

# **Fases del Sueño**

## **Trabajo Final ADS**

Realizado por Amparo Gálvez, Lidia Moreno y Alfredo Quiles

# Índice

01

**Introducción**

02

**Etiquetado de las Fases**

03

**Características**

04

**Modelado**

05

**Interpretabilidad**

06

**Conclusión**

# Introducción

## Datos Iniciales

- Se ha analizado el **Subgrupo 3** en el que hay **10 pacientes sanos**
- Dimensiones Iniciales:
  - 19 canales para cada paciente.
  - Diferencias en duración de grabación (~8 horas por paciente).
  - Frecuencia de muestreo: 200 Hz, asegurando alta resolución temporal.

## Formato Verificado:

- Archivos cargados como objetos RawEDF de MNE.
- Confirmación de:
  - 19 canales marcados como "buenos", sin señales dañadas.

**Conclusión:** Datos consistentes en estructura y adecuados para análisis.

# Eliminación de Canales Irrelevantes

**Motivación:** Simplificar el análisis y enfocarse en las señales clave para el estudio del sueño.

## Canales Seleccionados:

- **C4-A1, O1-A2 (EEG):** Capturan actividad eléctrica central y occipital del cerebro.
- **LOC-A2 (EOG):** Registra movimientos oculares (sueño REM).
- **X1 (EMG):** Evalúa el tono muscular (diferenciación REM/NREM).

## Verificación:

- Comprobación del formato de los datos tras eliminar canales para garantizar que la estructura resultante sea consistente.

**Resultado:** Optimización computacional y relevancia clínica.

# Cambio de Frecuencia de Muestreo

**Nueva Frecuencia:** 100 Hz (reducida desde 200 Hz).

## Motivación:

- Optimizar tiempo de cálculo.
- Trabajar más adelante con YASA.
- Mantener resolución suficiente para capturar características relevantes.

**Resultado:** Balance entre precisión y eficiencia computacional.

# Filtrado de Señales

**Filtro pasa banda:** 0.3 Hz – 49 Hz.

## Objetivo:

- Eliminar ruido de baja y alta frecuencia.
- Preservar frecuencias relevantes para el análisis del sueño (ondas delta, theta, alfa y beta).

**Conclusión:** Mejora en la calidad de las señales para una clasificación precisa.

# Acceso a los datos

**Escalado:** Datos convertidos a microvoltios ( $\mu\text{V}$ ).

**Segmentos eliminados:** Últimos 30 minutos de grabación según el artículo base.

# Visualización y Verificación

## Graficación de señales tras el preprocesamiento:

- Confirmación de calidad visual y comportamiento esperado.

## Verificación de dimensiones:

- 4 canales seleccionados, cada uno con duraciones esperadas tras el remuestreo.

**Conclusión:** Todas las verificaciones aseguran que los datos procesados son consistentes, fiables y adecuados para el análisis de las fases del sueño.

# Etiquetado de las Fases

Se trabaja con datos polisomnográficos, utilizados para estudiar las fases del sueño. Las etiquetas de las fases no siguen el formato estándar de YASA, lo que complicaba el análisis.

## Problema:

- Codificación original: 0 -> Vigilia (W); 1 -> N1; 2 -> N2; 3 -> N3; 5 -> REM
- Yasa requiere que las fases REM esté etiquetadas hasta 4 (5 no es válido).
- Los datos contenían irregularidades al final de los registros (últimas 30 muestras).

## Modificaciones:

- **Cambio de etiquetas:** Transformar la etiqueta 5 como 4, para ajustarla al estándar de YASA.
- **Eliminación de muestras finales:** Eliminar las últimas 30 muestras de cada registro.

## Impacto clínico:

- **Compatibilidad:** Los datos son utilizables en herramientas avanzadas como YASA, facilitando análisis automatizados.
- **Reducción de ruido:** Mejora de la precisión del análisis al eliminar inconsistencias.
- **Diagnóstico fiable:** Representación más exacta de las fases del sueño.

## Cod. YASA

0 = Wake  
1 = N1 sleep  
2 = N2 sleep  
3 = N3 sleep  
4 = REM sleep

# Características

Para mejorar la precisión en la clasificación se analizan señales EEG, EMG y EOG complementadas con características avanzadas. Estas capturan patrones y dinámicas de cada canal.

## Filtro Pasa-banda

- Aislar frecuencias relevantes.
- Eliminar ruido y centrar el análisis en rangos específicos.
- Mejora la calidad de las señales.

## Desviación Estándar

- Mide la dispersión de los datos.
- Identifica variabilidad en EMG y EOG.
- Detecta eventos atípicos.

## Varianza

- Cuantifica la variabilidad total en las señales.
- Visión global de los cambios en EEG, EMG y EOG.

## Percentiles

- Información sobre la distribución de los valores.
- Identifica diferencias sutiles entre etapas.

## PSD

- Análisis de espectros de frecuencias de las señales.
- Identifica las dinámicas de las ondas cerebrales para clasificar estados REM o NREM.



# Características:

## Análisis EMG

Evaluar actividad muscular para detectar movimientos asociados con trastornos del sueño.

## Análisis EEG

Evaluar la actividad cerebral para identificar patrones relacionados con las diferentes fases del sueño.

## Análisis EOG

Estudiar los movimientos oculares para identificar patrones asociados con las diferentes fases del sueño, especialmente la fase REM.



# Características:

## Análisis EMG

### Desviación Estándar (EMG\_STD):

- Valores altos indican actividad muscular más pronunciada
- Asociados a movimientos periódicos de las extremidades o bruxismo durante el sueño.
- Relevante para detectar trastornos relacionados con movimientos involuntarios en el sueño.

### Percentiles (P25, P50, P75):

- Muestran una variabilidad moderada en los movimientos musculares.
- Los valores altos pueden estar relacionados con trastornos como:
  - Síndrome de piernas inquietas.
  - Movimientos periódicos de extremidades.

# Características:

## Análisis EEG

### Desviación Estándar (EEG\_STD):

- Valores altos: Mayor actividad cerebral, asociada a sueño ligero o vigilia.
- Valores bajos: Fases de sueño profundo con menor variabilidad.

### Percentiles (P25, P50, P75):

- Mayor variabilidad: Observada en fases REM.
- Percentil 50: Relacionado con actividad cerebral promedio.

### Potencia de Bandas Espectrales:

- **Delta (Sueño profundo):** Potencia baja.
- **Theta (Sueño ligero):** Moderada con transiciones ligeras.
- **Alfa y Beta (Vigilia/Sueño ligero):** Picos en beta asociados con mayor actividad cerebral.

### Centro de Frecuencia en Banda Beta:

- Varía entre 19-24 Hz, típico de estados de vigilia o actividad elevada.

# Características:

## Análisis EOG

### Desviación Estándar (EEG\_STD):

- Refleja la variabilidad de los movimientos oculares.

### Percentiles (P25, P50, P75):

- Muestran amplitudes de los movimientos oculares.
- Percentil 75 más alto: Indica movimientos más amplios y variables, característicos de la fase REM.

### Potencia de Bandas Espectrales:

- **Delta y Theta (Sueño no REM):** Potencias bajas, reflejan fases de menor actividad ocular.
- **Alfa y Beta (Sueño REM):** Aumentan en REM, correlacionándose con actividad cerebral elevada y movimientos oculares rápidos.

# Modelado



# Estrategia de Partición de los Datos

## Datos de Entrada y Salida

- **Características de Entrada (X):**
  - Señales fisiológicas derivadas de EEG, EOG y EMG.
- **Etiquetas de Salida (Y):**
  - Fase del sueño: Clasificación multiclase con 5 categorías.

## División Fija de Datos

- **Conjunto de Entrenamiento:**
  - Registros de los pacientes 3 a 10.
  - Tamaño: 6754 muestras.
- **Conjunto de Prueba:**
  - Registros de los pacientes 1 y 2.
  - Tamaño: 1835 muestras.

## Motivación de la Estrategia

La partición fija asegura que los datos de prueba representen un desafío realista, evaluando la capacidad del modelo para clasificar correctamente fases del sueño en pacientes nuevos.

# Estrategia de Clasificación Multiclase

## Razón para Elegir Clasificación Multiclase

- **Naturaleza del Problema:** Clasificar etapas del sueño en 5 categorías distintas (0 a 4).
  - Cada clase representa una etapa fisiológica diferente con características únicas.
- **Preservación de Información:**
  - Evitar pérdida de detalles que ocurriría al agrupar clases o realizar binarización.
- **Relevancia Clínica:**
  - La identificación precisa de cada etapa es crucial para el análisis y diagnóstico de trastornos del sueño.

## Ventajas del Enfoque Multiclase con XGBoost

- **Soporte Nativo:** XGBoost permite manejar clasificación multiclase con multi:softprob y genera probabilidades para cada clase.
- **Manejo de Datos sin Normalización:**
  - XGBoost es robusto frente a diferencias de escala en las características.
  - Ahorro de tiempo y esfuerzo en preprocesamiento.
- **Capacidad de Generalización:**
  - Parámetros ajustados para evitar sobreajuste, garantizando un modelo confiable en datos no vistos.

## Resultado Final:

- El modelo multiclase permite capturar patrones específicos de cada etapa del sueño, mejorando su utilidad en aplicaciones clínicas y su capacidad de discriminación.

# Selección de Hiperparámetros y Validación Cruzada

## Reto Inicial: **Sobreajuste**

El modelo inicial mostraba un alto rendimiento en el entrenamiento pero un bajo rendimiento en validación.

## Ajuste de Hiperparámetros

- Tasa de Aprendizaje: Reducida a 0.005 para un entrenamiento más lento y preciso.
- Profundidad Máxima del Árbol: Limitada a 4 para evitar un modelo excesivamente complejo.
- Regularización L1 y L2: Añadir penalizaciones ( $\alpha = 0.5$ ,  $\lambda = 1.5$ ) para controlar la complejidad del modelo.
- Submuestreo: 70% de filas y columnas para mejorar la generalización.

## Validación Cruzada

Para evaluar el rendimiento del modelo y detectar problemas de sobreajuste, se implementó validación cruzada con 5 pliegues estratificados:

- Ventajas:
  - Permite evaluar el modelo en diferentes particiones del conjunto de entrenamiento.
  - Proporciona una estimación más robusta del rendimiento.

## El reto del sobreajuste fue abordado combinando:

- **Ajuste de hiperparámetros** para simplificar el modelo.
- **Validación cruzada** para asegurar que el rendimiento sea consistente en diferentes particiones de los datos.



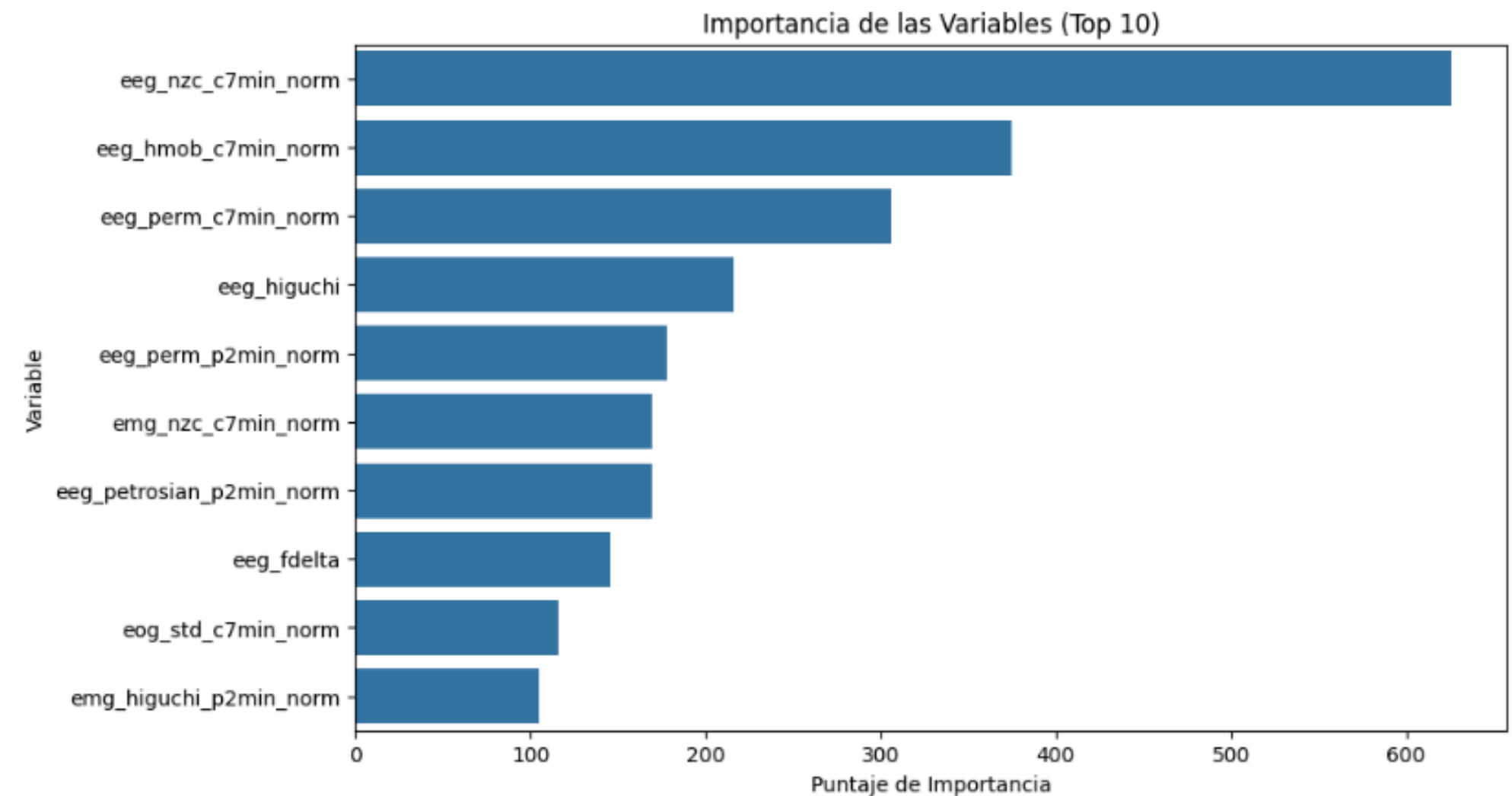
# Selección de Variables

**Estrategia:** Se utilizó el **método Gain** de XGBoost para identificar las características más importantes. Gain calcula el aporte promedio al valor de pérdida para divisiones que usan una característica.

- Ventajas:
  - Simplificación del modelo para hacerlo más interpretable.
  - Reducción de costos computacionales.

## Decisión:

Se seleccionaron las 10 características más importantes para evaluar el impacto en el rendimiento frente al modelo completo.



# Comparación de Resultados

## Modelo Completo

### Conjunto Desarrollo:

- **Accuracy medio:** 84.54%.
- **AUC promedio:** 0.9720.

### Conjunto de Test:

- **Accuracy:** 80.82%.
- **AUC promedio:** 0.9583.

### AUC por Clase (Test):

- **Clase 0:** 0.9841.
- **Clase 1:** 0.8920.
- **Clase 2:** 0.9465.
- **Clase 3:** 0.9785.
- **Clase 4:** 0.9905.

	precision	recall	f1-score	support
0	0.78	0.89	0.83	262
1	0.74	0.29	0.41	255
2	0.78	0.91	0.84	692
3	0.90	0.85	0.87	375
4	0.84	0.91	0.87	251
accuracy			0.81	1835
macro avg	0.81	0.77	0.77	1835
weighted avg	0.81	0.81	0.79	1835

- Mejor desempeño general en accuracy y AUC.
- Mayor capacidad para captar las variaciones de las diferentes clases.

## Modelo Reducido

### Desempeño en el Conjunto de Test:

- **Accuracy:** 76.00%.
- **Macro Avg (F1):** 0.71.
- **Weighted Avg (F1):** 0.74.

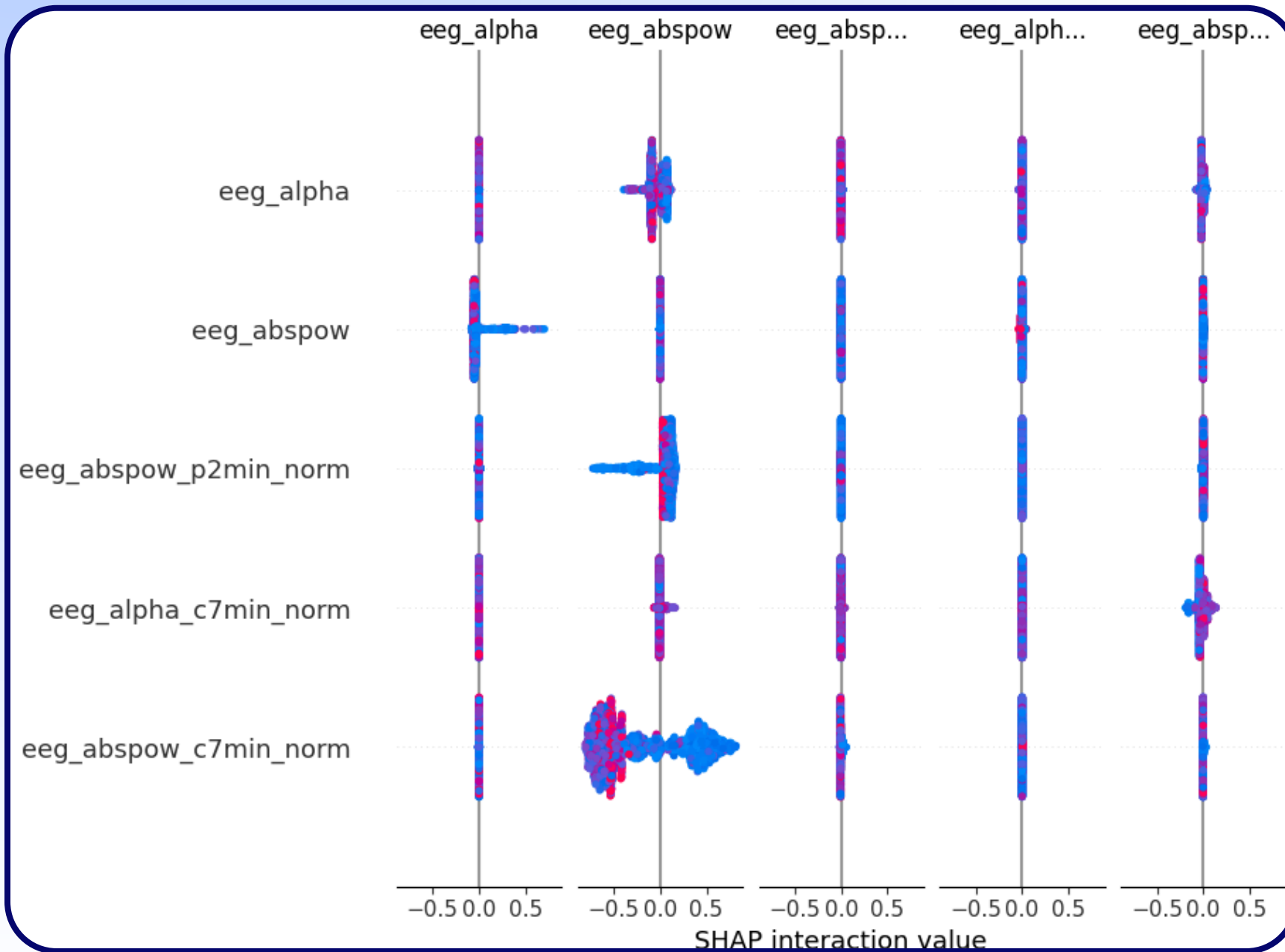
	precision	recall	f1-score	support
0	0.80	0.86	0.83	262
1	0.50	0.20	0.28	255
2	0.71	0.92	0.80	692
3	0.94	0.73	0.82	375
4	0.78	0.84	0.81	251
accuracy			0.76	1835
macro avg	0.75	0.71	0.71	1835
weighted avg	0.75	0.76	0.74	1835

- Reducción en el rendimiento global (Accuracy: 76%, AUC: menor).
- Facilita la interpretación y es más eficiente .

# Interpretabilidad



# Análisis con SHAP



- **Aplicación clínica:**

- SHAP permite visualizar el impacto de cada característica, facilitando la interpretación de las decisiones.
- Identificación de características clave

- **Objetivo:** Interpretar las predicciones del modelo. Permite entender cómo cada característica afecta a las decisiones y permite justificar los resultados clínicos.

- **Gráfico de dispersión:**

- **Eje horizontal:** Valores SHAP. Impacto de cada característica en la predicción.
- **Colores:** Valores de las características (azul: valores bajos, rojo: valores altos).
- **Densidad:** Combinaciones frecuentes de datos.

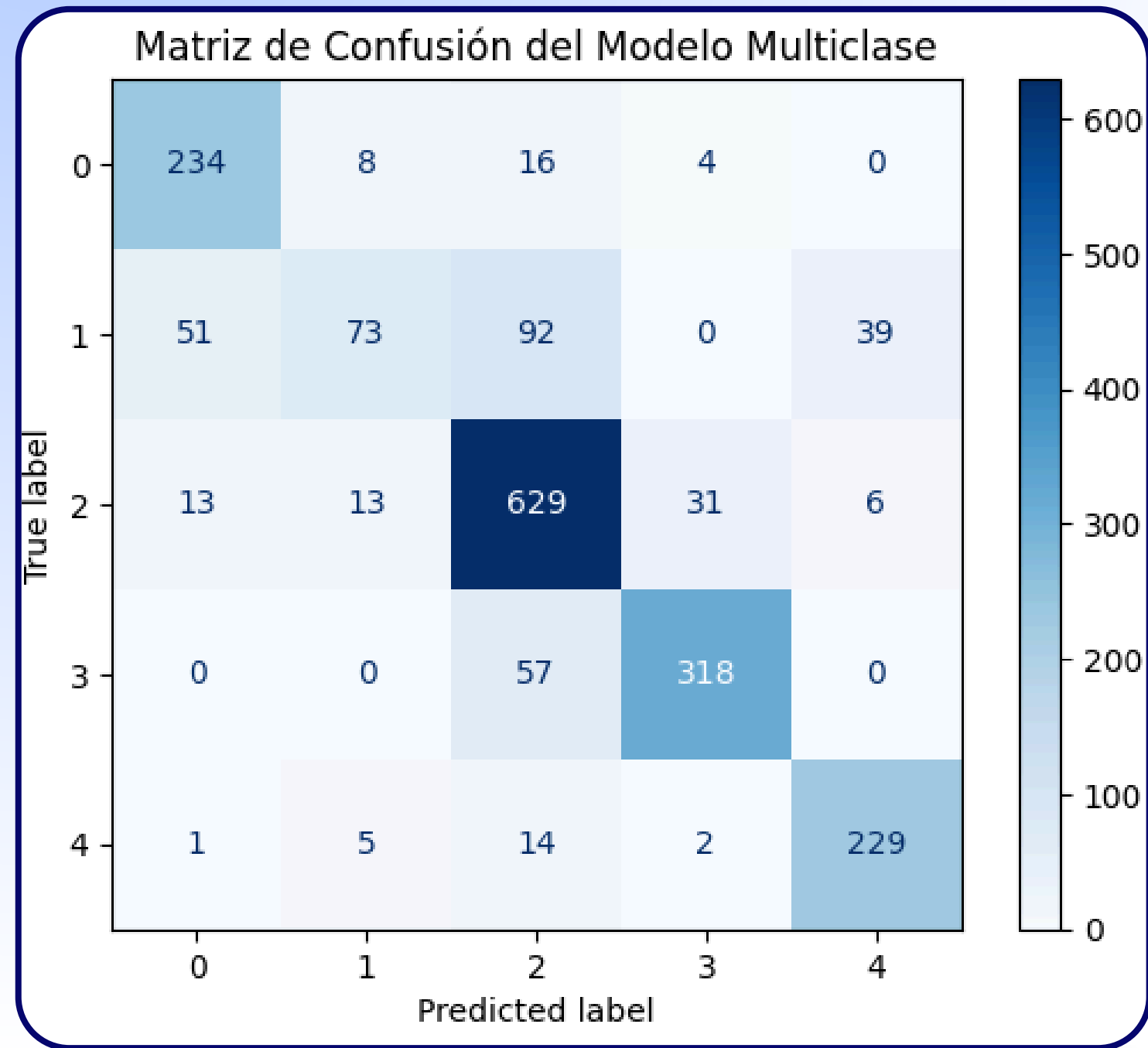
- **Análisis de dispersión:**

- **Mayor dispersión horizontal:** **eeg\_abspow\_c7min\_norm** y **eeg\_perm\_p2min\_norm**. Alta interacción con variables, con gran impacto en las predicciones.
- **Menor dispersión horizontal:** **eeg\_alpha**. Poca interacción con variables, con menor impacto en las predicciones.

- **Densidad y valores de características:**

- **Colores rosas:** Valores positivos de SHAP, es decir, valores altos en cuanto a la positiva contribución en la predicción.
- **eeg\_abspow\_p2min\_norm:** Altas concentraciones (rojo), indicando interacciones frecuentes, con gran impacto en la predicción.

# Matriz de Confusión



- **Aplicación práctica:**

- Herramienta visual para analizar el rendimiento del modelo en transiciones complejas entre fases del sueño.
- Identificar clases que requieren mejoras en las características.

- **Objetivo:** Evaluar el desempeño del modelo al mostrar las predicciones correctas e incorrectas para cada clase.

- **Ejes del gráfico:**

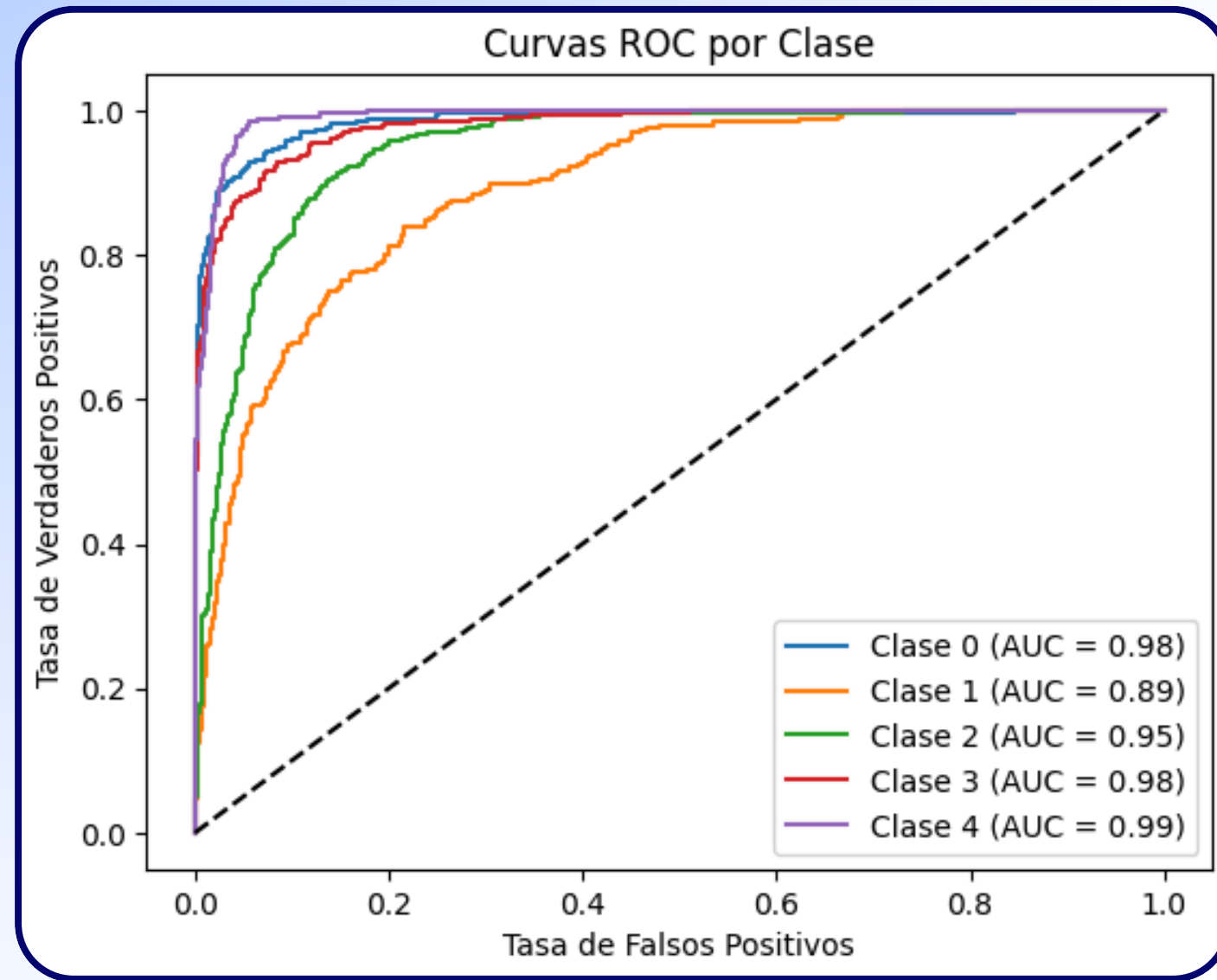
- **Eje vertical:** Etiquetas verdaderas.
- **Eje horizontal:** Etiquetas predichas por el modelo.
- Diagonal principal: Predicciones correctas.
- Fuera de diagonal: Errores de clasificación entre clases.

- **Interpretación de resultados:**

- **Clase 1:** Mayor dispersión de errores (51 clasificadas como clase 0 y 92 como clase 2).
- **Clase 2:** Mejor desempeño con errores mínimos (629 correctamente clasificadas).
- **Clase 3:** 318 correctamente clasificadas, pero 57 clasificadas como clase 2.
- **Clases 0 y 4:** Menor tasa de error.
- Las clases con más datos tienen mejor desempeño, mientras que las menos representadas presentan mayor confusión.



# Curvas ROC por clase



- **Observaciones:**

- **Clases 0, 3 y 4 (AUC>0.98):** Alta precisión para distinguirlas.
- **Clase 1:** Problemática. Dificultad para separarla, por lo que necesita ajuste.
- Modelo con un desempeño robusto en la mayoría de las clases. Pero la clase 1 requiere de una especial atención para mejorar la consistencia y precisión del modelo.

- **Objetivo:** Verificar el desempeño del modelo entre clases y ayudar a justificar ajustes en los umbrales de decisión, especialmente en fases difíciles de predecir.
- **Ejes del gráfico:**
  - **Eje horizontal (Tasa de Falsos Positivos):** Porcentaje de negativos clasificados como positivos.
  - **Eje vertical (Tasa de Verdaderos positivos):** Porcentaje de positivos correctamente clasificados.
  - **Diagonal:** Clasificador aleatorio con AUC=0.5. Cuanto más alejada esté la curva de esta línea, mejor desempeño del modelo.
- **Resultados por clase:**
  - **Clase 0 (AUC=0.98):** Excelente precisión. Alta sensibilidad y especificidad.
  - **Clase 1 (AUC=0.89):** Menor desempeño. Dificultar para diferenciar esta clase de otras.
  - **Clase 2 (AUC=0.95):** Muy buen rendimiento. Separación clara de esta clase.
  - **Clase 3 (AUC=0.98):** Excelente desempeño. Similar a la clase 0.
  - **Clase 4 (AUC=0.99):** Mejor rendimiento. Distinción de clase casi perfecta.

# Conclusión

Este proyecto ha desarrollado un modelo eficaz que representa un avance significativo para la clasificación automática de las fases del sueño, ofreciendo una herramienta precisa y clínicamente interpretable.

Aunque los resultados son prometedores, se recomienda realizar ajustes adicionales para mejorar el rendimiento en clases complejas y validar el modelo en datos externos para garantizar su aplicabilidad en diversos contextos clínicos.



# **Ronda de Preguntas**