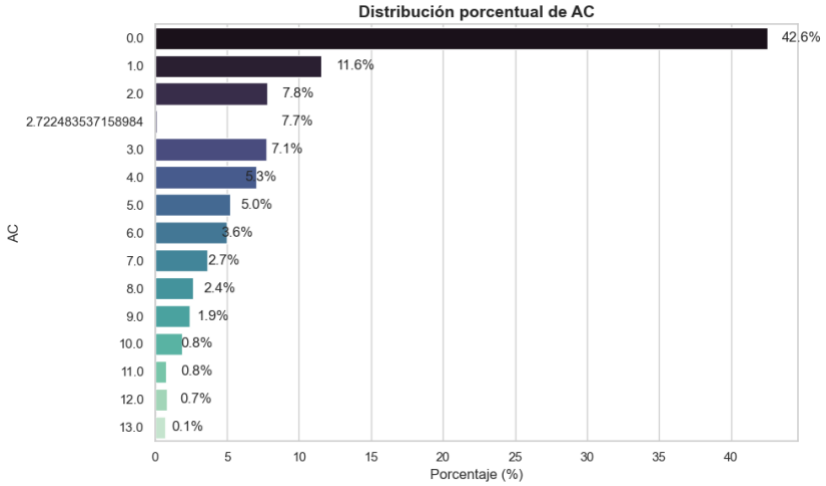
Sirven para observar qué categorías del estado fetal (NSP) aparecen con mayor o menor frecuencia.

Facilitan la comparación inmediata entre clases porque el ojo humano lee mejor longitudes horizontales.

Ayudan a detectar desbalance de clases, lo cual es importante para modelos predictivos (por ejemplo, si la clase Normal aparece mucho más que Suspect o Pathologic).



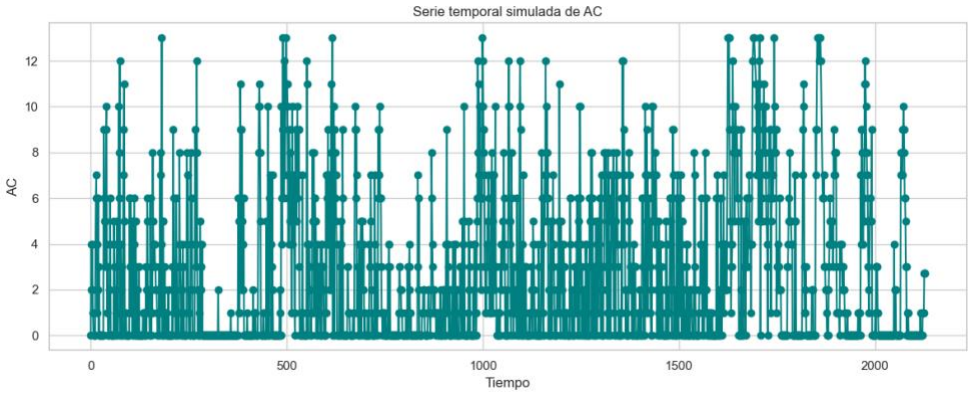


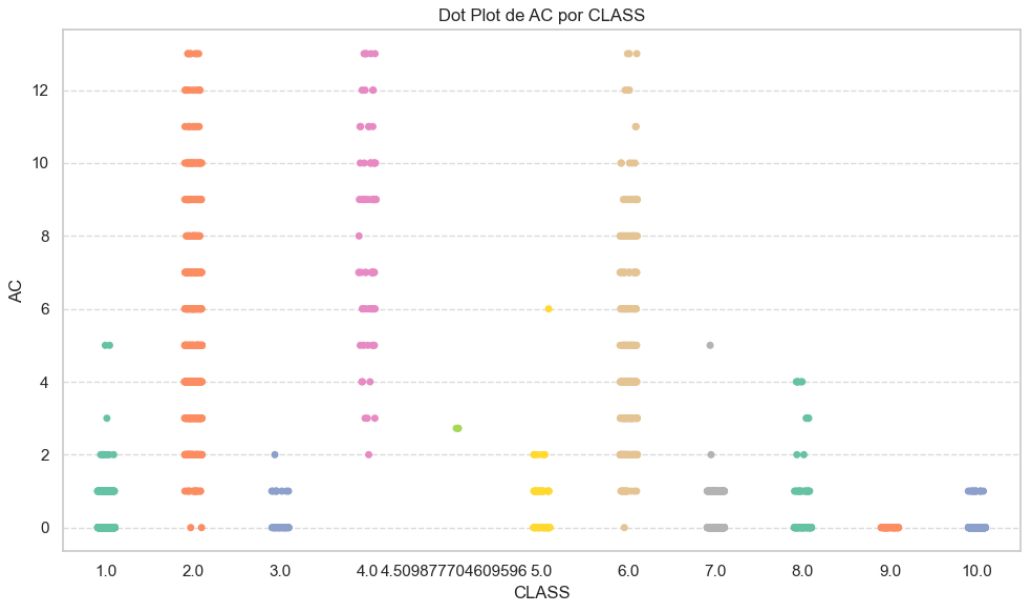

---

## 2.4 Lineas

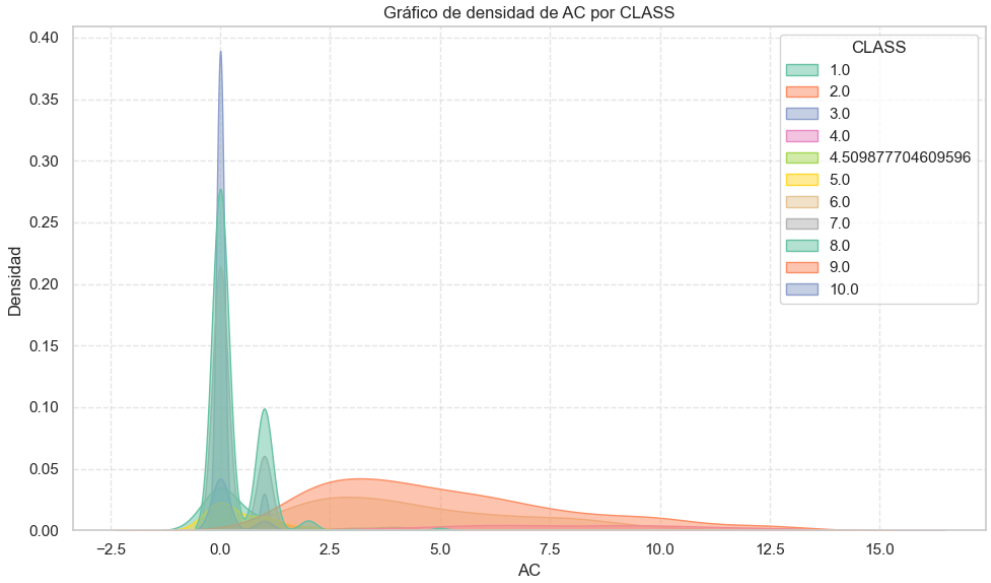
Las gráficas de líneas son muy útiles para mostrar la evolución de datos a lo largo del tiempo o cualquier otra variable continua



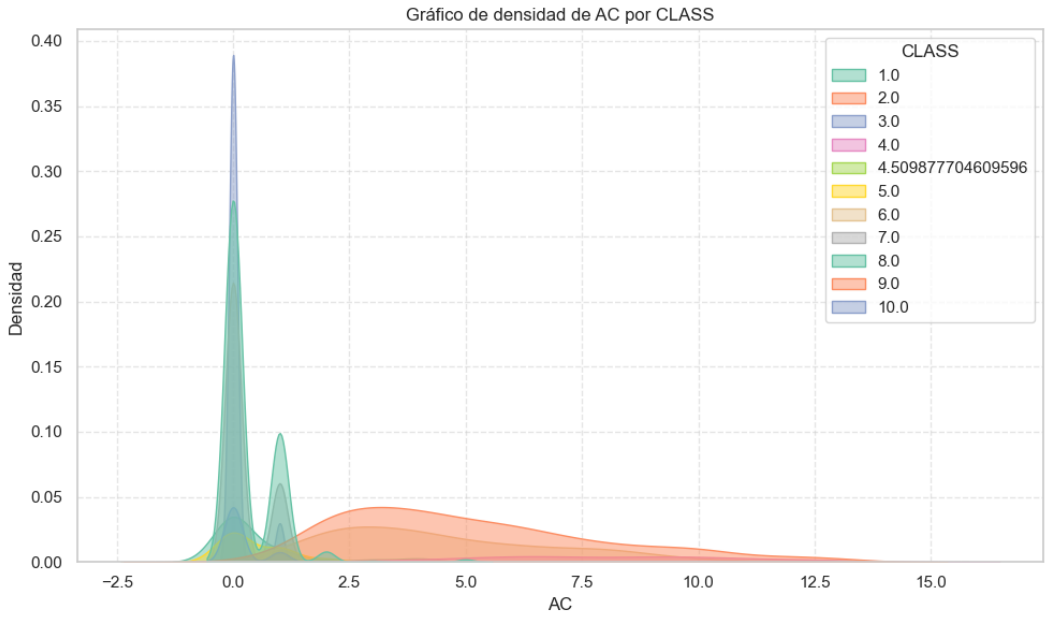
## 2.5 Dot Plots



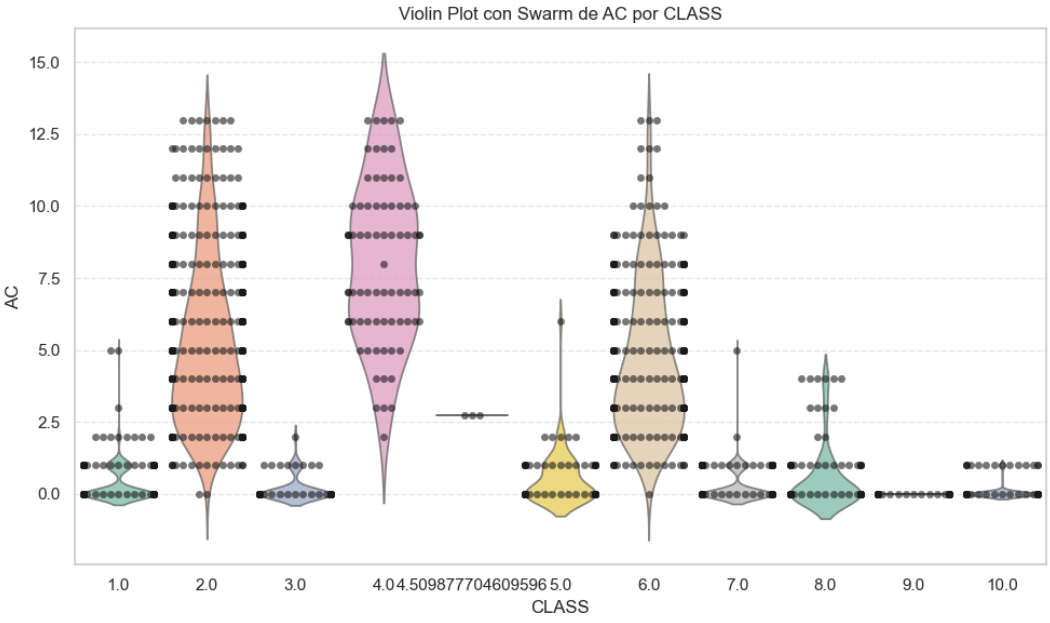

## 2.6 Densidad



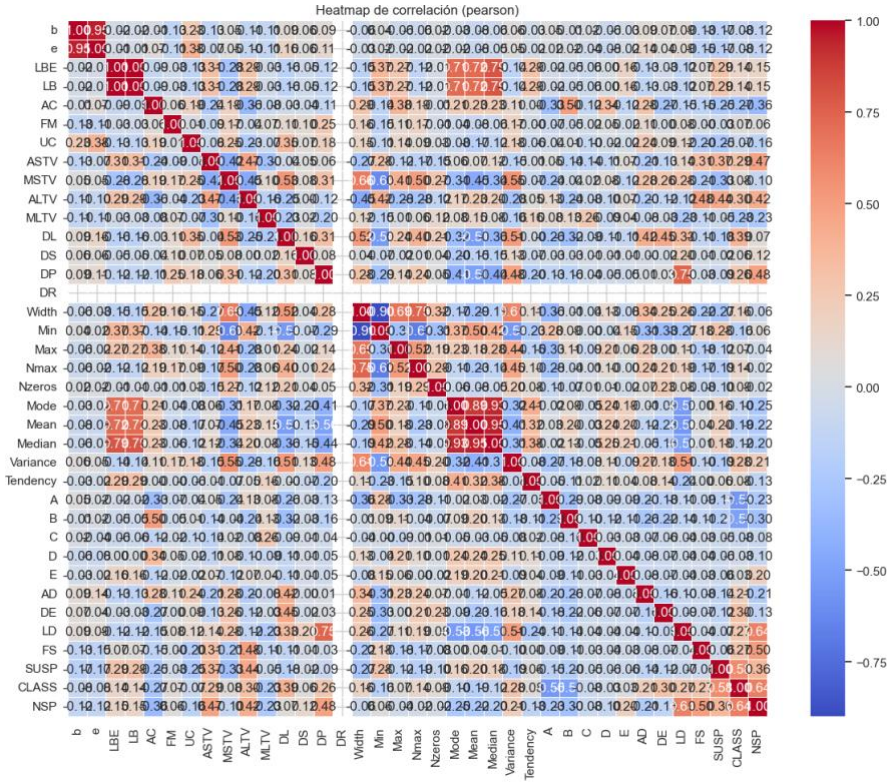
## 2.7 Densidad

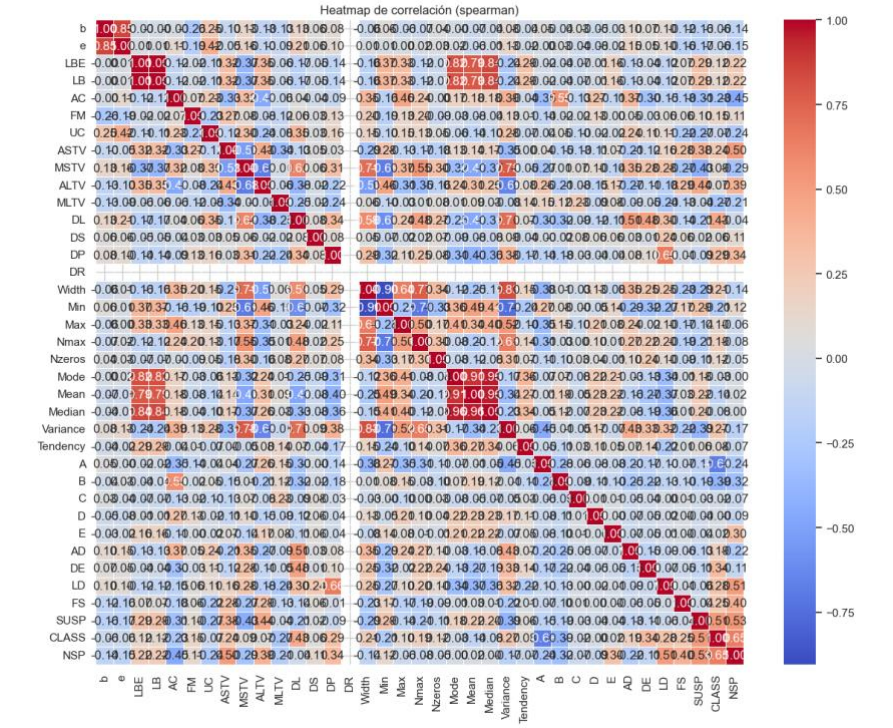



## 2.8 Violin



## 2.9 Heatmap


Heatmap de correlación (pearson)
b
e
LBE
LB
AC
FM
UC
ASTV
MSTV
ALTV
MLTV
DL
DS
DP
DR
Width
Min
Max
Nmax
Nzeros
Mode
Mean
Median
Variance
Tendency
A
B
C
D
E
AD
DE
LD
FS
SUSP
CLASS
NSP
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75


Heatmap de correlación (spearman)
b
e
LBE
LB
AC
FM
UC
ASTV
MSTV
ALTV
MLTV
DL
DS
DP
DR
Width
Min
Max
Nmax
Nzeros
Mode
Mean
Median
Variance
Tendency
A
B
C
D
E
AD
DE
LD
FS
SUSP
CLASS
NSP
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

## 3. Recomendaciones Analíticas

- La variable **LB** presenta diferencias importantes entre clases, por lo que podría aportar potencia discriminativa a modelos predictivos.
- Se detectaron outliers relevantes; se recomienda aplicar técnicas robustas como:
  - Winsorización
  - Transformaciones log o Yeo-Johnson
  - Escaladores robustos
- Se observa desbalance en la variable **NSP**. Se recomienda considerar:
  - Oversampling (SMOTE)
  - Class weights
- Algunas variables muestran correlación alta, lo que sugiere revisar colinealidad antes de modelar.
- Las distribuciones sugieren que algunas variables pueden necesitar normalización.