# Informe Técnico – Sistema de Análisis de Contaminación del Agua Coinvestigación1

Autor: Nicolás Ramírez Villamizar

Fecha: 24 de octubre de 2025

## Introducción

El presente informe técnico describe en detalle el funcionamiento e integración del sistema Coinvestigación1, una plataforma avanzada diseñada para el análisis de contaminación en muestras de agua mediante técnicas de voltametría y procesamiento inteligente de datos. Este sistema combina herramientas de ingeniería de software, aprendizaje automático, y procesamiento de señales, con el fin de estimar concentraciones en ppm (partes por millón) de contaminantes metálicos a partir de mediciones obtenidas con sensores PalmSens y archivos .pssession.

## Arquitectura General del Sistema

El sistema se compone de varios módulos interconectados que trabajan en conjunto para permitir la adquisición, procesamiento, análisis y visualización de los datos. A continuación se presenta una descripción general del flujo completo:

1. El dispositivo PalmSens genera archivos .pssession que contienen los ciclos voltamétricos de la muestra analizada.  
2. El módulo pstrace\_session.py procesa estos archivos, extrayendo los valores de corriente y voltaje.  
3. El sistema aplica un análisis de Componentes Principales (PCA) entrenado previamente sobre una base de referencia.  
4. El modelo predictivo RidgeCV estima los valores de concentración (ppm) de contaminantes.  
5. La interfaz gráfica (interfaz\_grafica.py) muestra los resultados, gráficas PCA y clasificaciones del agua.  
6. Todos los resultados se almacenan en una base de datos interna y se exportan en formato CSV.

## Módulo de Entrenamiento – train\_memory.py

Este módulo se encarga de preparar el modelo de aprendizaje automático. A partir del archivo Excel de referencia ('Datos de referencia\_laboratorio, matriz\_2\_UNAL 191925.xlsx'), el sistema limpia y transforma los datos, eliminando columnas vacías, corrigiendo valores NaN y normalizando los datos. Posteriormente, aplica un PCA para reducir la dimensionalidad de los datos conservando el 99% de la varianza y entrena un modelo RidgeCV multisalida que aprende a correlacionar los espectros voltámetricos con las concentraciones de laboratorio (mg/L o ppm).

Durante el entrenamiento se generan tres archivos esenciales en la carpeta 'models/':  
- scaler.pkl → escalador estándar utilizado para normalizar nuevos datos.  
- pca.pkl → modelo PCA entrenado para la reducción de características.  
- model.pkl → modelo RidgeCV final para la predicción de contaminantes.

## Módulo de Preprocesamiento – preprocess.py

Este módulo contiene las funciones necesarias para limpiar y transformar los datos crudos antes del entrenamiento o la predicción. Incluye la función baseline\_subtract(), que elimina el ruido de base de cada curva, y fit\_scaler(), que ajusta el escalado de los datos. Ambas funciones garantizan la coherencia del preprocesamiento tanto para el entrenamiento como para las futuras predicciones en campo.

## Módulo de Procesamiento de Sesiones – pstrace\_session.py

Este componente actúa como el núcleo de análisis en tiempo real. Se encarga de cargar los archivos .pssession generados por el equipo PalmSens y extraer sus mediciones. Cada archivo puede contener múltiples mediciones, y cada medición múltiples ciclos voltamétricos. El sistema selecciona el tercer ciclo como referencia, calcula las señales procesadas y aplica el PCA entrenado junto al modelo RidgeCV para obtener una estimación de ppm promedio. Además, clasifica la muestra según el nivel de contaminación detectado.

El módulo también genera un archivo CSV consolidado (matriz\_pca.csv) con los valores PCA, las predicciones ppm, el nivel de contaminación y la clasificación final del agua.

## Interfaz Gráfica de Usuario – interfaz\_grafica.py

La interfaz gráfica permite visualizar todo el proceso y sus resultados de manera intuitiva. Está desarrollada en Python con Tkinter y Matplotlib, presentando pestañas que permiten observar curvas voltamétricas, gráficas PCA y los resultados de predicción. Al cargar un archivo .pssession, el sistema procesa automáticamente la sesión, aplica el modelo entrenado y muestra el nivel de contaminación junto con las curvas y componentes principales.

El bloque 'Mostrar PCA' implementa una visualización de la varianza acumulada, mostrando cuánta información conserva cada componente. Además, se puede exportar la gráfica en formato PNG directamente desde la interfaz.

## Resultados y Ejecución del Sistema

Durante las pruebas realizadas, se procesaron varios archivos .pssession con distintas muestras de agua. El sistema logró identificar correctamente los niveles de contaminación, generando valores de ppm promedio cercanos a 1.0, lo que demuestra una adecuada calibración del modelo y la coherencia entre el PCA y el RidgeCV. Los logs muestran la correcta detección de mediciones, ajuste de longitudes, predicción final y generación del archivo CSV consolidado.

El flujo completo evidenció una integración estable entre el modelo entrenado, el procesamiento de sesión y la visualización en la interfaz. Cada componente trabaja de forma sincronizada, garantizando precisión y trazabilidad en cada análisis.

## Conclusión

El sistema Coinvestigación1 representa una implementación robusta y modular para el análisis inteligente de contaminación en agua. Integra procesamiento de señales, aprendizaje automático y visualización avanzada dentro de una arquitectura bien definida. Gracias al uso del PCA, el sistema logra una gran reducción de dimensionalidad manteniendo la información relevante, mientras que el modelo RidgeCV permite predecir con precisión la concentración de contaminantes. El resultado es una herramienta práctica, científica y técnicamente sólida para el monitoreo de calidad del agua.