

Limpieza (Segunda iteración)

Dataset de Registros de Ordeño

Ya que esta iteración se enfoca en los registros de ordeño de cada vaca. El objetivo fue construir un dataset limpio y estructurado a nivel de ordeño individual para analizar el comportamiento productivo y fisiológico de la vaca, y preparar la base para un modelo posterior de análisis temporal.

En primer lugar, se realizó la carga de todos los archivos CSV disponibles en la carpeta de ordeño correspondiente en Drive. Estos archivos fueron **concatenados en un único dataframe**, manteniendo la granularidad por evento de ordeño.

Posteriormente, se aplicó un pipeline de limpieza. Como parte de este pipeline, se eliminaron columnas consideradas **administrativas o de control**, que no aportan información útil para la modelación ni para el entendimiento fisiológico de la vaca. Entre estas columnas se encuentran:

- Usuario, RCS (* 1000 células / ml), Razón de la desviación, Ubre, Pezón, Patada, Incompleto, Pezones no encontrados

Estas variables se descartan por dos motivos principales:

1. **No aportan valor analítico directo** al comportamiento productivo (por ejemplo, campos de usuario o razones internas del sistema).
2. **Introducen ruido operativo** (marcadores de eventos como “patada”, “incompleto” o descriptores de calidad de registro) que no se utilizarán en el modelo final de producción.

Finalmente, se aplicó una limpieza estricta de filas con valores nulos. Esta decisión se toma de forma controlada, después de haber transformado y renombrado columnas críticas, con la intención de garantizar que el dataset resultante esté **libre de valores faltantes** y listo para análisis estadístico y modelado.