UNIVERSITAT POLITÈCNICA DE VALÈNCIA

ESCUELA TÉCNICA SUPERIOR INGENIEROS INDUSTRIALES VALENCIA

TRABAJO FIN DE GRADO EN INGENIERÍA BIOMÉDICA

CLASIFICACIÓN DE TRASTORNOS DEL SUEÑO A PARTIR DEL ANÁLISIS DE SEÑALES CEREBRALES MEDIANTE WAVELETS Y TÉCNICAS ESTADÍSTICAS DE ANÁLISIS MULTIVARIANTE.

AUTORA: Alba González Cebrián

TUTOR: Jesús Andrés Picó i Marco

COTUTORES: Alberto José Ferrer Riquelme
José Manuel Prats Montalbán

Curso Académico: 2015-16

AGRADECIMIENTOS

Este Trabajo Fin de Grado está dedicado a mi abuelo Rafael.

Me gustaría aprovechar la ocasión para agradecer la colaboración del Grupo de Ingeniería Estadística Multivariante de la UPV, por la gran ayuda que me han prestado al realizar el proyecto.

Gracias a mi familia por el apoyo incondicional en mis aventuras, a mis amigos por seguir guardándome un sitio entre ellos, a los profesores que me han ayudado y también, gracias a cierta persona que alimenta mi ilusión cada día.

ÍNDICE DEL TFG

RESUMEN

El sueño desempeña un papel crítico en la salud humana. Las alteraciones de este hábito se relacionan con patologías crónicas y afecciones a sistemas encargados de funciones básicas. Por ello, dichas alteraciones son un objetivo interesante en el sector del desarrollo de Sistemas de Ayuda al Diagnóstico y a la Decisión Clínica, enfocado hacia una Medicina basada en métodos estandarizados, reproducibles y por tanto, robustos. El reconocimiento de patrones indicadores de la actividad cerebral durante el sueño, es clave para el diagnóstico, encontrándose dicha actividad registrada en el Electroencefalograma (EEG).

En el presente trabajo se ha realizado la caracterización y clasificación de trastornos del sueño a partir de los EEGs registrados en Polisomnografías. Mediante la Transformada *Wavelet* Discreta, se han obtenido series de coeficientes referentes a distintas bandas de frecuencia con información sobre el registro. Posteriormente se ha realizado un Análisis de Componentes Principales con los parámetros calculados sobre las series de coeficientes. Finalmente se ha obtenido una matriz de observaciones formada por tantos vectores fila de características, como individuos. Esta ha sido empleada en modelos K-Vecinos Más Próximos y Análisis Discriminante basado en Mínimos Cuadrados Parciales, para el análisis y estudio de las diversas patologías.

El desarrollo de este tipo de métodos, incluyendo análisis tiempo-frecuencia sin enventanado junto con técnicas de Análisis Multivariante, constituyen una estrategia interesante para la compresión de grandes volúmenes de datos sin pérdida de información relevante. Este tipo de herramientas nace de la voluntad tanto de mejorar la gestión de estas patologías, como de esclarecer el papel onírico de uno de los órganos con mayor importancia e intriga: el cerebro.

Palabras Clave: Sueño, Electroencefalograma, Transformada *Wavelet*, Análisis Multivariante.

RESUM

El somni posseeix un paper crític en la salut humana. Les alteracions d'aquest hàbit es relacionen amb patologies cròniques i d'altres afeccions a sistemes encarregats de tasques bàsiques. Per això, aquestes alteracions són un objectiu interessant en el sector del desenvolupament de Sistemes d'Ajuda al Diagnòstic i a la Decisió Clínica, enfocats cap a una Medicina basada en mètodes estandarditzats, reproduïbles i per tant, robusts. El reconeixement de patrons indicadors de l'activitat cerebral durant el somni, és clau per al diagnòstic, trobant-se aquesta activitat registrada a l'Electroencefalograma (EEG).

En el present treball s'ha realitzat la caracterització i classificació de desordres del somni partint de dades dels EEGs obtinguts en Polisomnografies. Mitjançant la Transformada *Wavelet* Discreta, s'han obtingut sèries de coeficients referents a diferents intervals de freqüència amb informació del registre. Posteriorment s'ha realitzat un Anàlisi de Components Principals amb el paràmetres calculats sobre les sèries de coeficients. Finalment s'ha obtingut una matriu d'observacions formada per tants vectors fila de característiques, com individus. Aquesta s'ha emprat en models K-Veïns Més Pròxims i Anàlisi Discriminant basat en Mínims Quadrats Parcials, per tal d'analitzar i estudiar les diverses patologies.

El desenvolupament d'aquest tipus de mètodes, incloent anàlisi temps-freqüència sense finestra junt amb tècniques d'Anàlisi Multivariant, constitueixen una estratègia interessant per a la compressió de grans volums de dades sense perdre informació rellevant. Aquesta classe de ferramentes naix de la voluntat tant de millorar la gestió d'aquestes patologies, així com d'esclarir el paper oníric d'un dels òrgans amb més importància i intriga: el cervell.

Paraules clau: Somni, Electroencefalograma, Transformada *Wavelet*, Anàlisi Multivariant.

ABSTRACT

Sleeping plays a main role in human health. Alterations of this routine are related to chronic diseases and affections in systems which are in charge of basic tasks. Thus, these disorders are an interesting topic in terms of developing Clinical Decision Support Systems, which are focused on achieving standardized, reproducible and therefore robust medical methodologies. Recognising patterns which identify brain activity while sleeping, is key for diagnosis of sleep disorders, being that activity recorded in the Electroencephalography (EEG).

The goal of this project is the characterization and classification of sleeping disorders using EEG from Polysomnography records. Using the Discrete Wavelet Transform, series of coefficients with information about different frequency bands of the EEG signal were obtained. Afterwards a Principal Component Analysis was performed with the parameters calculated about Wavelet Transform coefficients. Finally, an observation matrix was obtained, having as many row vectors as individuals. This matrix was imputed in the generation of K-Nearest Neighbour models and Partial Least Square – Discriminant Analysis, for the analysis and study of various illnesses.

The development of this kind of methods, including time-frequency analysis avoiding the use of a window along with Multivariate Analysis techniques, lead in interesting strategies for large data compression without losing relevant information. Tools with this philosophy are raised by the aim of improving the management of diseases, as well as because they may be an opportunity in order to clarify the oneiric role of an important and intriguing system: the brain.

Key words: Sleep, Electroencephalography, Wavelet Transform, Multivariate Analysis.

Clasificación de trastornos del sueño a partir del análisis de señales cerebrales mediante *wavelets* y técnicas estadísticas de análisis multivariante

Autora Alba González Cebrián

Tutor Jesús Andrés Picó i Marco

Co-tutores Alberto José Ferrer Riquelme, José Manuel Prats Montalbán

Fecha 06/07/2016

Titulación Grado en Ingeniería Biomédica

ESCOLA TÈCNICA SUPERIOR ENGINYERS INDUSTRIALS VALÈNCIA

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

ÍNDICE DE LA MEMORIA

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE ECUACIONES

1. INTRODUCCIÓN

1.1. ESTUDIO Y DIAGNÓSTICO DE TRASTORNOS DEL SUEÑO.

El sueño es una rutina que ha despertado intriga e interés históricos. Inicialmente relacionado con el descanso, el sueño ha pasado de estado de bajo consumo energético, a ser considerado hoy en día como un estado cíclico con momentos de actividad comparables a la realizada durante la vigilia. Pese a que las necesidades de descanso poseen una alta variabilidad interpersonal e intrapersonal, un hábito de sueño que incumpla el requisito diario mínimo situado entre cuatro y cinco horas, merma funciones básicas como la secreción hormonal, la regulación de la temperatura corporal y la transmisión nerviosa.

Las patologías o desórdenes del sueño, pueden clasificarse fundamentalmente en dos grupos mayoritarios:

- Parasomnias. Corresponden a comportamientos atípicos durante el sueño, como terrores nocturnos o sonambulismo durante la fase REM.
- Disomnias. Hacen referencia a afecciones en la cantidad, horario o calidad del sueño, incluyendo insomnio, narcolepsia, apnea del sueño o desordenes de los ritmos circadianos.

La determinación de una patología concreta cuando hay indicios y sospechas de anomalías en el hábito de sueño, pasa por la exploración del paciente, siendo común en la mayoría de los casos el registro de la actividad corporal durante el sueño, llamado Polisomnografía (PSG). Ésta constituye la principal fuente de información para el estudio de este hábito, pese a que no siempre figura como la práctica clínica establecida para el diagnóstico de ciertas patologías. La naturaleza del trastorno que se desee estudiar, justifica el empleo de unas u otras técnicas para la evaluación del sueño [1]. Aun así, las PSGs son recomendadas para la mayoría de estudios, puesto que incluyen una variedad de registros que permiten monitorizar simultáneamente distintos tipos de actividad a lo largo de toda una noche de sueño. Entre dichos registros, es habitual encontrar Electrocardiogramas (ECG), Electromiografías (EMG) o Electroencefalogramas (EEG).

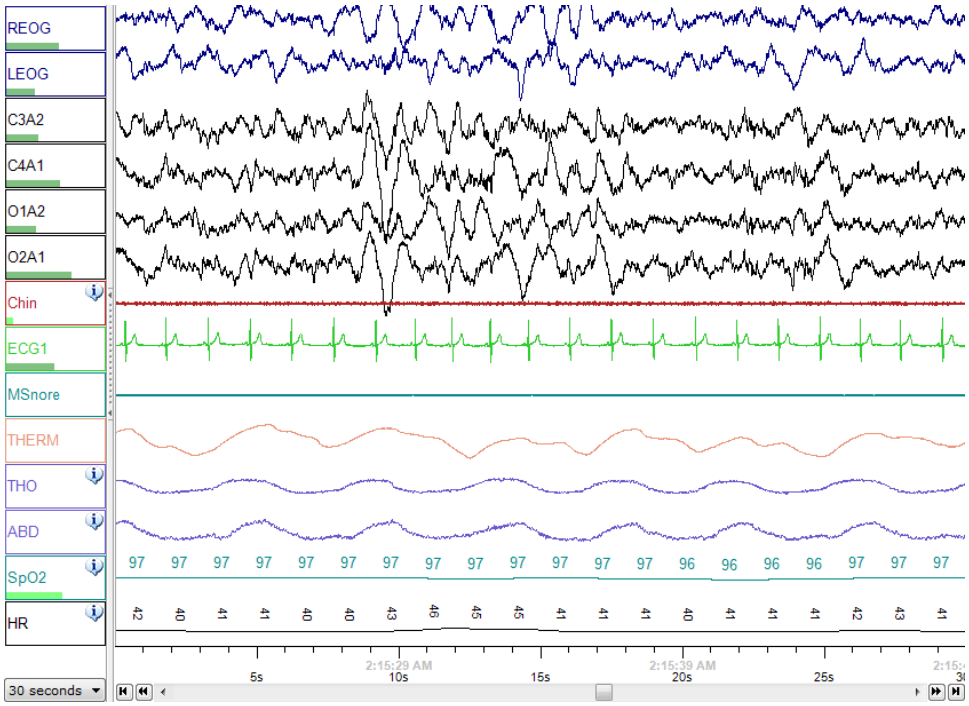

Figura 1. Ejemplo de Polisomnografía de 30 segundos con Electrooculograma (REOG, LEOG), Electroencefalograma (C3A2, C4A1, O1A2, O2A1), Electrogardiograma (ECG), registro del ronquido (MSnore), de la temperatura (THERM), seguimiento de la presión Torácica y Abdominal (THO, ABD), de la saturación de Oxígeno en sangre (SpO2) y de la Frecuencia cardíaca (HR).

El EEG contiene información sobre la actividad encefálica, íntimamente relacionada con una función normal o patológica del sueño. Por esta razón, es uno de los registros obligatoriamente incluidos en la PSG. El sistema 10-20 es la recomendación de la Federación Internacional de Electroencefalografía para la captación del EEG, siendo el sistema más usado en la actualidad. Especifica la disposición de 21 electrodos alrededor de la superficie encefálica, en contacto con la piel del sujeto en cuestión (Figura 2). Para lograr una distribución que cubra varias zonas del córtex, la superficie encefálica se divide en 5 o 6 puntos equiespaciados un 10 o 20% de la distancia entre referencias anatómicas en lados opuestos.

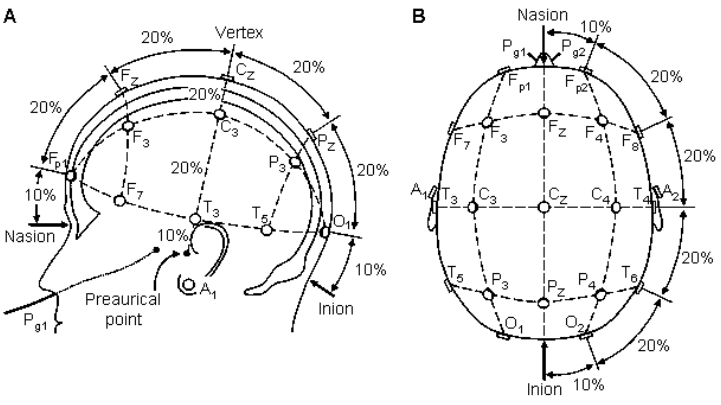

Figura 2. Sistema internacional 10-20 desde vista lateral izquierda (A) y superior (B).

Las mayúsculas de cada electrodo determinarán el lóbulo en el que esté situado ("F" frontal, "C" central, "T" temporal, "P" parietal y "O" occipital), concretándose puntos más específicos por los subíndices ("p" punto polar y "z" la línea media, del inglés *zero*). Los electrodos impares refieren al lado izquierdo y los pares al lateral derecho. La masa empleada como referencia, conectada al amplificador, son los lóbulos auriculares A_1 y A_2.

Una vez la señal ya ha sido registrada, múltiples parámetros pueden cuantificar el comportamiento de una señal, siendo amplitud y frecuencia los dos más básicos. El EEG es una señal cuyos rangos fisiológicos oscilan entre los 10 µV y los 100 µV de amplitud cuando se capta con electrodos de superficie, y los 0.1 y 30 Hz de frecuencia [2]. La variabilidad de la amplitud por los factores que le afectan (interferencias, atenuación de la señal), hace que sea la frecuencia el parámetro escogido para representar el funcionamiento cerebral. En general se pueden distinguir 5 ondas distintas comprendidas en un ancho de banda desde los 0 Hz hasta algunas que superen los 30 Hz (Tabla 1).

Tabla 1. Clasificación de señal EEG en base a la frecuencia en ondas α, β, θ, δ o γ.

Nombre de la onda	Ancho de banda	Función/estado del sueño asociado
Delta (δ)	0.5 – 2 Hz	Sueño profundo (fases 3 y 4), sobretodo en región frontal
Theta (θ)	3 – 7 Hz	Infantes, adultos con estrés emocional, frecuencia más abundante durante el sueño
Alfa (α)	8 – 13 Hz	Relajación y fase 1 del sueño, producida en la región occipital.
Beta (β)	14 – 30 Hz	Activación de zonas de la corteza, fase REM
Gamma (γ)	> 30 Hz	Procesamiento activo

Puesto que el córtex cerebral se divide en distintas áreas asociadas a diferentes funciones, equivale a varias fuentes que coexisten temporalmente, siendo el EEG una señal variante en el tiempo y el espacio. La procedencia de señales desde distintas localizaciones, requiere tener en cuenta qué electrodos se deben escoger para un estudio de la actividad cerebral. Por otro lado, esta superposición de distintas ondas, suele dar como resultado señales con frecuencias altas (activación de varias zonas) pero amplitudes pequeñas debido a la suma de distintas fuentes de señal. Sin embargo, cuando hay más zonas dañadas o inactivas, la contribución a la señal de un menor número de zonas en funcionamiento se refleja en un EEG con menores frecuencias pero mayores amplitudes. Así pues, tramos donde la señal parece completamente errática, suelen ser un mejor indicador en términos de salud, que tramos con un patrón más definido, típicos de fenómenos patológicos como la epilepsia (Figura 3).

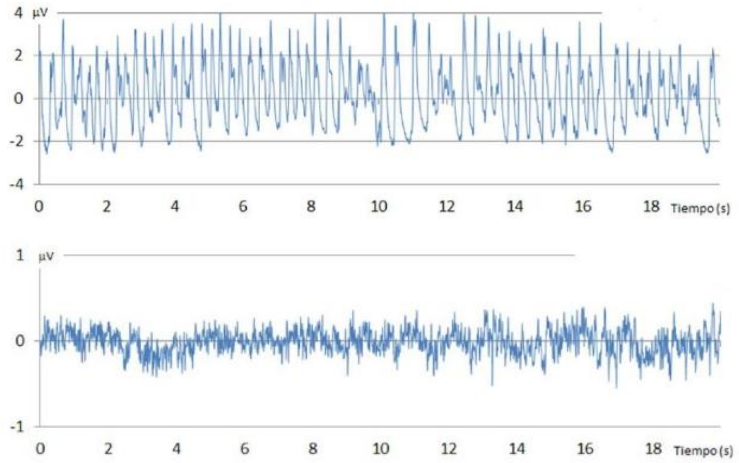

Figura 3. Imágenes de registros EEG durante 20 segundos de un sujeto sano (inferior) y un sujeto con episodio epiléptico (superior). Puede apreciarse la apariencia más errática y aleatoria en el sujeto sano, presentando la señal menores amplitudes en comparación con la imagen superior, en la que el periodo de la señal es más claro y las amplitudes mayores. [3]

En base a las distintas ondas (Tabla 1), se construyen unidades mayores de actividad cerebral que son las empleadas en la clasificación de los estadios del sueño La clasificación inicial en cuatro estados elaborada por A.Rechtschaffen y A.Kales en 1968, fue actualizada en 2007 por la American Academy of Sleep Medicine (AASM). Esta nueva clasificación (Tabla 2) consta de tres fases iniciales no-*REM* (NREM) y una *REM* (del inglés *Rapid Eye Movement*). Las primeras tres fases reciben el nombre de su posición en el ciclo (N1, N2 y N3). Los acontecimientos fisiológicos difieren en cada una de las fases, así como la duración, estimándose que el 50% del sueño pertenece a la fase N2, el 20% a la fase *REM* y el 30 % restante se distribuye entre las demás fases [4].

Los complejos K y *spindles* son patrones de ondas características de la fase N2 del sueño. Los *spindles* poseen una frecuencia entre 12 y 14 Hz, con una duración superior a los 0.5 – 3 segundos. Los complejos K son ondas lentas, con perfil afilado formado por una primera deflexión negativa y seguidas de una positiva, cuya duración debe superar el medio segundo. Ambos comportamientos se producen de forma mayoritaria en la región vértex-central.

Tabla 2. Identificación de fases del sueño en los registros EEG. Actualización de las fases de R & K (1986) por la AASM en 2007.

Fase del sueño	Descripción
N1	Más de media *epoch* contiene ondas theta. Presencia de ondas con morfología de vértice.
N2	Actividad theta. Complejos K y *spindles* ocasionales.
N3	Más del 20% de la *epoch* con ondas delta. Amplitud ≥ 75 µV.
REM	Frecuencias mixtas, amplitud baja, similar al estado de vigilia. Posibles ondas con morfología de diente de sierra.

Así pues, el comportamiento cíclico del sueño se refleja como variaciones en el tipo de actividad cerebral desempeñada, repitiéndose con una cadencia de 90 minutos aproximadamente, las fases descritas en la Tabla 2. La metodología estándar para el análisis de las fases del sueño, pasa por la obtención del Hipnograma (Figura 4). El procedimiento establecido es la división del electroencefalograma en segmentos de unos 30 segundos de duración, llamados *epochs*. El tipo de onda con mayor contribución a la energía de la señal en este tramo, se asigna como la actividad cerebral dominante durante esos 30 segundos del registro.

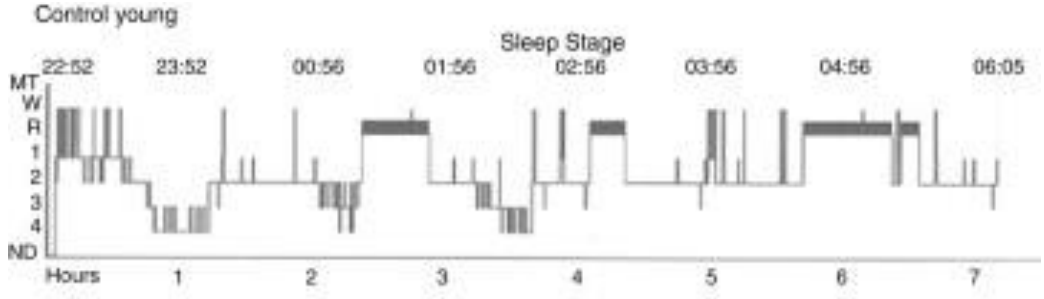

Figura 4. Hipnograma con la evolución temporal de las fases del sueño. En el eje vertical las distintas fases y en el horizontal las horas. Nótese la clasificación empleando las normas de R & K con cuatro fases No Rem (1, 2, 3, 4) y una fase REM (R), siendo W el estado de vigilia.

Tras la obtención de la evolución de las fases del sueño, el sueño de cada sujeto es cuantificado mediante el cálculo de parámetros que permitirán posteriormente, establecer una comparación entre los valores obtenidos y los esperados para individuos similares al caso de estudio particular. Dichas medidas, la mayoría propuestas cuando las normas de R & K estaban todavía vigentes, han sido recogidas y adaptadas por la AASM, encontrando:

- Tiempo total de sueño. Equivale a la suma de las duraciones en minutos de las fases N1, N2, N3 y REM.
- Tiempo en cama o tiempo total del registro.
- Latencia del sueño. Tiempo transcurrido desde que se apaga la luz en la sala de realización de la PSG, hasta que comienza la primera fase del sueño.
- Latencia del estadio REM. Tiempo transcurrido desde que empieza la primera fase del sueño, hasta la primera *epoch* clasificada como REM.
- Tiempo de inicio de vigilia tras el sueño. Expresado en minutos, es el tiempo acumulado de estados *W* (despierto) que tienen lugar desde que se inicia el sueño, hasta el despertar final del individuo.
- Eficiencia del sueño. Expresada en porcentaje, se calcula como el cociente entre el Tiempo total de sueño y el Tiempo total de registro.
- Porcentaje de tiempo en cada fase. Calculado como el cociente expresado en porcentaje, entre el tiempo calculado en total para cada fase, y el tiempo total de sueño.

Así pues, los parámetros comentados, una vez son obtenidos, no tienen valor diagnóstico por sí solos. Su valía reside en las diferencias que el profesional médico observe respecto a valores de referencia que representen a individuos sanos con un perfil que concuerde con el del sujeto estudiado.

1.2. IMPORTANCIA E IMPACTO SOCIAL DEL ESTUDIO DEL SUEÑO.

La relevancia de la metodología descrita reside en su aplicación como método diagnóstico de trastornos del sueño, cuya prevalencia se estima entre el 35 y 40% de la población adulta en USA [7], con 40 millones de pacientes crónicos y otros 20 anuales [4]. De acuerdo con la Sociedad Española del Sueño (SES), un tercio de la españoles padece algún trastorno del sueño, y las cifras de afectados a nivel mundial se estiman cercanas al 50% [6]. Además, un indebido hábito de sueño puede tener consecuencias críticas en el crecimiento a lo largo de la infancia y, aunque no se ha demostrado una clara relación entre falta de sueño y fallecimiento, sí hay evidencia clínica sobre una mayor incidencia de patologías severas y crónicas como enfermedades cardiovasculares o psiquiátricas [7]. Así pues, un correcto estudio del sueño conlleva una mejor gestión del impacto que generan en los sistemas sanitarios.

Según la *American Sleep Association*, la huella económica impresa por los trastornos del sueño alcanza cifras estimadas de 16 billones de dólares americanos en cuanto a costes médicos directos. La sospecha es que los indirectos son todavía mayores, asociados entre otros a pérdida de productividad, relación directa con patologías crónicas o por ser un claro factor de riesgo en la ocurrencia de accidentes de tráfico. La causa de su encarecimiento apunta en dirección a las fases previas al tratamiento, que no acarrea grandes dificultades organizativas, siendo la fase de diagnóstico de la enfermedad lo que mayor cantidad de recursos del sistema sanitario consume.

Con el fin de aligerar un estudio largo y múltiple que requiere de la atención médica durante toda una noche, como es el caso de la PSG, se han desarrollado equipos que abordan esta prueba mediante telemonitorización del paciente. Existen distintos tipos de Polisomnografías, pudiendo realizarse en un laboratorio del centro de salud (Tipo I), o mediante un dispositivo portátil instalado en la casa del propio paciente (Tipos II, III y IV). La variabilidad dentro de cada tipo viene dada por las condiciones del entorno donde se realiza la medida, pero también por la mayor o menor adaptación del paciente a la prueba y por el profesional que realice la evaluación del registro. Actualmente, ningún caso de PSG portátil es recomendado de forma única para el diagnóstico, siendo necesarios los estudios Tipo I en aquellos casos en los que una PSG sea la prueba diagnóstica establecida.

Un diagnóstico correcto es necesario para la gestión exitosa de estas patologías, pero la interpretación manual de los registros genera una gran variabilidad en los resultados obtenidos. Los esfuerzos por parte de organizaciones como la AASM en la divulgación de protocolos y metodologías estándar son un motor para el consenso y mejora en la calidad de los estudios del sueño. Sin embargo, la falta de una Normativa o Protocolo que unifique el procedimiento establecido en cada centro, basándose en métodos cuantitativos y objetivos, repercute negativamente en aspectos fundamentales, como la reproducibilidad de los diagnósticos establecidos en base al estudio manual de los registros. Las diferencias generadas por la falta de un método estándar son, por tanto, un tema con el que el estudio del sueño ha de lidiar todavía a día de hoy [8].

Aunque sí hay un consenso general en que las PSG de Tipo I son el estándar en cuanto a realización de esta práctica clínica, esta afirmación se contradice con el ánimo creciente de trasladar el cuidado y supervisión de los pacientes a sus domicilios. Pese a la menor fiabilidad de los registros obtenidos fuera de un entorno hospitalario, la telomonitorización está presente y es considerada una línea futura clara, de acuerdo con valoraciones realizadas por la AASM respecto al uso de la Telemedicina en la gestión de trastornos de sueño [9]. Esta integración entre tecnología y salud, ya ha tomado posiciones más avanzadas en otros campos, como el del Electrocardiograma o la Pulsioximetría. Ambas prácticas tienen ya variantes considerablemente fiables y que son realizadas mediante dispositivos portátiles, como es el caso del Holter, que registra la actividad cardíaca permitiendo hacer vida normal a un paciente durante unas 24 horas.

Yendo un paso más allá en la optimización del seguimiento de enfermedades, es cada vez más común hablar de métodos faciliten la emisión del diagnóstico, cimentándose una de las variantes de los Sistemas de Ayuda a la Decisión. Así pues, el reconocimiento de patrones y el desarrollo de algoritmos para la clasificación de éstos es un tema recurrente en la bibliografía sobre los llamados Sistemas de Ayuda al Diagnóstico. Se cree posible que el análisis informatizado de estas señales permita un examen más exhaustivo de la información que contienen, acrecentando en un futuro el valor de estas exploraciones [10].

Por ahora, la ausencia de un criterio basado en un análisis numérico y riguroso de la señal requiere que la PSG esté constituida por registros que representen actividades bioeléctricas diferentes, pues pretender captar cambios más sutiles en el comportamiento de una señal a simple vista es muchas veces una tarea imposible. La necesidad de examinar otros registros como el Electroculograma o el Electromiograma, aumenta el peso del trabajo manual y cualitativo a la hora de diagnosticar, volviendo a obtener resultados variables y poco reproducibles.

No obstante, la interpretación de señales encontrándose patrones característicos mediante el uso de herramientas matemáticas y estadísticas es un tema con el que el ámbito ingenieril se enfrentó en su momento logrando establecer todo un campo de Análisis de Señales. Estudios realizados a lo largo de las últimas décadas han logrado aplicar de forma exitosa técnicas ingenieriles al estudio de la actividad biológica, aunque en el caso del EEG todavía a nivel pre-clínico. Las perspectivas futuras de este campo son alentadoras, no solo ofreciendo un colchón de seguridad en el diagnóstico por su precisión y resultados cuantificables, sino también haciendo evolucionar al estudio y comprensión de la fisiología humana.

2. OBJETIVOS

El objetivo del presente trabajo es el análisis y modelado de señales cerebrales registradas en Electroencefalograma, para su caracterización y establecimiento de modelos de compresión y clasificación basados en la implementación de la Transformada *Wavelet* y técnicas estadísticas de análisis multivariante.

Los distintos módulos de los que se compone el trabajo y que constituyen los hitos hacia el objetivo final, son:

- Transformación del espacio temporal al espacio tiempo-frecuencia para el análisis de la señal EEG mediante uso de la Transformada *Wavelet*.
- Desarrollo de un modelo de compresión basado en Análisis de Componentes Principales (PCA) para reducir la dimensionalidad de los datos obteniendo una estructura del tipo *individuos x variables.*
- Implementación y comprobación de los resultados obtenidos con diferentes modelos estadísticos multivariantes de clasificación en base a las características extraídas sobre la señal EEG.

3. MATERIALES Y MÉTODOS

3.1. ESTRUCTURACIÓN DE LOS DATOS.

3.1.1. Obtención de los datos, cribado y selección.

El elemento de trabajo que contiene toda la información, alrededor del cual gira todo el trabajo, son los datos. Desde el inicio hasta llegar a la extracción de variables relevantes para una clasificación, el conjunto de datos pasa por varias etapas y procesos cuya finalidad puede ser la compresión de la información, el procesado de las variables para hacerlas comparables entre sí, o la transformación de los datos a dominios de otra naturaleza. El hilo de operaciones realizadas sobre los datos en crudo, hasta llegar a la implementación de los modelos de compresión, constituye la etapa conocida como preprocesado de los datos.

El punto de partida es la Base de Datos (BBDD) desde la cual es descargada la información. Para este proyecto, la naturaleza de la señal a analizar restringía la cantidad de bases de datos a aquellas que contuviesen registros sobre señales biomédicas. Por ser una BBDD con acceso libre a los repositorios y contar con archivos en un formato que puede ser fácilmente trabado en Matlab, Physionet.org se escogió como repositorio.

Los archivos empleados provienen de la base de datos de Patrones Cíclicos Alternantes (*Cyclic Alternating Patterns*), elaborada por el Centro de Trastornos del Sueño del *Maggiore Ospedale* de Parma, Italia [11]. Este banco de registros, contiene en total 108 EEGs de Polisomnografías realizadas a 16 pacientes sanos y 92 con distintas patologías del sueño. Las patologías son: bruxismo, epilepsia nocturna del lóbulo frontal, narcolepsia, movimientos periódicos de piernas, trastorno de la fase REM, trastornos respiratorios durante el sueño (apnea). El formato de los registros es el estándar europeo .edf (*European Data Format*).

Por tratarse de trastornos en la conducta del sueño, el examen realizado consiste en una Polisomnografía, las patologías ditadas son exploradas realizando una Polisomnografía. La extensa duración de estos registros, tomados con una frecuencia de muestreo de 512 Hz, genera matrices iniciales cuyas dimensiones temporales superan el orden de 10^7 columnas. Incluso suponiendo una cantidad de horas baja, en comparación con las 8-9 horas normales, por tratarse de pacientes con trastornos del sueño:

$$6\ horas\ de\ sueño\ \times 3600\frac{segundos}{hora} \times 512\frac{muestras}{segundo} = 11059200\frac{muestras}{paciente \cdot electrodo}$$

La gran cantidad de muestras a analizar por cada electrodo para cada paciente, ha supuesto un reto a la hora de cargar los datos para poder trabajar con ellos en el espacio de trabajo. El problema ha sido abordado mediante el uso del objeto *matlab.io.MatFile* [13], un objeto Matlab que permite trabajar con grandes volúmenes de datos (ver Anejo 1). Pese a que no soporta comandos básicos como el indexado, o funciones que lo empleen como variable de entrada, la ventaja que ofrece esta clase es la posibilidad de aunar toda la información referente a cada paciente en una sola estructura. Esta estructura cuenta con tantas variables como individuos. Si un comando llama a uno de los pacientes almacenados en el objeto, o a parte de la información de uno, en la memoria del programa solo se cargará la parte requerida para el comando ejecutado, sin necesidad de tener toda la estructura cargada en la memoria del programa.

La captación de la actividad cerebral mediante electrodos de superficie distribuidos en distintas localizaciones de la cabeza del sujeto conlleva la variación espacial de la señal captada, viéndose distinta según el electrodo del que provenga el registro analizado. Por tanto, para realizar una comparación entre registros, el primer punto fue asegurar que las comparaciones se establecían entre información procedente del mismo conjunto de electrodos.

El segundo criterio a tener en cuenta fue la frecuencia de muestreo con la que el registro fue tomado. Este parámetro del registro condiciona el uso de la Transformada *Wavelet* para el paso del dominio temporal al dominio de tiempo-frecuencia. Por la naturaleza del Análisis Multiresolución (AMR), cada nivel de compresión obtenido tras la transformación, contiene la información referente a una banda de frecuencias.

Tras descargar los datos de Physionet.org, se obtuvo un fichero en formato .edf que contenía una estructura .mat con los valores numéricos de los registros, junto con una estructura .hdr con la información de las cabeceras. Si bien los datos numéricos son el objeto principal de trabajo, la información existente en las cabeceras fue necesaria también por varios motivos.

En primer lugar, la utilidad de las cabeceras residió en su uso para la visualización de los electrodos correspondientes a cada fila en la estructura numérica de cada paciente. Este punto es esencial en la determinación de los electrodos que se escogerán como fuentes de la señal. Para escogerlos, se exportaron los campos referentes a la distribución de electrodos en las cabeceras de cada paciente a un fichero Excel.

Tras visualizar la distribución, se encontró que no todas las filas hacían referencia al mismo electrodo, ni tampoco a la misma señal. La existencia de distintas señales se debe a la naturaleza del estudio Polisomnográfico, que cuenta además del EEG con Electroculogramas (EOG) y Electromiogramas (EMG) entre otros. En cuanto a la señal EEG, pese a que el registro fue realizado mediante el sistema 10-20, no constaba la información referente a los 21 electrodos que componen este sistema. A las diferencias en cuanto la posición de los electrodos entre los individuos, se añadió la existencia de registros no válidos, la mayoría por la falta de algunos canales, o bien por ser registros vacíos que dieron error cuando se cargó la estructura numérica en Matlab.

En segundo lugar, en las cabeceras también se puede encontrar información sobre la frecuencia de muestreo con que un registro fue tomado. La frecuencia de muestreo para la captación del EEG fue de 512 Hz. Sin embargo, algunos pacientes tenían en su cabecera correspondiente a este dato, valores inferiores o superiores. Pese a que una solución podría haber sido el remuestreo de los registros con frecuencias de muestreo diferentes, se optó por un enfoque más conservador, dado que estas diferencias se daban en casos puntuales y se desconocía hasta qué punto podría perderse información relevante de los registros. Definitivamente, el principal obstáculo y la fase más costosa fue lidiar con las inconsistencias en los electrodos de procedencia del registro EEG.

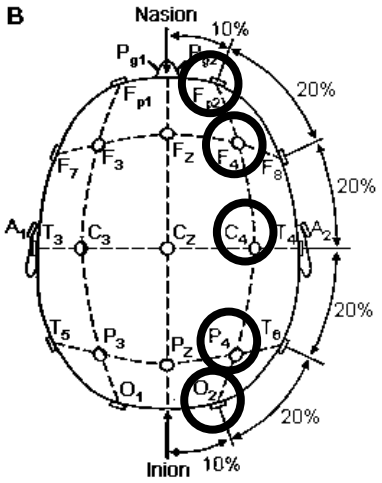

Figura 5. Distribución de los electrodos escogidos para el análisis de los registros.

Finalmente, tras este cribado de los datos, se pasó de un número de 108 pacientes con entre 15 y 21 filas por registro, a un conjunto de 80 pacientes con 4 filas por archivo, correspondientes a los 4 electrodos más comunes en la mayoría de pacientes, captando la señal EEG con una frecuencia de muestreo de 512 Hz. Estos electrodos eran, siguiendo la nomenclatura del SI 10-20, los Fp2-Fp4, F4-C4, C4-P4 y P4-O2 (Figura 5).

Leyenda
bruxismo
insomnio
normal
narcolepsia
epilepsia nocturna del lóbulo frontal
movimientos repetitivos de piernas
trastorno en la conducta del sueño REM
apnea del sueño

Motivos de eliminación de paciente:
No electrodos comunes
Frecuencia de muestreo ≠ 512 Hz
Insuficiente taamaño de su clase

Set inicial	
Paciente	Patología
1	BRUX1
2	BRUX2
3	INS1
4	INS2
5	INS3
6	INS4
7	INS5
8	INS6
9	INS7
10	INS8
11	INS9
12	N1
13	N2
14	N3
15	N4
16	N5
17	
18	
19	N8
20	N9
21	N10
22	N11
23	
24	
25	
26	

27	N16
28	NARCO1
29	NARCO2
30	NARCO3
31	NARCO4
32	NARCO5
33	NFLE1
34	NFLE2
35	NFLE3
36	NFLE4
37	NFLE5
38	NFLE6
39	NFLE7
40	NFLE8
41	NFLE9
42	
43	
44	NFLE12
45	NFLE13
46	NFLE14
47	NFLE15
48	NFLE16
49	NFLE17
50	NFLE18
51	
52	
53	NFLE21
54	NFLE22

55	
56	NFLE24
57	
58	
59	NFLE27
60	NFLE28
61	NFLE29
62	NFLE30
63	
64	NFLE32
65	
66	NFLE34
67	NFLE35
68	NFLE36
69	NFLE37
70	NFLE38
71	NFLE39
72	NFLE40
73	PLM1
74	PLM2
75	PLM3
76	PLM4
77	PLM5
78	PLM6
79	PLM7
80	PLM8
81	PLM9
82	PLM10

82	PLM10
83	RBD1
84	RBD2
85	RBD3
86	RBD4
87	RBD5
88	RBD6
89	RBD7
90	RBD8
91	RBD9
92	RBD10
93	RDB11
94	RBD12
95	RBD13
96	RBD14
97	RBD15
98	RBD16
99	RBD17
100	RBD18
101	RBD19
102	RBD20
103	RBD21
104	RBD22
105	SBD1
106	SBD2
107	SDB3
108	SDB4

Figura 6. Resumen de la inspección inicial de las cabeceras. Las leyendas de la zona superior hacen referencia a las razones por las que se eliminó ese registro (izquierda) y a las patologías en las que se clasifica cada individuo según el diagnóstico médico, adjuntado también como campo en las cabeceras (derecha). En la zona inferior, se encuentran los 108 registros con la casilla derecha del color referente al resultado tras la inspección. Los registros en buen estado, mantienen el color de la patología.

Una vez conocidos los electrodos a cargar de la estructura numérica de cada paciente, se redujo significativamente la cantidad de información a manejar. Sin embargo, trabajar con esta cantidad masiva de datos fue uno de los principales retos a superar llegado este punto, debido a la larga duración de los registros. Teniendo en cuenta la frecuencia de muestreo de 512 Hz, y la duración de una noche de sueño alrededor de 6-7 horas, se alcanza al orden de decenas de millón en la duración del registro realizado en cada uno de los electrodos.

Por otro lado, para la construcción del conjunto de datos final, los pacientes deben estar como distintos individuos, cada uno en una fila de la matriz de observaciones. Para conseguir una matriz con esta estructura, en primer lugar se igualó la dimensionalidad de los datos en cada paciente, puesto que no es posible almacenarlos directamente en una única matriz por la diferente dimensión en el número de columnas.

3.1.3. Organización de los datos.

Aunque a partir de este punto se tiene la información estructurada, esta disposición de los datos no es óptima debido a que toda la señal sería introducida en la función que calculase a continación la descomposición mediante la Transformada *Wavelet* Discreta. Por ello, se reorganizan los datos dentro de cada variable paciente, organizándose en tantas filas como tramos de 30 segundos o *epoch*s y tantas columnas como electrodos. Cada coordenada (*epoch*, electrodo) dentro de paciente *pa###*, será una matriz con la porción de registro correspondiente.

Inicialmente, se parte de un conjunto en el que cada paciente constituye una matriz de dimensiones $Electrodos \times Muestras_{EEG}$. El tipo de estructura en la que se basan los algoritmos empleados en Análisis Multivariante que serán aplicados posteriormente a los datos, es del tipo $Individuos \times Variables$ (Figura 7). En este caso, los individuos serán cada uno de los pacientes, y las variables serán características extraídas sobre estos. Por tanto, la estrategia seguida tiene como objetivo la compresión de la información referente a un paciente, partiendo de una matriz, hasta llegar a un vector de variables.

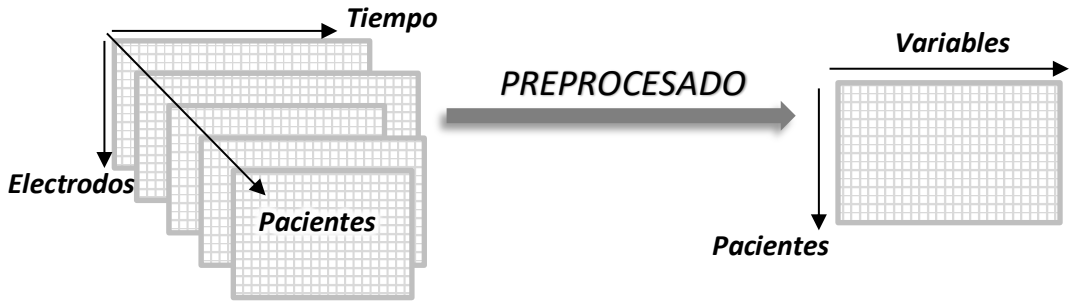

Figura 7. Descripción gráfica del objetivo perseguido con el preprocesado de los datos, partiendo de una estructura tridimensional, para llegar a una bidimensional. Puede apreciarse las diferentes duraciones de los registros en cada paciente, debido a que cada sujeto durmió un número de horas distinto al realizar el estudio.

Además de la reducción de dimensionalidad para la consecución de la estructura final, existe una necesidad de compresión de información para poder operar a nivel de software, dada la problemática explicada en el apartado *Obtención de los datos y carga en el espacio de trabajo*. Tal y como se señala en el apartado anterior, por el factor crítico con el que cuenta un escenario basado en el diagnóstico de enfermedades, se ha buscado otra estrategia que permita la compresión de datos, desde un punto de vista más conservador. Imitando la metodología seguida por el personal médico, cada registro ha sido discretizado en *epoch*s (Figura 8), que son ventanas con una longitud temporal de 30 segundos, y consituyen el estándar establecido para el análisis de señales neurológicas.

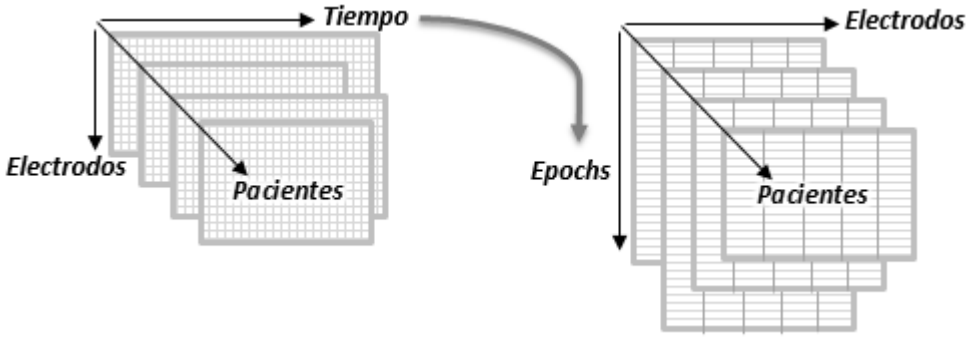

Figura 8. Representación gráfica de la primera reorganización del conjunto de datos inicial, agrupando las muestras temporales en *epoch*s. La distinta duración de los registros se verá ahora reflejada en distinto número de *epoch*s para cada paciente.

Tener los datos en esta disposición facilitó la transformación al espacio tiempo-frecuencia mediante el uso de la Transformada *Wavelet*. Una vez se obtuvo la transformación (Figura 8), fueron necesarios más pasos en la organización óptima de los datos (Figura 9).

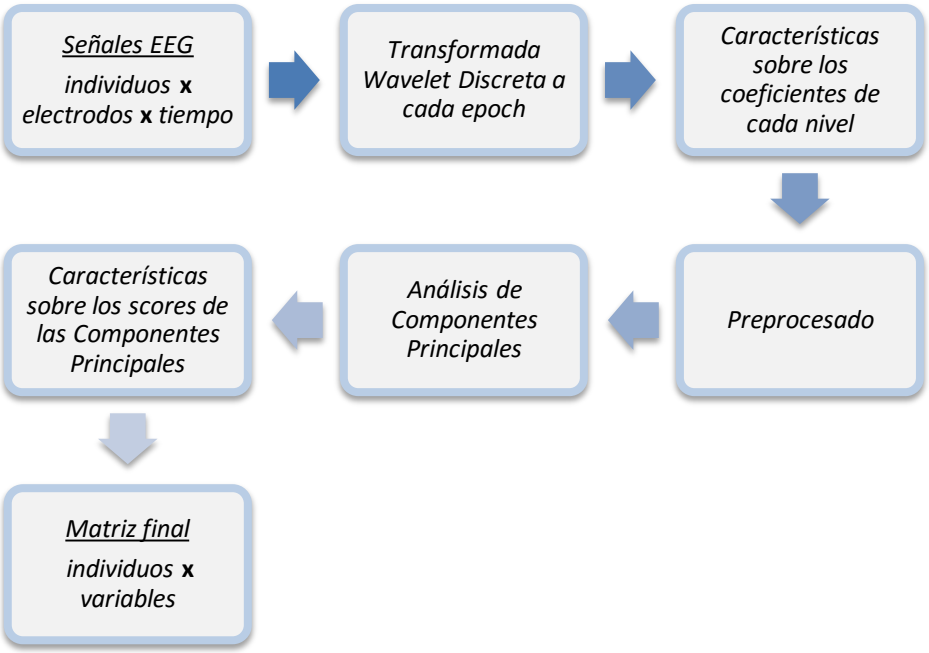

Figura 9. Sucesivas transformaciones sobre el conjunto de datos hasta llegar a la estructura final deseada.

3.2. TRANSFORMADA WAVELET.

La Transformada *Wavelet* (TW) es una herramienta matemática que, partiendo de una señal temporal, obtiene unas series de coeficientes que caracterizan el comportamiento de dicha señal en el dominio tiempo-frecuencia. Trabajar en este espacio implica tener información no solo sobre las frecuencias que componen una señal, sino sobre cuándo se producen cambios en cada una de las componentes frecuenciales. Poder localizar temporalmente cambios en el comportamiento de la señal, reflejados en variaciones de su frecuencia, es de especial interés cuando se trata del estudio de señales no estacionarias y variantes tanto en el tiempo como en el espacio, siendo este el caso del Electroencefalograma (EEG).

Por otro lado, tal y como se ha descrito en el apartado 3.1.3. Organización de los datos., la reducción de la dimensionalidad de los datos es tanto un objetivo de este trabajo como una necesidad particular debido a la extensa duración de los registros EEG. Ello requiere de estrategias potentes en la compresión de información, que no conlleven pérdida de la misma, puesto que esto podría llevar en última instancia a un diagnóstico erróneo. La Transformada *Wavelet* es un buen candidato para este fin, siendo ampliamente conocida por su capacidad de compresión.

Frente a métodos tradicionales como la Trasformada de Fourier (TF), la TW ofrece resolución variable dentro de la ventana de convolución con la señal. Esto es posible gracias al elemento que define esta transformada, llamado ondícula, o más comúnmente *wavelet*. Matemáticamente, la Transformada *Wavelet* Continua (TWC) de una señal definida en el dominio temporal $\left(s(t) \in L^2(\mathbb{R})\right)$, se define como la convolución entre la señal y la función *wavelet* $\Psi_{a,b}(t)$ (Ecuación 1).

$$W_s(a,b) = |a|^{-\frac{1}{2}} \int_{-\infty}^{\infty} s(t)\Psi^*_{a,b}(t)dt \ , con\ \Psi_{a,b}(t) = \Psi\left(\frac{t-b}{a}\right)$$

Ecuación 1. Transformada *Wavelet* Continua de una señal $\left(\mathbf{s(t)} \in \mathbf{L^2}(\mathbb{R})\right)$ con la *Wavelet* madre $\boldsymbol{\Psi}$, dilatada y trasladada dado el par de coeficientes (*a, b*).

La *wavelet* madre (Ψ) es dilatada y trasladada con el factor de $a\ (a > 0)$ y b, respectivamente. Los valores posibles vienen determinados por la expresión $a_m = a_0^m$ y $b_{m,n} = nb_0a_0^m$, donde m es la localización en frecuencia y n la localización en el dominio temporal. Si los valores de m y n pueden tomar un rango discreto de valores, se obtiene la Transformada *Wavelet* Discreta (TWD) (Ecuación 2).

$$W_s(a,b) = |2|^{-\frac{m}{2}} \int_{-\infty}^{\infty} s(t)\Psi^*\left(\frac{t-2^m n}{2^m}\right)dt$$

Ecuación 2. Transformada *Wavelet* Discreta de una señal $\left(\mathbf{s(t)} \in \mathbf{L^2}(\mathbb{R})\right)$.

Así pues, la *wavelet* madre viene siendo una función caracterizada por un comportamiento oscilatorio determinado. Puesto que hay multitud de oscilaciones posibles alrededor de la línea base, la existencia de diversas morfologías de la ondícula, da lugar distintas familias de *wavelets*, cada una con su *wavelet* madre (Figura 10). Cada familia de funciones está a su vez formada por un conjunto de *wavelets* con distinto número de momentos de desvanecimiento, que está relacionado con el orden de la función *wavelet*. A mayor orden, la *wavelet* tendrá más puntos de derivada nula, y por tanto, una morfología más compleja.

La elección de la familia de *wavelets* viene determinada por las características de la señal. Esa elección es un factor clave, puesto que el análisis de la señal devolverá distintas características. Esto es debido a que diferentes zonas de la señal serán mayor o menormente ponderadas dependiendo de la morfología de la *wavelet* madre con la que se realice la convolución, obteniéndose resultados directamente proporcionales a la convolución entre *wavelet* y señal.

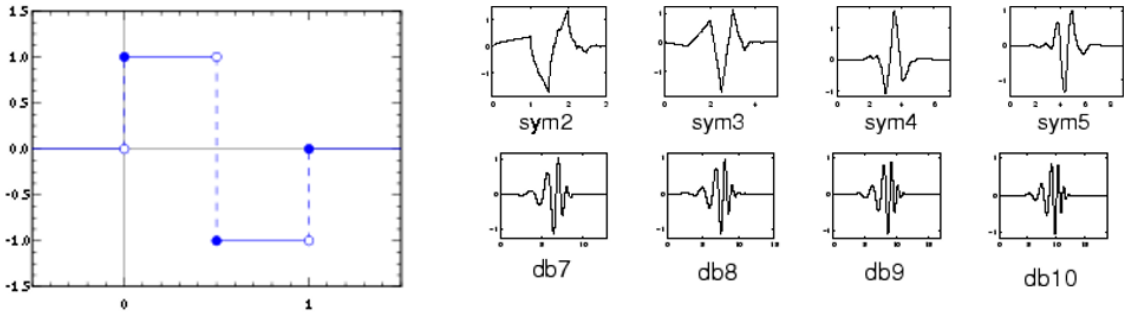

Figura 10. A la izquierda el primer *wavelet* creado por Alfred Haar en 1909, siendo hoy en día el *wavelet* más simple posible. A la derecha, están los *wavelets* de la familia Daubechies (inferior) con momentos de desvanecimiento del 7 al 10, y de la familia Symlet (superior) con momentos de desvanecimiento del 2 al 5.

Pese a que la TWC ha sido catalogada como una buena herramienta representativa, su versión Discreta es preferible en el uso para algoritmos de clasificación [13]. Este hecho concreta un poco más la elección de la familia de *wavelets*, pero lo cierto es que no hay un claro consenso sobre qué familia de *wavelets* es la que proporciona un mejor análisis de la señal EEG. Finalmente, se ha escogido la *Daubechies* 4 (db4) por ser una *wavelet* ampliamente empleada para extracción de características de EEG y su uso en algoritmos de clasificación, y por haber obtenido mejor resultado ha obtenido en términos de precisión, de acuerdo a una revisión completa sobre el uso de *wavelets* para la detección de patrones en el EEG a lo largo de los últimos años [13].

Cuando la TWD es aplicada a una señal, dicha señal será descompuesta en N niveles. El número máximo de niveles de descomposición, aumenta con la frecuencia de muestreo de la señal. Esto es porque la señal almacenada en los coeficientes de bajo nivel será diezmada cada 2 muestras, conforme avancen los niveles de descomposición. De modo que el último nivel de compresión con la información de bajo nivel será equivalente a haber muestreado la señal original cada 2^N muestras, siendo N el nivel de descomposición (Ecuación 3).

$$N_{niveles\ de\ descomposición} = \log_2(f_m) \rightarrow longitud\ señal_{Nivel\ N} = \frac{longitud\ señal\ original}{2^N}$$

Ecuación 3. Relación entre el número de niveles de compresión al aplicar la TWD y la frecuencia con la que la señal ha sido muestreada, que es equivalente a hablar en términos de longitud (en muestras) de la señal.

Para cada nivel de descomposición, se obtiene un par de series de coeficientes (Figura 11). Una de ellas contendrá los coeficientes con información sobre la variación de componentes de la señal en altas frecuencias (H), también llamada información de detalle. La otra serie obtenida contendrá la señal original muestreada cada 2^j muestras, siendo j el nivel de descomposición. Esta última serie mantiene la información de la señal referente a frecuencias más bajas (L), constituyendo un promedio o suavizado de la señal original.

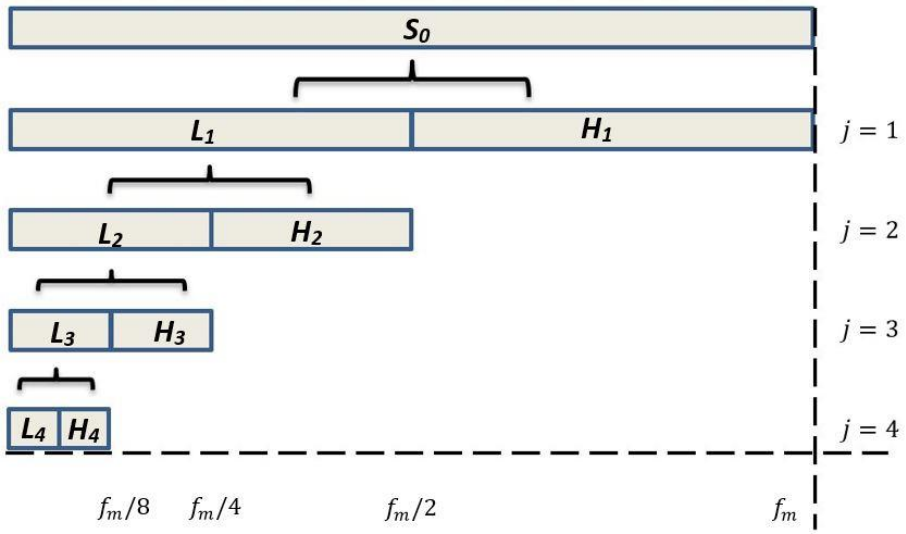

Figura 11. Diagrama en árbol de la *DWT* de la señal S_0, siendo *L* la aproximación con los coeficientes de bajo nivel, y *H* la información de detalle. En el eje horizontal están las sub-bandas contenidas en cada vector de coeficientes, reduciéndose los valores de frecuencia conforme aumentan los niveles de descomposición, representados en el eje vertical.

Por el Teorema de Nyquist, sea una señal $s(t)$ con una frecuencia máxima f_m, que es muestreada con una frecuencia f_s, la relación que debe haber entre la frecuencia de muestreo de la señal, y la frecuencia máxima de esta, para poder reconstruir la señal $s(t)$ a partir de la señal muestreada, es del doble de la frecuencia máxima de la señal f_m (Ecuación 4).

$$f_s \geq 2 \cdot f_m$$

Ecuación 4. Relación entre la frecuencia máxima de una señal y la frecuencia a la que esta debe ser muestreada para evitar el fenómeno de *aliasing*.

En cada nivel de detalle, se refleja la actividad cerebral con referente a una frecuencia máxima de $f_s/2^{j+1}$, siendo *j* el nivel de descomposición de la señal. Gracias al análisis previo de las cabeceras, se supo que todos los registros poseían una frecuencia de muestreo de 512 Hz. De acuerdo con la expresión de la Ecuación 4, y con la Transformada *Wavelet* Discreta de *Daubechies* 4, se determinó un número de nueve niveles de descomposición (descomposición hasta los 0.5 Hz que contienen información sobre actividad cererbral). Para optimizar el espacio en memoria y agilizar el proceso, en una única función se calcularon la TWD y los estadísticos descriptivos de las series de coeficientes obtenidas, para cada *epoch*. La variable *W_EPOCH* en el código es renovada en cada iteración con los coeficientes de la transformación de cada *epoch*.

Tabla 3. Estructura con coeficientes de los 9 niveles de detalle y del promediado del 9º nivel, para una *epoch* de un paciente.

1 EPOCH = 15360 muestras

Serie de coeficientes	H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8	H_9	L_9
Núm. muestras	7683	3845	1926	966	486	246	126	66	36	36

El tiempo de ejecución del código fue de 15360 segundos. A continuación se muestra el ejemplo de una descomposición *wavelet* en los 10 niveles (nueve de detalle y uno de suavizado), para una *epoch*, mediante el uso de la *wavelet* discreta Daubechies 4 (Anejo 2).

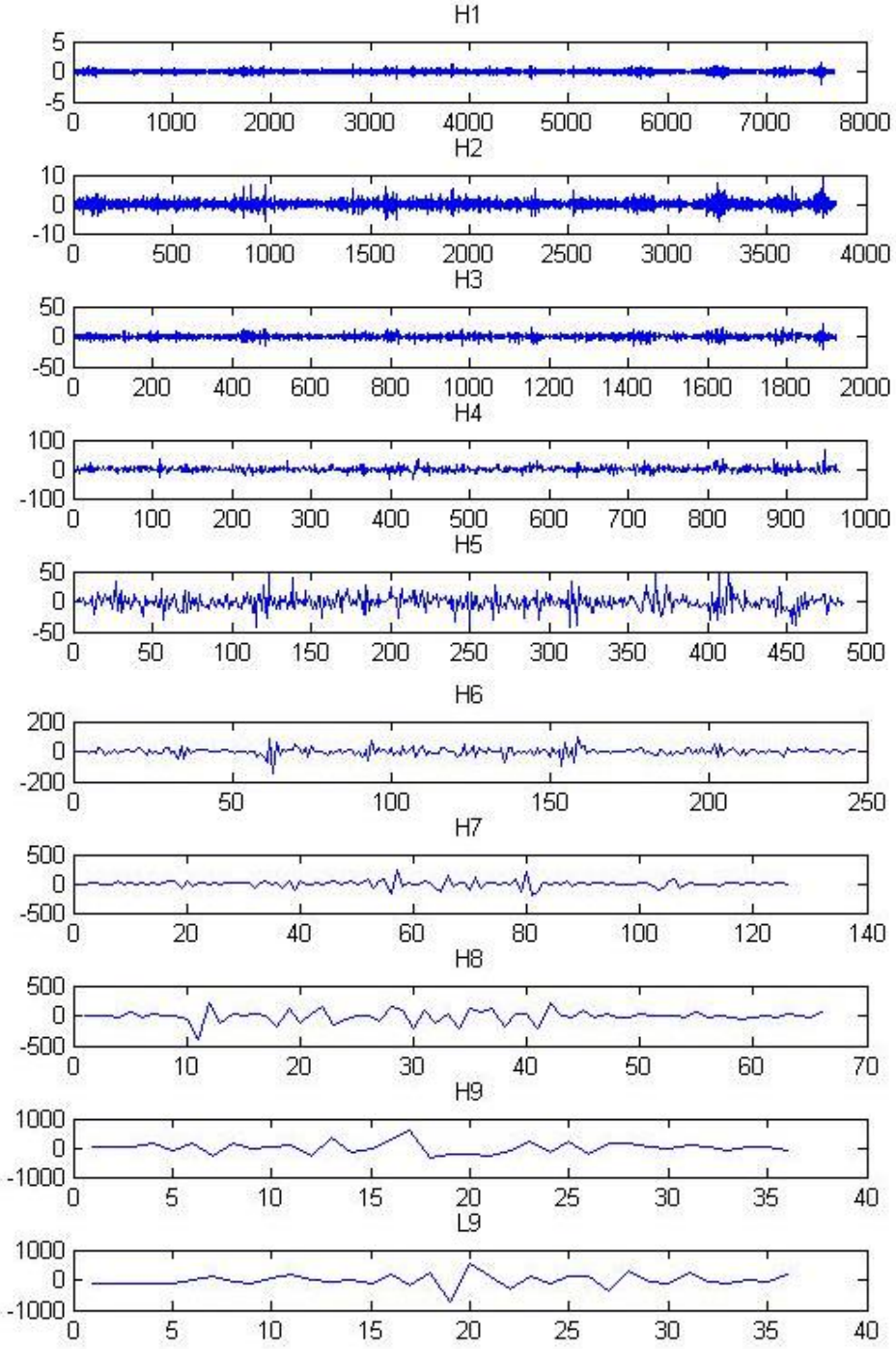

Figura 12. Descomposiciones sucesivas de la señal contenida en una *epoch*, reduciéndose a la mitad el número de muestras en cada serie de coeficientes de detalle. En el eje horizontal se representa el número de elementos de cada vector de coeficientes, que se reduce a la mitad conforme aumenta el nivel de compresión. En el eje vertical, los valores de los coeficientes.

3.3. EXTRACCIÓN DE CARACTERÍSTICAS.

3.3.1. Parámetros sobre los coeficientes de la TWD.

Una vez se obtuvo la descomposición en las distintas bandas de frecuencia, se almacenaron las series de coeficientes con información sobre el nivel de detalle, añadiendo los de coeficientes con información sobre bajas frecuencias obtenidos para el último nivel. De cada una de estas series se extrajeron varios parámetros, teniendo información de distinto orden sobre la distribución de los coeficientes. Los parámetros o características calculados fueron: mediana, desviación típica, coeficiente de asimetría, mínimo, máximo, energía y entropía (Anejo 3).

La mediana es un parámetro de posición alternativo a la comúnmente empleada media aritmética. Se obtiene como el valor central de los observados, lo cual la hace más robusta frente a la media en caso de poblaciones asimétricas, donde algunos valores extremos u *outliers* pueden influir en el valor de la media. Así pues, la mediana x_m se calcula como el valor que ocupa la posición m-ésima en la serie ordenada del conjunto de N individuos, variando el índice *m* según si el número de observaciones es par o impar (Ecuación 5).

$$x_m = \begin{cases} m = \dfrac{N+1}{2} \; si\; N\; es\; impar \\ m = \dfrac{x_{N/2+1} + x_{N/2}}{2} \; si\; N\; es\; par \end{cases}$$

Ecuación 5. Fórmula de la mediana (x_m) para una muestra de N elementos**.**

Ligada al estudio de la variabilidad en una población, la desviación típica es una de las medidas de dispersión más conocida y utilizada. Su valor refleja la proximidad o lejanía de los individuos de una población respecto al valor medio de esos mismos datos. Como ventaja frente a su valor al cuadrado, es decir la varianza, la desviación típica mantiene las unidades en las que se expresan las características cuya variabilidad es estudiada.

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}$$

Ecuación 6. Fórmula de la desviación típica para una muestra de N elementos y media $\bar{x}$.

Valores mayores de desviación típica indicarían cambios importantes en magnitud de la señal, intuyéndose una morfología más abrupta por distancias mayores respecto al promedio (Ecuación 6). Dichos valores altos pueden ser un posible síntoma de presencia de picos a lo largo de la serie de coeficientes. Análogamente, valores pequeños de desviación típica suelen ir ligados a comportamientos "tranquilos" o de muy altas frecuencias, en las que la mayor parte de la señal sea ruido gaussiano.

El coeficiente de asimetría es un estadístico que describe la forma de la distribución de los datos informando sobre si la muestra se distribuye de forma simétrica o no respecto a su media. El coeficiente toma valores positivos o negativos según si la muestra tiene colas alargadas hacia la derecha o la izquierda de la media (Ecuación 7), respectivamente.

$$CA = \frac{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^3}{N-1}}{s^3}$$

Ecuación 7. Fórmula del Coeficiente de Asimetría en la serie de N elementos, promedio $\bar{x}_j$ y desviación típica s_j .

Los parámetros máximo y mínimo en un conjunto de observaciones reflejan los extremos entre los cuales se distribuyen las observaciones sobre los datos. Estos valores extremos, sin un preprocesado previo, son candidatos a *outliers* o valores atípicos para una distribución normal de las observaciones. En la distribución de los valores que tomen los coeficientes de la TWD, los máximos y mínimos en cada nivel de detalle determinan a los rangos de valores en cada serie de coeficientes. Su cálculo sobre los datos de los coeficientes se realizó, sin previa obtención del valor absoluto, y de forma global (Ecuación 8).

$$H(x)\begin{cases} x_a = \min_H si\ x_a \leq H(x)\ \forall x \\ x_b = \max_H si\ x_b \geq H(x)\ \forall x \end{cases}$$

Ecuación 8. Expresiones de mínimo (x_a) y máximo (x_b) absolutos de la serie *H*.

Cuando una señal se lleva al dominio de la frecuencia mediante la transformación de su expresión temporal en la suma de señales de distintas frecuencias, ponderadas cada una de ellas por un coeficiente, la energía puede calcularse como el sumatorio de los cuadrados de los coeficientes. Puesto que cada nivel de descomposición tras la obtención de la TW (H_j, L_j) viene representado por una serie finita de coeficientes, se obtuvo la energía de la banda de frecuencias referente a cada nivel de descomposición, en base a las series de coeficientes de cada nivel (Ecuación 9).

$$E = \sum_{i=1}^{N} x_i^2$$

Ecuación 9. Expresión empleada en el cálculo de la energía para una serie de N elementos.

Por último, la entropía suele ser descrita como una medida del desorden en el conjunto de valores de una función. El cálculo de la entropía de Shannon en el ámbito de la información sobre señales, expresa la incertidumbre en un conjunto de variables aleatorias. Dada una secuencia s, si al aumentar su tamaño disminuye la relación entre variables, entonces aumenta la aleatoriedad de la serie y por tanto la entropía en valor absoluto.

$$S_j = s_i^2 \log(s_i^2)$$

Ecuación 10. Fórmula para el cálculo de la entropía de la serie de coeficientes del nivel *j*, con *i* = 1…N elementos.

Teniendo en cuenta la longitud de la *epoch* y la frecuencia de muestreo del registro ($30\ seg \cdot 512\ Hz = 15360\ muestras$), se puede llegar a la longitud del vector de coeficientes de detalle para la primera descomposición mediante la TWD. Así pues, puede obtenerse una idea acerca de la compresión que supone la ejecución de este paso, caracterizando vectores de longitudes aproximadas a 10^3 (en los niveles de descomposición más tempranos) mediante un conjunto de 7 parámetros.

3.3.3. Información sobre los sujetos.

Una vez se dispone de la información de cada individuo en un solo vector, se le ha añadido a ese vector con características sobre el paciente, variables con información sobre el sujeto, que podrían ser vinculantes a la hora de realizar una clasificación de los pacientes. Estas variables han sido: la duración del registro, la edad del paciente y el género.

La duración se expresa en horas y puede obtenerse de directamente de las cabeceras buscando la información referente a ese campo, o indirectamente empleando la información sobre la frecuencia de muestreo (ya cargada llegado este punto), y calcular la duración en tiempo a partir de la longitud del registro y la frecuencia de muestreo.

Tanto la edad como el género del paciente son obtenidos directamente de las cabeceras de los registros, en caso de que se quisiese disponer de la información. La edad está almacenada en años, mientras que el sexo es una variable que puede tomar valor “M” si el sujeto es hombre (del inglés *Male*) o “F” si se trata de una mujer (del inglés *Female*). Para poder trabajar con estas variables en formato texto, se genera una variable Dummy que puede tomar valor 0 o 1 según si el campo con información sobre el género tiene valor “M” o “F” (Anejo 4).

3.4. PREPROCESADO DE LOS DATOS.

Los valores de los coeficientes obtenidos tras la transformación al espacio tiempo-frecuencia, son empleados para el cálculo de los parámetros descritos en el apartado 3.3.1. Parámetros sobre los coeficientes de la TWD. Un aspecto técnico importante en la aplicación de métodos de estudio Multivariante, como es el caso del Análisis de Componentes Principales posteriormente aplicado (3.5.1. Análisis de Componentes Principales), es que para estudiar de forma veraz la variabilidad de los datos, es necesario trabajar con variables que sean comparables entre sí. El acondicionamiento de los datos en crudo, se conoce como etapa de preprocesado. Es una fase crítica que condicionará resultados y conclusiones posteriores, habiendo varios preprocesados posibles. La elección del más conveniente depende en gran parte del conocimiento sobre la naturaleza de los datos con los que se trabaje.

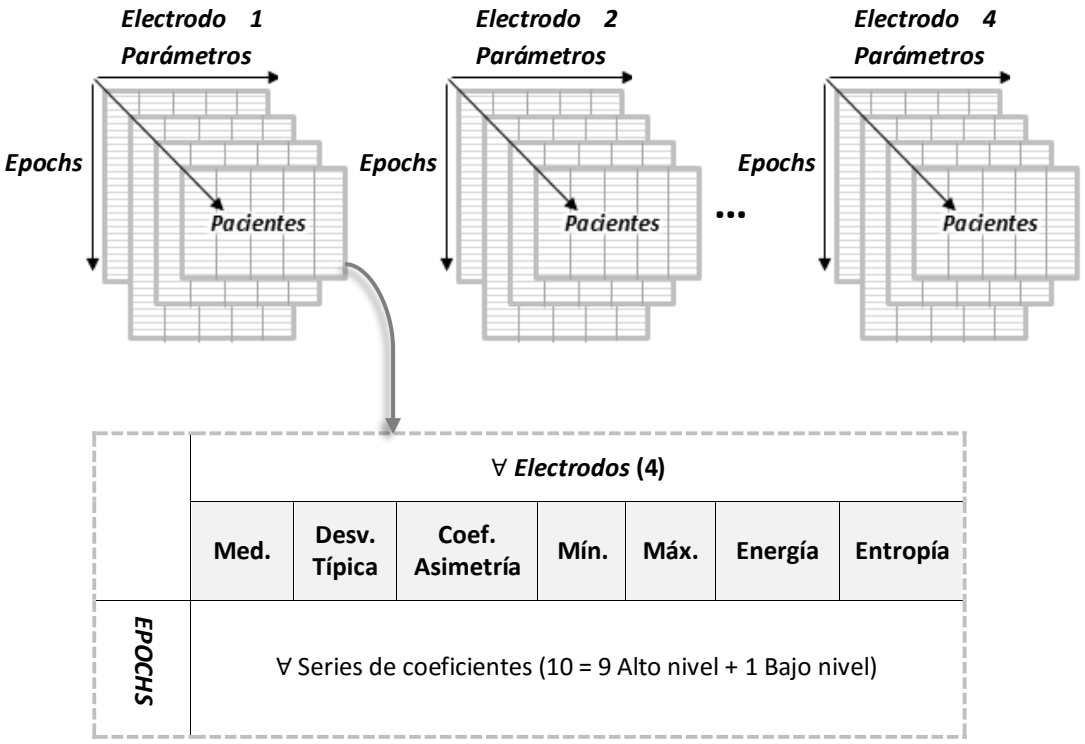

Figura 13. Organización de los estadísticos descriptivos de los coeficientes de detalle de la Transformada *Wavelet*, para cada nivel de descomposición, *epoch*, electrodo y paciente.

3.4.1. Datos faltantes.

Distintas herramientas estadísticas han sido desarrolladas para suplir la ausencia de información, dada la recurrencia del problema originado por los datos faltantes. Una de las posibles soluciones es la imputación de datos faltantes mediante la estimación de la variabilidad de la matriz con los datos faltantes, basándose en un Análisis de Componentes Principales. El algoritmo estima los valores de los datos faltantes teniendo en cuenta la información existente en la matriz inicial. La implementación de este método se ha realizado empleando una Toolbox libre para Imputación de Datos Faltantes [15], desarrollada en el Departamento de Estadística e Ingeniería Operacional y Calidad, por el Grupo de Ingeniería Estadística Multivariante (GIEM) de la Universitat Politècnica de València (UPV).

Una visualización de las celdas vacías (valor NaN), permitió hacer un rápido *screening* de la población. En los sujetos 1, 15, 17 y 31 hay valores NaN. Dichos valores suponen aproximadamente un tercio de las matrices de datos. Puesto que se poseen 3/4 de la información de cada individuo, puede emplearse un programa desarrollado para trabajar con datos faltantes, rellenando de valores los elementos con valor NaN.

Previamente a calcular los datos faltantes, se comprobó la configuración óptima de los parámetros. El número de componentes consideradas repercute en el ajuste de los datos generados que suplen a los valores faltantes al resto de datos. Un mayor número de componentes puede obtenerse, pero a costa de mayor coste computacional. Para optimizar el valor de este parámetro, se calculó el error de la nueva matriz, con 5, 6, 7, 8, 9 y 10 componentes, obteniéndose la diferencia entre el resultado con cada parámetro y con el consecutivo. Los resultados mostraron que las diferencias eran en todos los casos nulas aproximadamente, con un orden de 10^{-15}. Por el menor coste que implica, se escogieron 5 componentes. Seguidamente se calcularon los datos faltantes en los pacientes indicados. Ejecutando las líneas de comando anteriores que permitieron visualizar los individuos erróneos, se comprobó que todas las posiciones de las matrices de datos, contenían información y no estaban vacías.

3.4.2. Autoescalado

La estrategia más sencilla para el preprocesado pasa por someter al conjunto de datos inicial a un centrado y posterior escalado dividiendo entre la desviación típica de cada variable. El centrado consiste en la sustracción a cada vector columna (variable original) de su valor medio. De esta forma se puede asegurar que el promedio de cada variable estará centrado en cero.

Por otro lado, la magnitud de las variaciones dictaminará cuán importantes sean a la hora de explicar la variabilidad de los datos, según el modelo. Para equiparar los rangos en los que se puedan distribuir las distintas variables, y hacerlas comparables entre sí, es común realizar un escalado. Normalmente este escalado consiste en la división de los datos entre la desviación típica de la variable a la que pertenezcan, habiendo sido previamente centrados, también respecto a la media de la variable en cuestión. Este tipo de preprocesado, permite aproximar la distribución de todos los datos a una Normal de media nula y varianza unitaria (Figura 9).

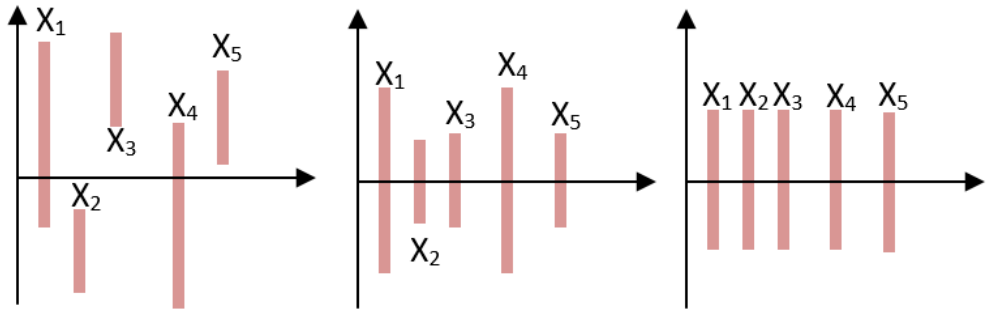

Figura 14. Representación esquemática del efecto del autoescalado sobre los datos. A la izquierda se muestran las distintas variables (X) situadas en distintos puntos del espacio. En la segunda imagen, se ha realizado el centrado de los datos, teniendo todas las variables media 0. En la última imagen, los datos han sido escalados tras haber sido centrados, lo que les aproxima a la distribución normal de una variable.

3.4.3. Escalado por bloques

A diferencia del Autoescalado, el Escalado por bloques pretende respetar las diferencias entre variables. De forma similar al procedimiento anterior, el primer paso de esta estrategia también consiste en centrar los datos. La diferencia está en el factor por el que se dividen estos datos una vez centrados, siendo esta vez, la desviación típica de los datos de un mismo bloque (Figura 15). En el autoescalado por bloques, por tanto, es necesario seccionar el conjunto de muestras en varios bloques.

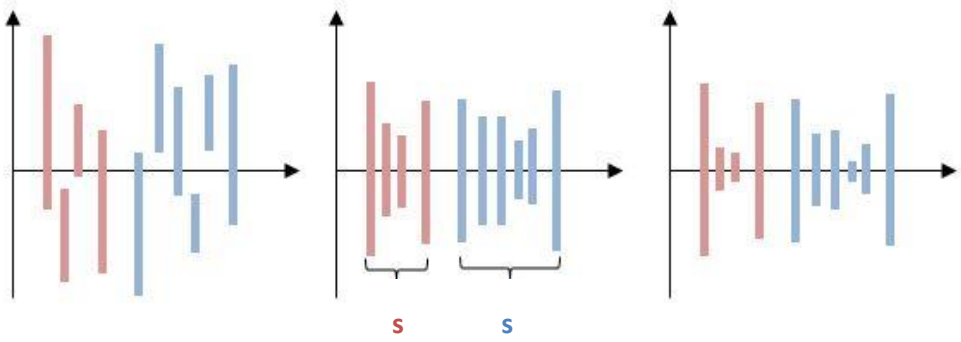

Figura 15. Representación del efecto del escalado por bloques. Cada color representa un bloque, y cada barra una variable. Los datos son centrados de la primera a la segunda imagen. A continuación se calcula la desviación típica de cada bloque para escalar todas las variables pertenecientes a ese bloque, manteniendo las diferencias entre varianzas.

Mientras que con el autoescalado todas las variables se asemejan a una Gaussiana con varianza unitaria, esto no ocurre en el caso del escalado por bloques. Si se aplica un autoescalado variables que presenten pequeños cambios y supongan ruido, estarán siendo amplificadas frente al resto de variables que sí posean mayor variabilidad. Sin embargo, mediante el escalado por bloques las variables con varianzas significativamente superiores respecto a la global serán magnificadas. Es decir, aplicando escalado por bloques, cuando PCA analice las variables buscando aquellas que mejor expliquen las diferencias entre individuos, se habrán conservado las diferencias entre las variabilidades introducidas por cada variable.

Así pues, se decidió realizar un Escalado por bloques tras la obtención de parámetros sobre los coeficientes *wavelet*. De este modo, evitando la pérdida de información sobre la varianza en los coeficientes de cada nivel de compresión, seguirá siendo mayor la varianza en las series donde la señal tenga mayores variaciones y por tanto, corresponda a frecuencias inferiores y con información relacionada con la actividad cerebral durante el sueño.

3.5. MODELOS DE COMPRESIÓN DE VARIABLES.

Pese a haber reducido considerablemente el número de variables aplicando los métodos descritos hasta este punto, existen técnicas más sofisticadas que permiten realizar una selección de aquellas características con información relevante sobre los factores influyentes en el proceso estudiado. Este tipo de estudio puede realizarse mediante varias técnicas [15] [16], como Análisis de Componentes Principales (*PCA*) o Análisis de Componentes Independientes (*ICA*), que constituyen una herramienta verdaderamente útil para el análisis de EEG, siendo ampliamente empleadas para el estudio de no sólo de esta señal, sino de otras actividades fisiológicas.

3.5.1. Análisis de Componentes Principales

El Análisis de Componentes Principales (*Principal Component Analysis,* PCA) es una técnica estadística Multivariante que permite estudiar la relación entre variables de un conjunto de datos, y cómo éstas influyen en la variabilidad entre los individuos. Extrayendo información contenida en la propia estructura del conjunto de observaciones, PCA es un modelo útil para la compresión de información, reduciendo la dimensionalidad del conjunto de datos.

La finalidad es construir nuevas variables denominadas latentes o Componentes Principales, que expliquen las principales fuentes de variabilidad la matriz de datos. PCA realiza una búsqueda de esas fuentes de variabilidad en los datos, denominándolas Componentes Principales (CP), y estando formadas por combinaciones lineales de las variables iniciales. Así pues, la primera Componente Principal obtenida tras el análisis, será una nueva variable latente que explicará la mayor parte de variabilidad entre la muestra estudiada. Para no recaer en la redundancia de información, el algoritmo que genera las nuevas Componentes Principales, tiene como condición el asegurar la ortogonalidad entre éstas.

Sea una matriz $\boldsymbol{X}$ de m variables aleatorias representando una población de n individuos con $m > n$, PCA buscará las r variables latentes que expresen el comportamiento de las m variables iniciales, siendo $r \leq m$. Cada CP se representa por las proyecciones de los n individuos sobre las nuevas variables latentes (*scores,* $\boldsymbol{t}$) y por las proyecciones de las variables originales sobre cada componente principal (*loadings,* $\boldsymbol{p}$). Los vectores de *loadings* pueden obtenerse como vectores propios de la matriz $\boldsymbol{X}^T\boldsymbol{X}$, asegurándose así la ortogonalidad entre las Componentes Principales, y junto con la imposición de norma unitaria para cada vector $\boldsymbol{p}$, otorga al conjunto de *loadings* la condición de ortonormalidad.

$$\boldsymbol{p}_1 \perp \boldsymbol{p}_2 \; ; \; \|\boldsymbol{p}\| = 1 \rightarrow \boldsymbol{p}^{-1} = \boldsymbol{p}^T \; (1)$$

$$\boldsymbol{T} = \boldsymbol{X}\boldsymbol{P} \overset{(1)}{\rightarrow} \boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^T$$

$$\boldsymbol{X} = \boldsymbol{t}_1\boldsymbol{p}_1^T + \boldsymbol{E}_1 \rightarrow \boldsymbol{X} = \boldsymbol{t}_1\boldsymbol{p}_1^T + \boldsymbol{t}_2\boldsymbol{p}_2^T + \boldsymbol{E}_2 \rightarrow \boldsymbol{X} = \boldsymbol{T}_r\boldsymbol{P}_r^T + \boldsymbol{E}_r \; (8)$$

Ecuación 11. Expresión de la descomposición de la matriz inicial **X** en el producto de la matriz de *score*s (**T**) por la matriz de *loading*s (**P**), representando este producto las Componentes Principales, más un residuo **E**.

La variabilidad en $\boldsymbol{X}$ que no ha logrado ser explicada por la primera componente, se agrupa en el residuo para la primera aproximación, $\boldsymbol{E_1}$(Ecuación 8). Ese término se computa como la diferencia entre la matriz original y la aproximación con la primera componente. El término $\boldsymbol{E_1}$ será analizado en la siguiente iteración. Su CP será la segunda del modelo explicativo de la $\boldsymbol{X}$, expresada mediante un vector de *scores* ($\boldsymbol{t_2}$) y de *loadings* ($\boldsymbol{p_2}$). Así se genera el espacio $\boldsymbol{T}$ con las nuevas variables latentes, de dimensionalidad menor que el espacio de datos original, $\boldsymbol{X}$ (Figura 16).

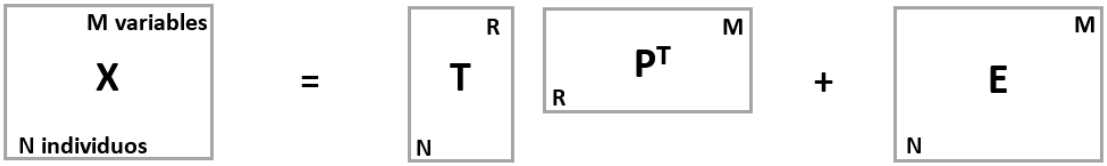

Figura 16. Expresión matricial de la descomposición de la matriz de datos original **X**, en un conjunto de *scores* **T** resultado de proyectar cada individuo sobre el nuevo espacio de variables latentes, **R**.

La explicación gráfica del algoritmo PCA, suele emplearse por ser considerada bastante intuitiva, dado que PCA puede entenderse como una transformación lineal del sistema de referencia inicial, imponiendo como nuevos ejes, las componentes principales (Figura 17). Los *scores* representan las nuevas coordenadas de cada individuo. Así pues, la primera coordenada hará referencia a la primera componente, siendo la que mayor variabilidad explique.

El número de Componentes Principales a obtener, puede estimarse de distintas formas. Una de ellas, es mediante la suma de la variabilidad explicada por cada componente. De forma gráfica, mediante un frente de Pareto se puede visualizar hasta qué CP se debe llegar para explicar suficientemente la variabilidad de los datos. De nuevo, esta es otra cuestión subjetiva que dependerá del proceso analizado y del propósito del análisis en cuestión. Una descripción de distintas estrategias para escoger el número de componentes, se describe en [17].

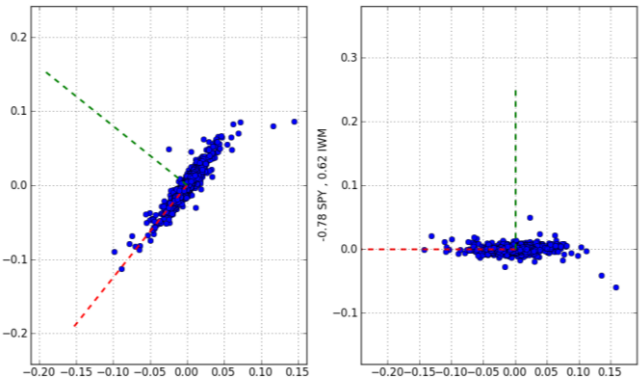

Figura 17. La interpretación intuitiva de PCA es su uso como transformación lineal del sistema de coordenadas inicial, a unos nuevos ejes que sean las Componentes Principales, siendo la Primera Componente, la que mayor parte de variabilidad entre los datos explique. En la imagen puede verse a la izquierda la distribución inicial de los datos en el espacio, siendo la CP_1 la línea discontinua roja y la CP_2 la verde. En la imagen de la derecha, son las dos primeras CP las que forman el nuevo sistema de coordenadas de los datos.

Los resultados del PCA pueden emplearse como características predictivas en modelos de clasificación. Sin embargo, existen ciertas limitaciones de este método. En primer lugar, PCA eliminará redundancia entre los datos, estableciendo las combinaciones lineales que aglomerarán las variables iniciales implicadas en una misma fuente de variabildiad, agrupándolas en la primera CP. No obstante, podría ser que variables estuviesen relacionadas de una forma más sutil, y por tanto las Componentes Principales extraídas no aislasen

completamente las distintas fuentes de variabilidad en los datos [18]. Además, la inexistencia de un método fijo para el preprocesado de los datos puede marcar la diferencia entre un PCA exitoso y un PCA que no desvele información sobre la estructura de los datos.

Tras la aplicación de PCA, se introdujeron en el conjunto de datos los parámetros descritos en el apartado *Extracción de características,* aplicados en este caso a los *scores* de las Componentes Principales seleccionadas, el porcentaje de variabilidad explicada y el de variabilidad sin explicar para el número de CP escogido.

Tabla 4. Distribución espacial de las variables en la matriz de observaciones **X**. T_1 y T_2 son los siete parámetros obtenidos de los *scores* de las dos primeras Componentes obtenidas tras el PCA. P_1 y P_2 son los vectores de *loadings* de esas mismas variables latentes, conteniendo información sobre los pesos de las características obtenidas para cada electrodo, en cada componente obtenida con PCA. Las columnas Var.Ex. y Var.No.Ex son las variabilidades explicadas y sin explicad al quedarse solo con las dos primeras componentes de PCA. Finalmente, el útlimo grupo hace referencia al Tiempo de sueño del paciente, a su Edad y a su Sexo.

Variables	**T_1T_2**	**P1**				**P2**				**Var .Ex**	**Var. No Ex.**	**T,E,S**
		E. 1	**E. 2**	**E. 3**	**E. 4**	**E. 1**	**E. 2**	**E. 3**	**E. 4**			
Columnas	14	84	154	224	294	364	434	504	574	575	576	579

Con la matriz formada por individuos representados por estas características (Tabla 4) se construye la matriz de observaciones **X**. Esta matriz será la empleada como entrada para la generación de los modelos explicados en el siguiente apartado.

3.6. MODELOS DE CLASIFICACIÓN.

Un modelo de clasificación realiza una predicción de la clase a la que pertenece uno de los sujetos de una población de la cual se han extraído ciertas características o variables, en base a las cuales se establece la discriminación entre individuos. Para la obtención de un modelo, existen diversos algoritmos de generación, basados en estrategias matemáticas diferentes y obteniendo distintos resultados.

Una implementación completa de un sistema de clasificación pasa por dos fases, una primera de entrenamiento en la que se construye el modelo y una posterior en la que se valida el modelo con nuevas observaciones, conocida como fase de validación. Debido al bajo número de individuos en el conjunto de observaciones, en este caso se realizó la fase de entrenamiento únicamente.

La validación del modelo generado ha sido una validación cruzada realizada de forma interna por el software empleado para el análisis estadístico Multivariante de los datos. La validación cruzada consiste en dividir la matriz introducida al algoritmo, en distintas particiones. Cada una de ellas es empleada como conjunto de validación del modelo generado por el resto de particiones.

Los modelos empleados en el trabajo fueron un K-Vecinos Más Próximos (KNN) y un Análisis Discriminante basado en Mínimos Cuadrados Parciales (PLS-DA). Con el segundo modelo, se realizó en primer lugar la clasificación con un conjunto de treinta individuos con cinco de cada clase, y posteriormente se realizó un segundo análisis para cada clase (con número desigual de individuos en cada una) frente al resto.

3.6.1. Modelo K-Vecinos Más Próximos.

El método *K-Nearest Neighbor* (KNN) aborda la clasificación de una población de *n* individuos asignándoles la familia o clúster de entre las existentes. El criterio para la elección de uno de los clústeres posibles es que este suponga la clase más cercana al individuo clasificado. La cercanía o lejanía se calcula mediante una distancia, siendo posible escoger entre varias métricas posibles: distancia Euclídea estandarizada, distancia de Mahalanobis, Correlación, etc. La distancia escogida para la clasificación con KNN, fue la distancia coseno. Esta distancia disminuye conforme aumenta la semejanza entre los individuos, es decir, la distancia será menor cuanto más similares sean las características que representan a cada individuo.

La estrategia seguida por el algoritmo de clasificación, tiene como objetivo minimizar la varianza intraclase (esto es, entre los individuos de una misma clase) y maximizar la interclase (entre los modelos representativos de cada clase). El avance de la clasificación puede producirse también de distintas formas. En concreto, la opción utilizada por el software empleado para la clasificación en este trabajo ejecuta el algoritmo *Exhaustive Search* (Figura 20). Como su nombre indica, este algoritmo realiza una búsqueda exhaustiva entre los individuos del conjunto de datos para clasificar muestra. Calcula la distancia del individuo a clasificar, a todos los individuos del conjunto de datos, devolviendo como solución los *k* vecinos que hayan obtenido una distancia menor, junto con la distancia obtenida.

La validación cruzada empleada en la implementación del modelo siguió una estrategia *leave-one-out*, basada en hacer las particiones de los datos separando un individuo como conjunto de validación y empleando al resto como conjunto de entrenamiento. Este método de validación es el más conveniente teniendo en cuenta el bajo número de individuos de cada clase, siendo éste el típico escenario en el que es aconsejable emplear el algoritmo *leave-one-out*, aún a costa de un elevado coste computacional, que es todavía mayor cuando se hace doble validación, como es el caso.

3.6.2. Análisis Discriminante basado en Mínimos Cuadrados Parciales.

A raíz del método *Partial Least Squares,* empleado inicialmente para la obtención de los coeficientes de regresión en la generación de un modelo explicativo de los datos, se genera una variante de éste cuya aplicación es realizar un Análisis Discriminante entre los individuos. A diferencia de PCA, en su concepción original, el método PLS obtiene las variables que explican en mayor grado la variabilidad en los datos teniendo en cuenta la relación entre la matriz de variables explicativas (**X**) y la matriz de variables respuesta (**Y**). Esto es gracias a la introducción de nueva información además de los datos en sí mismos, añadiendo la matriz respuesta que contiene tantos posibles valores para una variable categórica, como clases de sujetos.

El método PLS equivale el estudio interno de las fuentes de variabilidad tanto para **X** como para para **Y**, con el objetivo de encontrar fuentes comunes que maximicen la covarianza entre la matriz de observaciones (**X**) y la respuesta (**Y**).

Una de las ventajas de emplear PLS-DA frente a PCA, es que en caso de emplear el segundo, puede que variables que no tengan una gran variabilidad, sean descartadas como candidatas a formar parte de las Componentes obtenidas, o al menos de las primeras. Sin embargo, PLS-DA al contar con información sobre la respuesta del proceso, es capaz de detectar variables en X que pese a variar poco, afecten a la respuesta en Y. En otras palabras, PLS-DA considera importantes variables a las cuales el proceso es sensible. Esto es que con variaciones de dicha variable se producen cambios en la respuesta del proceso. Si dichas variables a las que el proceso es sensible no varían tanto como el resto de variables de la matriz X, con PCA se correría el riesgo de no considerar importantes dichas variables.

Así pues, para la generación del modelo de clasificación PLS-DA, se debe contar con una matriz de variables explicativas (*X*), que en este caso particular será la matriz de *n x m* donde *n* sea el número de pacientes y *m* el de variables. Además, se imputa el vector con la variable respuesta (Y). Esta variable respuesta, es en realidad una matriz de ceros y unos con tantas filas como individuos y tantas columnas como clases. Un individuo valor uno o cero según si pertenece o no, respectivamente, a la clase representada por cada columna. En este caso, habrá seis columnas referentes a las seis clases de patologías consideradas: Insomnio, Sano, Narcolepsia, Epilepsia Nocturna del Lóbulo Frontal, Movimientos Periódicos de Piernas y Trastorno en la Conducta del Sueño REM.

4. RESULTADOS

4.1. MODELO K-VECINOS MÁS PRÓXIMOS

Los resultados obtenidos en la clasificación KCC fueron evaluados empleando distintas métricas, así como vecinos. Finalmente, se han dejado los dos mejores resultados, ambos calculados empleando la distancia coseno, que representa una medida de la similitud entre dos individuos, mediante la comparación del vector de características que los define. El algoritmo fue configurado aplicando el método *Exhaustive Search*, la distancia Coseno y Doble Validación Cruzada *leave-one-out* con un ratio de entrenamiento/validación de 0.7.

La forma de analizar una tabla de clasificación es estudiando los elementos con cierta coordenada en las filas, caigan en la misma coordenada para las columnas. Éste es el resultado deseado debido a que las filas representan las clases reales de los individuos, y las columnas la clase que le es asignada, formándose una diagonal si el ratio de aciertos es del 100%.

Tabla 5. Tablas de clasificación para conjunto de 30 individuos con k = 2 vecinos. Izquierda: Validación cruzada, porcentaje de aciertos: 33.33%. Derecha: Doble validación, porcentaje de aciertos: 44.44%.

	1	**2**	**3**	**4**	**5**	**6**
1	1	0	0	0	2	0
2	0	3	0	0	1	0
3	1	0	0	1	1	0
4	0	0	1	3	0	0
5	1	2	0	1	0	0
6	1	0	0	2	0	0

	1	**2**	**3**	**4**	**5**	**6**
1	2	0	0	0	0	0
2	0	0	0	0	1	0
3	0	1	1	0	0	0
4	0	0	0	1	0	0
5	0	1	0	0	0	0
6	0	0	0	1	1	0

Tabla 6. Tablas de clasificación para conjunto de 30 individuos con k = 4 vecinos. Izquierda: Validación cruzada, porcentaje de aciertos: 33.33%. Derecha: Doble validación, porcentaje de aciertos: 66.67%.

	1	**2**	**3**	**4**	**5**	**6**
1	3	0	0	0	1	0
2	0	0	0	1	2	0
3	0	1	0	1	1	0
4	0	1	0	3	0	0
5	1	0	1	1	1	0
6	1	0	0	2	0	0

	1	**2**	**3**	**4**	**5**	**6**
1	1	0	0	0	0	0
2	0	1	0	0	1	0
3	0	0	1	0	0	1
4	0	0	0	1	0	0
5	0	0	0	0	1	0
6	1	0	0	0	0	1

El ratio de aciertos obtenido con k = 2 vecinos y doble validación es del 33.33%, mientras que con 4 vecinos, del 66.67%. Esto quiere decir que la varianza interclase no es lo suficientemente alta como para definir claramente una frontera entre las distintas agrupaciones correspondientes a las distintas clases. Por tanto, cogiendo un número menor de vecinos, las probabilidades de asignar al elemento a clasificar, una clase correcta o incorrecta, son similares.

Sin embargo, de los resultados también se desprende que sí existe la suficiente semejanza intraclase como para reconocer agrupaciones de elementos. Así pues, al aumentar el número de vecinos considerados en la clasificación de un nuevo individuo, aumenta la probabilidad de clasificarlo correctamente, debido a que la cercanía de los sujetos de su misma clase, será mayor que la de elementos de otras clases. Esta es la razón por la cual con k = 4 vecinos, elementos de las clases 2, 5 y 6 (sanos, movimiento periódico de piernas y trastorno de la fase REM), son clasificados como tal, mientras que para k = 2 vecinos, sólo lo son sujetos con insomino, narcolepsia y epilepsia nocturna (1, 3 y 4 respectivamente).

Por otro lado, en los resultados plasmados en las tablas de clasificación, también puede verse qué tipo de error en la asignación de clases se ha cometido. En este caso, se puede observar que elementos de la clase 2, son clasificados como de la clase 5; de la clase 3 como clase 6 y de la clase 6 como clase 1. Pese a que los resultados descritos sean errores de clasificación, podrían ser un indicio de que la información referente a la patología, ha persistido todo el proceso matemático hasta la llegada al clasificador.

El motivo de esta sospecha, reside en que entre narcolepsia y trastorno de la fase REM, existe relación clínica que indica existencia en ambos casos de pérdida de atonía muscular en la fase REM (el sujeto realiza los movimientos que protagoniza en su sueño) [19] .

Sin embargo, el bajo número de individuos no es suficientemente grande como para poder afirmar la sospecha. Serían necesarios estudios con una mayor población, para así poder complementar una realidad a nivel clínico, que es en este caso, el desarrollo de estudios que permitan confirmar que estadios iniciales de pérdida de atonía muscular durante la fase REM (trastorno del comportamiento en la fase REM), pueden ser un indicador previo a enfermedades más severas como la narcolepsia.

4.2. ANÁLISIS DISCRIMINANTE BASADO EN MÍNIMOS CUADRADOS PARCIALES

Para la interpretación de los resultados mostrados a continuación, debe conocerse el significado de los gráficos con los coeficientes R^2 y Q^2, así como de los gráficos con los valores Observados frente a los Predichos para las observaciones de cada clase.

El coeficiente R^2 (azul en los gráficos) es la bondad de ajuste del modelo construido y el coeficiente Q^2 (rojo en los gráficos) la bondad de predicción, variando ambos según del número de componentes que el modelo tenga en cuenta. Ambas bondades serán mayores conforme más próximas sean al valor 1. Sin embargo dichos coeficientes no tienen por qué aumentar al unísono. En los resultados que se muestran a continuación es recurrente el aumento de la bondad de ajuste con un mayor número de componentes, mientras que la bondad de predicción queda estancada sin crecer aunque lo haga la cantidad de componentes consideradas. La interpretación de este suceso, es que en dichos casos se obtiene un modelo capaz de explicar muy bien la variabilidad en los datos (por tanto tendrá una R^2 cercana a 1), pero sobreajustado a esos mismos datos, teniendo dificultades para explicar nuevas observaciones (por tanto tendrá un poder predictivo Q^2 muy limitado).

Con el fin de intentar paliar la diferencia entre el poder explicativo y el predictivo, se realizó una selección de las características que contuviesen una cantidad de información y predictiva considerada importante para la discriminación posterior de individuos. Estas variables se pueden evaluar mediante la obtención de Gráficos de Importancia de las Variables (*Variable Importance Plots).* Las características con valores menores a la unidad, no son "importantes", mientras que las que igualen o superen dicho valor, sí lo serán.

Por otro lado, en base al modelo considerado se elaboran predicciones de las clases a las que pertenecen las observaciones. La respuesta contenida en la matriz Y, es una variable *Dummy* que tomará el valor 1 si el individuo al que hace referencia, pertenece a la clase correspondiente a cada columna. Los valores observados se representan con una línea discontinua situada en uno o en cero, según el valor Y de cada individuo para cierta clase.

Una discriminación será óptima cuando los valores predichos para cierta clase ajusten a la línea discontinua situada en uno (valores observados para dicha clase), y los valores predichos para los individuos que no pertenezcan a esa clase, se sitúen entorno a la línea discontinua situada en cero. Si la discriminación prima la sensibilidad, priorizará detectar a todos los individuos de cierta clase, normalmente a costa de reducir la especificidad, que equivale a incluir individuos en dicha clase que no lo sean.

Este equilibrio entre ambas suele dar lugar a una solución de compromiso entre sensibilidad y especificidad, priorizando una u otra según las circunstancias de la aplicación del modelo. Dichos valores de sensibilidad y especificidad, variarán según se defina la frontera de decisión entre individuos pertenecientes y no-pertenecientes a cierta clase. En este caso la frontera establecida fue un umbral en 0.5 para evaluar la discriminación en base a las predicciones del modelo PLS, construyendo así una matriz de confusión.

Tabla 7. Estructura de una Matriz de Confusión (Valor Observado, Valor Predicho) en la que los resultados corresponden a una Clasificación correcta ((1,1); (0,0)), a Falsos Negativos (1,0) o a Falsos Positivos (0,1).

Valores Predichos **Valores Observados**	**1** **(Pertenece a clase)**	**0** **(No pertenece a clase)**
1 **(Pertenece a clase)**	Clasificación correcta	Falso Negativo
0 **(No pertenece a clase)**	Falso Positivo	Clasificación correcta

Los errores cometidos en la clasificación pueden corresponder a Falsos Positivos o Falsos Negativos. Los primeros hacen referencia a los individuos que no son de una clase pero son clasificados como pertenecientes a dicha clase. La fracción de Falsos Positivos aumenta cuando lo hace la sensibilidad del clasificador y decae si disminuye la especificidad. Los Falsos Negativos son aquellas observaciones que pese a pertenecer a cierta clase, son clasificados como tal. De forma complementaria al caso anterior, la cantidad de Falsos Negativos aumenta junto con la especificidad de la discriminación y disminuye cuando lo hace la sensibilidad.

Finalmente, analizando un modelo PLS-DA elaborado en base a cada clase, se pretendió observar qué variables poseen una mayor relación con qué patologías (Tabla 8). Para la construcción de los modelos siguientes, se tuvieron en cuenta todos los individuos existentes de cada clase que se encontrasen en la matriz de observaciones X.

Tabla 8. Número de individuos de cada clase empleadas en la obtención de modelo PLS para cada clase frente al resto, de un total de

Clase	Insomnio	Sanos	Narcolepsia	Epilepsia Nocturna	Mov. Per. Piernas	Trast. S. REM
Número de individuos	7	6	5	21	9	22

Para cada modelo PLS construido, se muestran las componentes consideradas junto con las bondades de ajuste y predicción del modelo, así como el gráfico de valores Observados frente a Predichos, con una matriz de confusión resumiendo la discriminación realizada en base al modelo con un umbral de 0.5.

En la construcción y validación de modelos normalmente el umbral se fija en base a la discriminación óptima de los datos empleados en la fase de entrenamiento. Posteriormente, las observaciones del conjunto de validación son proyectadas por el modelo sobre el espacio de predicciones. La situación óptima es que el umbral siga discriminando correctamente los datos de validación.

Sin embargo, debido al bajo número de individuos tal y como se explica en el apartado 3.5, en este caso se realizó la implementación del modelo únicamente con el conjunto de entrenamiento. Por tanto no cabía la posibilidad de validar un umbral fijado ad-hoc para cada clase mediante el conjunto de validación. Siguiendo un criterio conservador, para no obtener simplemente la solución más conveniente para cada clase sin ser validada posteriormente, se primó el poder comparar los resultados del modelo con los datos de entrenamiento, y se determinó un umbral común en 0.5 para evaluar las predicciones sobre el conjunto de entrenamiento, lo que constituye un criterio conservador que trata de reducir la clasificación basada únicamente en el propio conjunto de entrenamiento (optimista).

4.2.1. Conjunto con 30 individuos.

La obtención del primer modelo PLS se caracteriza por valores bajos de R^2 y de Q^2 (Figura 18). Una vez hecha la selección de características con *VIP* ≥ 1, se puede comprobar un aumento considerable en la bondad predictiva del modelo con tres y cuatro componentes (Figura 19), aún a costa de la pérdida de cierta capacidad explicativa (descenso de R^2 con un valor superior a 0.4 a un valor aproximado de 0.4 en el modelo con cuatro componentes).

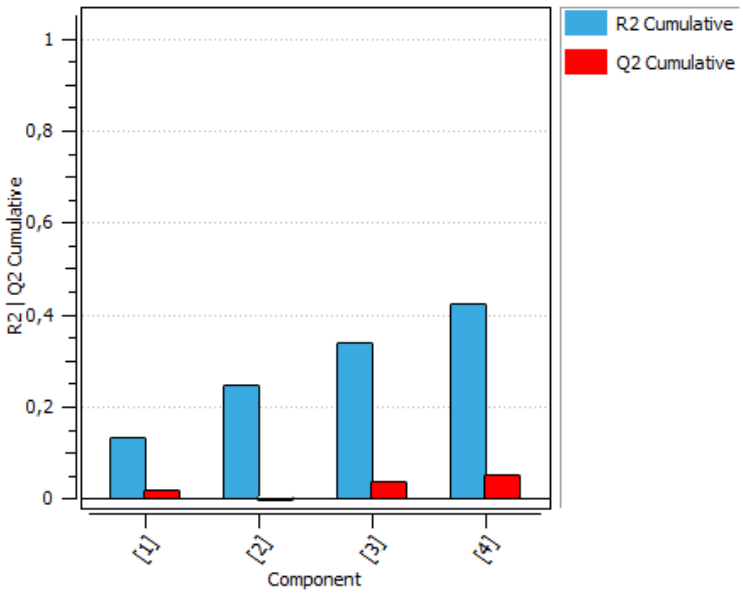

Figura 18. Suma acumulada de los coeficientes R^2 y Q^2 explicando y prediciendo la variabilidad en la matriz de observaciones X, del modelo PLS con 4 componentes para los 30 individuos.

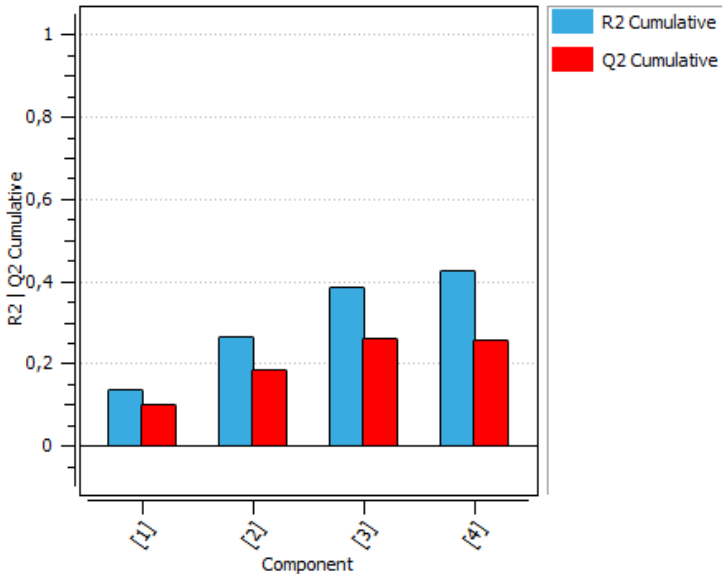

Figura 19. Suma acumulada de los coeficientes R^2 y Q^2 explicando y prediciendo la variabilidad en la matriz de observaciones X, del modelo PLS con 4 componentes para los 30 individuos, *VIPs* ≥ 1.

Puesto que para el modelo con cuatro componentes disminuye la capacidad predictiva respecto a la del modelo con tres componentes, se tendrá en cuenta finalmente el modelo con tres componentes (Figura 20). Los valores de bondad de ajuste y de predicción para este son de 0.3837 y 0.2569 respectivamente.

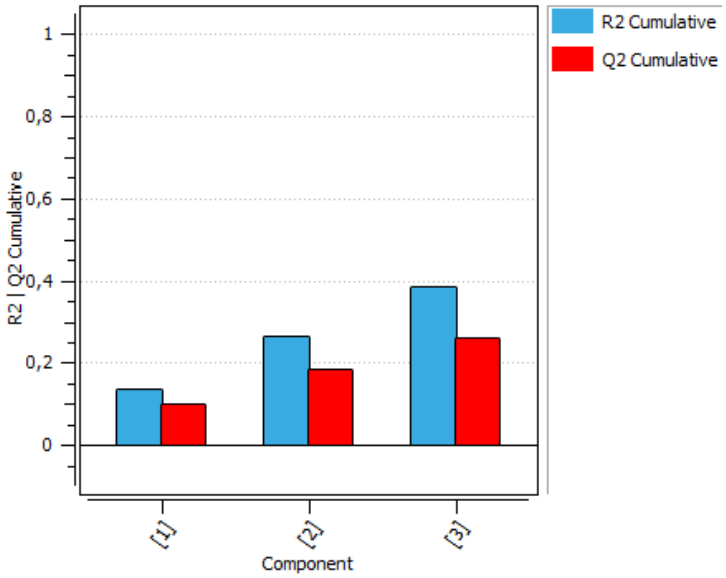

Figura 20. Capacidad explicativa y predictiva del modelo PLS con tres componentes y características con *VIP* ≥1, considerado para realizar el Análisis Discriminante entre los 30 individuos.

Para este modelo de tres componentes (Figura 20; $R^2 \approx 0.38$, $Q^2 \approx 0.27$), se construyó un modelo PLS mediante el cual se estableció posteriormente un Análisis Discriminante entre los distintos individuos del conjunto. Dicha discriminación se realizó estableciendo un umbral en un valor de 0.5, obteniendo una matriz de confusión con la que evaluar la distinción de cada clase. Los valores predichos por el modelo son graficados siguiendo un código de colores para cada clase (Tabla 9).

Tabla 9. Leyenda para clases empleada en los gráficos Observados vs. Predichos obtenidos con clasificador PLS-DA.

Clase	Color individuos predichos
Insomnio	▲
Sano	▲
Narcolepsia	▲
Epilepsia Nocturna del Lóbulo Frontal	▲
Movimientos Periódicos de Piernas	▲
Trastorno de la fase REM	▲

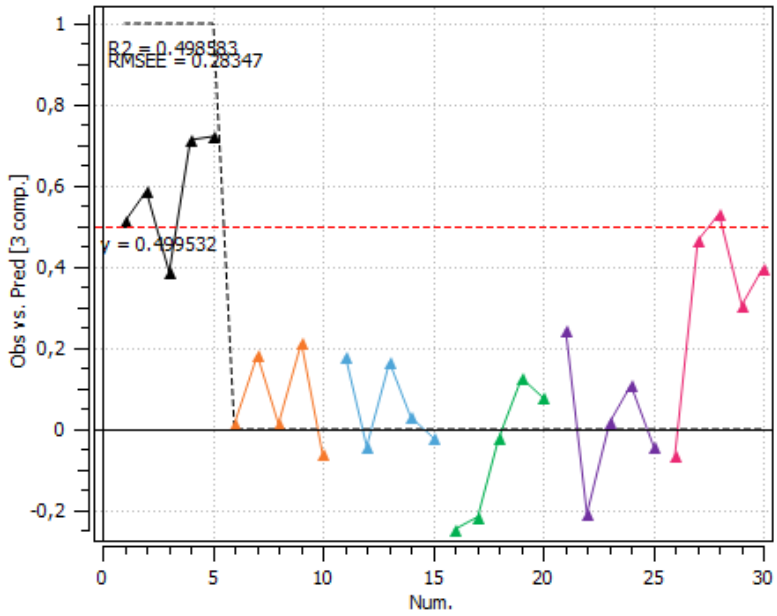

Figura 21. Valores observados frente a predichos para un conjunto de 30 individuos mediante modelo PLS. Clase 1: Insomnio.

Tabla 10. Matriz de confusión obtenida con umbral 0.5 para la discriminación de individuos con Insomnio en base a modelo PLS con 3 componentes y 6 clases.

Obs. \ Pred.	Insomnio	No-Insomnio
Insomnio	4	1
No-Insomnio	1	24

El resultado ideal en la matriz de confusión, hubiese sido una diagonal con 5 (en Insomnio observado con Insomnio predicho) y 25 (en No-Insomnio observado con No-Insomnio predicho). Sin embargo, se obtuvo un Falso Negativo (Tabla 10) para una predicción de Insomnio (Figura 21; observación 3, $\tilde{y} \approx 0.4$). También se obtuvo un Falso Positivo para una observación de la clase seis (Trastorno en la conducta de la fase REM, observación 28, $\tilde{y} \approx 0.55$). Puesto que la observación 3 no correspondía a un dato anómalo, el error de clasificación se achacó a un fallo en la predicción del modelo.

El porcentaje de aciertos fue del 80% para la clase Insomnio, con un error del 20% (Falso Negativo) y del 4% para el resto de clases (Falso Positivo). La clase con la que mayor solapamiento obtiene la clase Insomnio es la clase correspondiente a la patología Trastorno en la conducta de la fase REM.

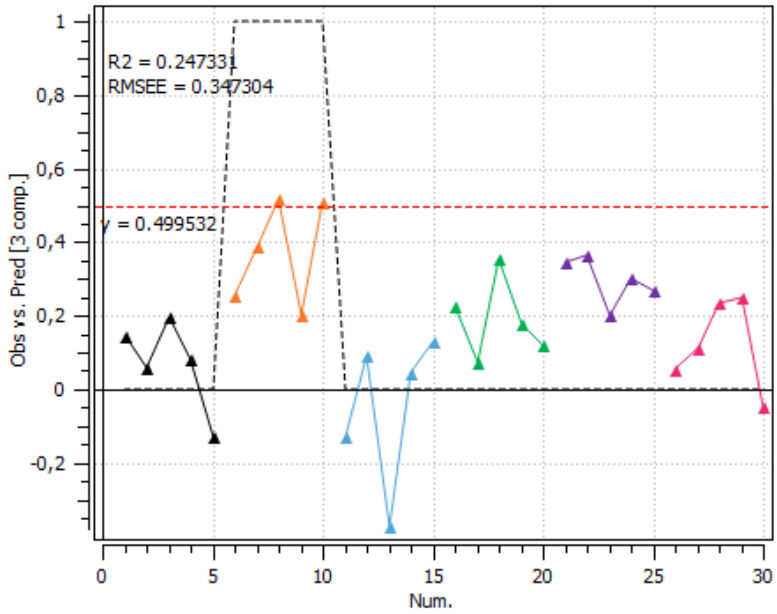

Figura 22. Valores observados frente a predichos para un conjunto de 30 individuos mediante modelo PLS. Clase: Sanos.

Tabla 11. Matriz de confusión obtenida con umbral 0.5 para la discriminación de individuos Sanos en base a modelo PLS con 3 componentes y 6 clases.

Obs. \ Pred.	Sano	No-Sano
Sano	2	3
No-Sano	0	25

En comparación con la clase Insomnio, la discriminación de los individuos Sanos tuvo un resultado con mayor error, aproximando de forma poco veraz los valores observados para dicha clase, puesto que incluso los dos individuos correctamente clasificados como Sanos, poseen valores predichos muy cercanos a 0.5. Se obtuvieron tres Falsos Negativos para las observaciones 6, 7 y 9 ($\tilde{y} \approx 0.25$, $\tilde{y} \approx 0.4$, $\tilde{y} \approx 0.2$, respectivamente). Sin embargo, no se obtuvo ningún Falso Positivo, siendo correcta la discriminación de todos los individuos no pertenecientes a la clase Sanos (Tabla 11).

El porcentaje de aciertos fue del 40% para la clase Insomnio, con un error del 60% (3 Falsos Negativos) y del 0% para el resto de clases (Falso Positivo). Destaca el solapamiento total de la case Sanos con la clase 5 (Figura 22). Esta similitud en el EEG, podría deberse a que el Movimiento Periódico de Piernas (clase 5) se manifiesta más notablemente en forma de actividad muscular, siendo de gran ayuda el análisis de la Electromiografía cuando se pretende diagnosticar este trastorno del sueño.

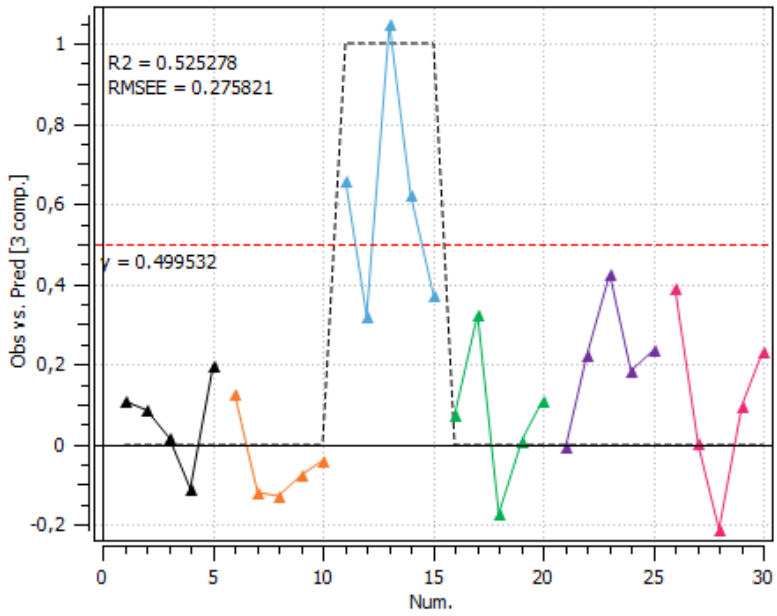

Figura 23. Valores observados frente a predichos para un conjunto de 30 individuos mediante modelo PLS. Clase: Narcolepsia.

Tabla 12. Matriz de confusión obtenida con umbral 0.5 para la discriminación de individuos con Narcolepsia en base a modelo PLS con 3 componentes y 6 clases.

Obs. \ Pred.	Narcolepsia	No-Narcolepsia
Narcolepsia	3	2
No-Narcolepsia	0	25

Para las observaciones de la clase Narcolepsia, mejoran las predicciones del modelo. Se obtuvieron dos Falsos Negativos para las observaciones 12 y 15 ($\tilde{y} \approx 0.30$, $\tilde{y} \approx 0.35$, respectivamente). De nuevo no se obtuvo ningún Falso Positivo.

El porcentaje de aciertos fue del 60% para la clase Insomnio, con un error del 40% (2 Falsos Negativos). Estas dos observaciones, coincidieron con valores predichos de las clases 4, 5 y 6 (Figura 23), correspondientes a las patologías Epilepsia Nocturna, Movimiento Periódico de Piernas y Trastorno de la Conducta del Sueño REM. Como en la clase anterior, no hubo ningún fallo en la clasificación de las observaciones no pertenecientes a la clase Narcolepsia (Tabla 12).

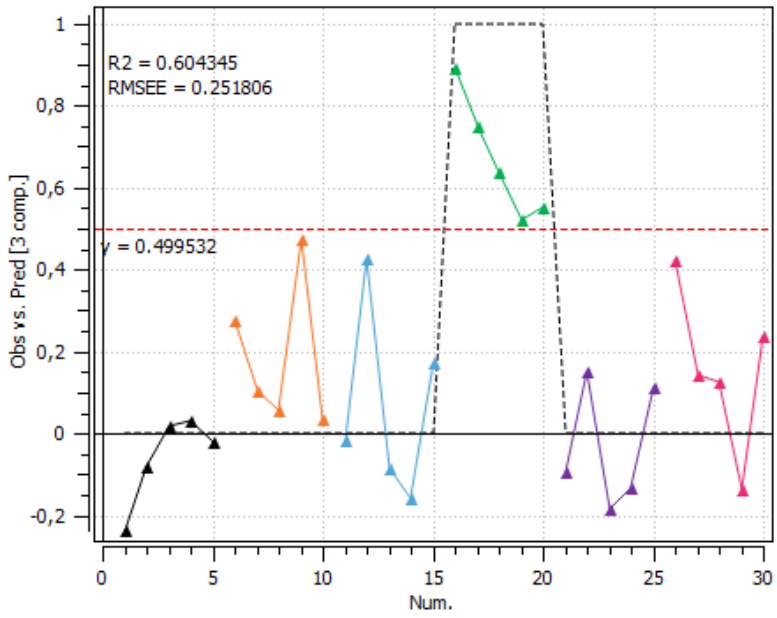

Figura 24. Valores observados frente a predichos para un conjunto de 30 individuos mediante modelo PLS. Clase: Epilepsia Nocturna del Lóbulo Frontal.

Tabla 13. Matriz de confusión obtenida con umbral 0.5 para la discriminación de individuos con Epilepsia Nocturna en base a modelo PLS con 3 componentes y 6 clases.

Pred. / Obs.	Epilepsia Nocturna	No-Epilepsia Nocturna
Epilepsia Nocturna	5	0
No-Epilepsia Nocturna	0	25

La clase Epilepsia Nocturna del Lóbulo Frontal, fue la mejor discriminada de entre las patologías consideradas. Además de obtener el valor más próximo de la predicción a la respuesta real (Figura 24; observación 16, $\tilde{y} \approx 0.80$), el porcentaje de aciertos fue del 100% tanto para la clase Insomnio, como en la discriminación de los individuos no pertenecientes a dicha clase.

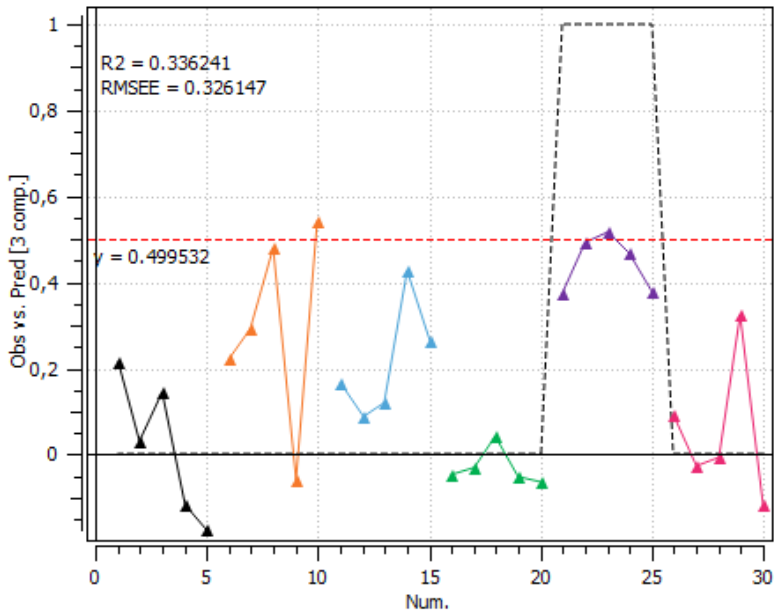

Figura 25. Valores observados frente a predichos para un conjunto de 30 individuos mediante modelo PLS. Clase: Movimientos Periódicos de Piernas.

Tabla 14. Matriz de confusión obtenida con umbral 0.5 para la discriminación de individuos con Movimientos Periódicos de Piernas en base a modelo PLS con 3 componentes y 6 clases.

Obs. \ Pred.	Mov. Per. Piernas	No-Mov. Per. Piernas
Mov. Per. Piernas	1	4
No-Mov. Per. Piernas	1	24

Tal y como se observa en la discriminación de las observaciones de la clase Sanos (Figura 22), el solapamiento entre sus individuos y los de la clase Movimiento Periódico de Piernas es total (Figura 25). Debido a ello, de los dos individuos clasificados como clase 5, uno de ellos sí corresponde a Movimiento Periódico de Piernas, mientras que otro es de la clase Sanos.

El porcentaje de aciertos fue del 20% para la clase Movimiento Periódico de Piernas, con un error del 80% (4 Falsos Negativos), mientras que sólo hubo un Falso Positivo de los 25 individuos no pertenecientes a la clase 5 (error del 4% para observaciones no pertenecientes a Movimiento Periódico de Piernas).

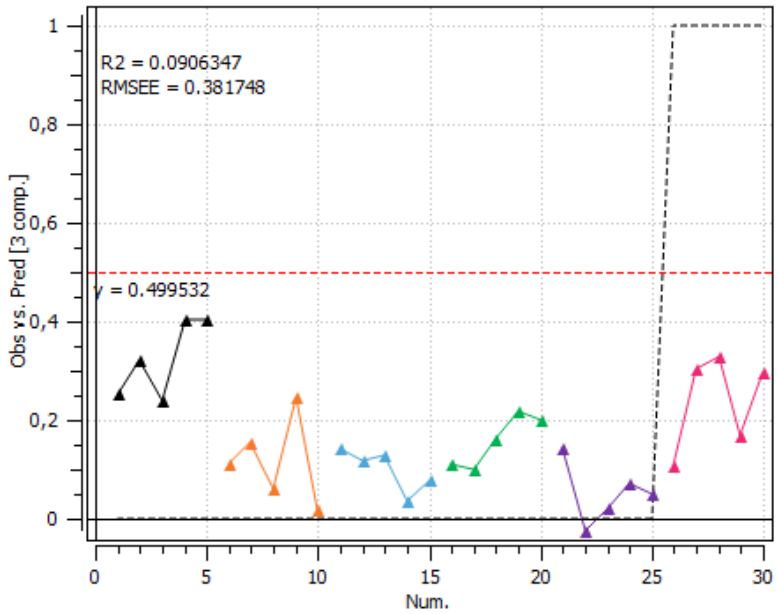

Figura 26. Valores observados frente a predichos para un conjunto de 30 individuos mediante modelo PLS. Clase: Trastorno en la Conducta del Sueño REM.

Tabla 15. Matriz de confusión obtenida con umbral 0.5 para la discriminación de individuos con Trastorno en la Conducta del Sueño REM en base a modelo PLS con 3 componentes y 6 clases.

Obs. \ Pred.	Trast. Cond. REM	No-Trast. Cond. REM
Trast. Cond. REM	0	5
No-Trast. Cond. REM	0	25

Por último, la peor discriminación obtenida fue la de individuos con Trastorno en la Conducta del Sueño REM, puesto que no se detectó a ningún individuo de esa clase como tal, aunque tampoco fueron clasificados en la clase 5, individuos de otras clases (Tabla 15).

4.2.1. Una clase frente al resto.

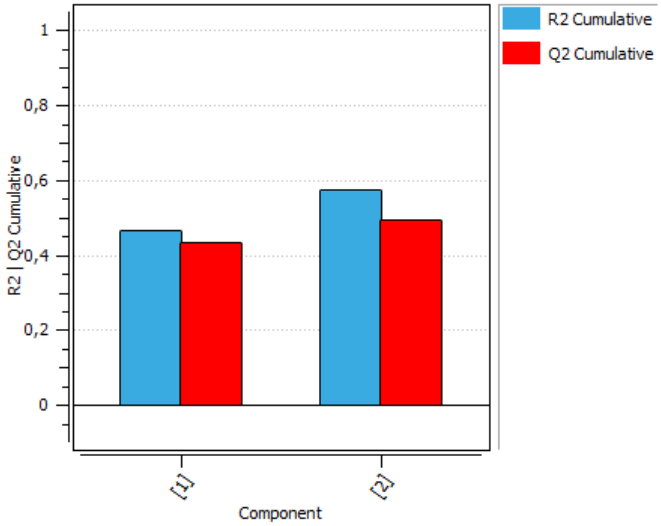

Figura 27. Suma acumulada de los coeficientes R^2 y Q^2 explicando y prediciendo la variabilidad en la matriz de observaciones X, del modelo PLS con 1 y 2 componentes para los 7/78 individuos con Insomnio, *VIPs* ≥ 1.

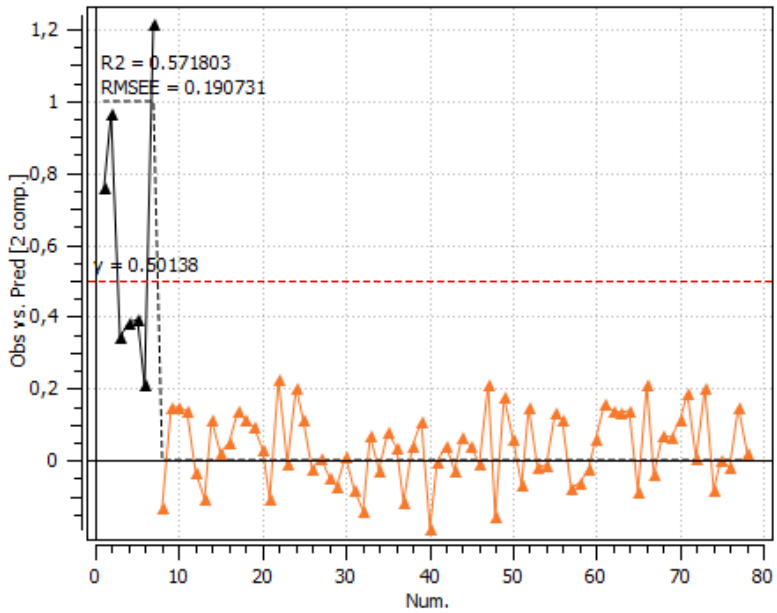

Figura 28. Valores observados frente a predichos para Insomnio (7) vs. No-Insomnio (71). Modelo PLS con 2 componentes.

Tabla 16. Matriz de confusión obtenida con umbral 0.5 para la discriminación de individuos con Insomnio en base a modelo PLS con 2 componentes.

Obs. \ **Pred.**	**Insomnio**	**No-Insomnio**
Insomnio	3	4
No-Insomnio	0	71

El modelo PLS obtenido con dos componentes (Figura 27; $R^2 \approx 0.67$, $Q^2 \approx 0.5$) mejora la capacidad explicativa y predictiva que el de tres componentes para discriminar las seis clases (Figura 20). El nuevo porcentaje de acierto de 42.85% para los individuos con Insomnio es menor que con el modelo PLS anterior (Tabla 10). Sin embargo, la discriminación de las observaciones que no pertenecen a la clase Insomnio mejora, ya que los valores predichos son más cercanos al valor observado cero (Figura 28) y en los errores cometidos en la clasificación, no se encuentra ningún Falso Positivo (Tabla 16).

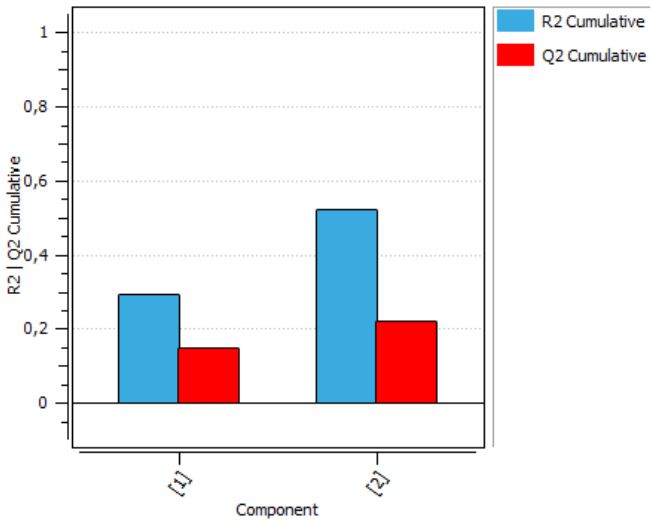

Figura 29. Suma acumulada de los coeficientes R^2 y Q^2 explicando y prediciendo la variabilidad en la matriz de observaciones X, del modelo PLS con 2 componentes para los 6/78 individuos Sanos, *VIPs* ≥ 1.

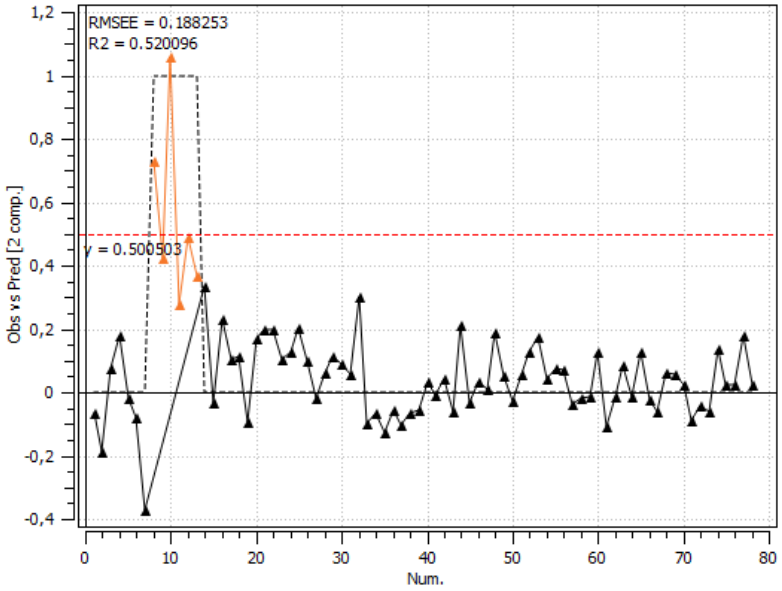

Figura 30. Valores observados frente a predichos para Sanos (6) vs. No-Sanos (72). Modelo PLS con 2 componentes.

Tabla 17. Matriz de confusión obtenida con umbral 0.5 para la discriminación de individuos Sanos en base a modelo PLS con 2 componentes.

Obs. \ Pred.	Sanos	No- Sanos
Sanos	2	4
No- Sanos	0	72

El modelo PLS empleado para discriminar individuos Sanos respecto al resto, se obtuvo con dos componentes (Figura 29; $R^2 \approx 0.55$, $Q^2 \approx 0.25$). Si bien es cierto que el porcentaje de aciertos del 50% no mejora el resultado del Análisis Discriminante con el modelo PLS previo (Tabla 11), con el nuevo modelo PLS ya no se produce el solapamiento con todas las observaciones de la clase 5 (Figura 30). De esta mejoría se desprende que el aumento del número de individuos ha jugado un papel crítico a la hora de construir el modelo PLS, haciéndose discriminaciones más veraces cuanta más información se esté teniendo en cuenta.

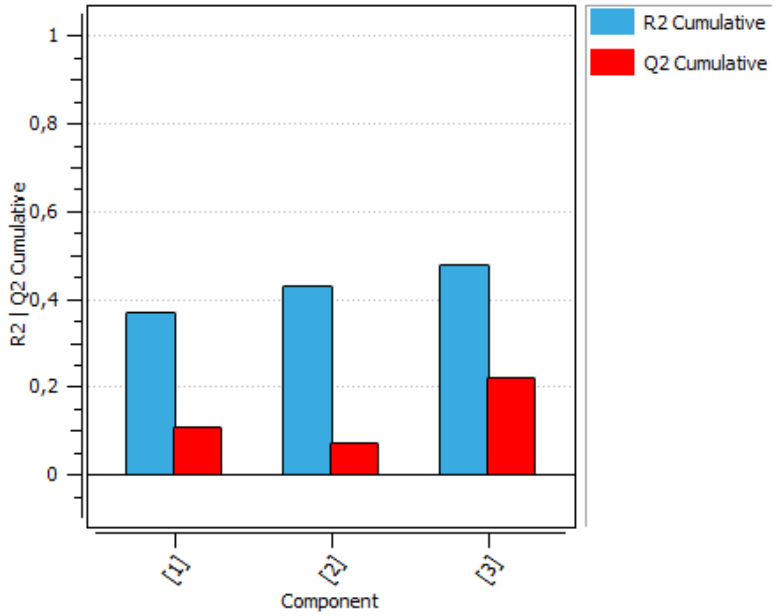

Figura 31. Suma acumulada de los coeficientes R^2 y Q^2 explicando y prediciendo la variabilidad en la matriz de observaciones X, del modelo PLS con 1, 2 y 3 componentes para los 5/78 individuos con Narcolepsia, *VIPs* ≥ 1.

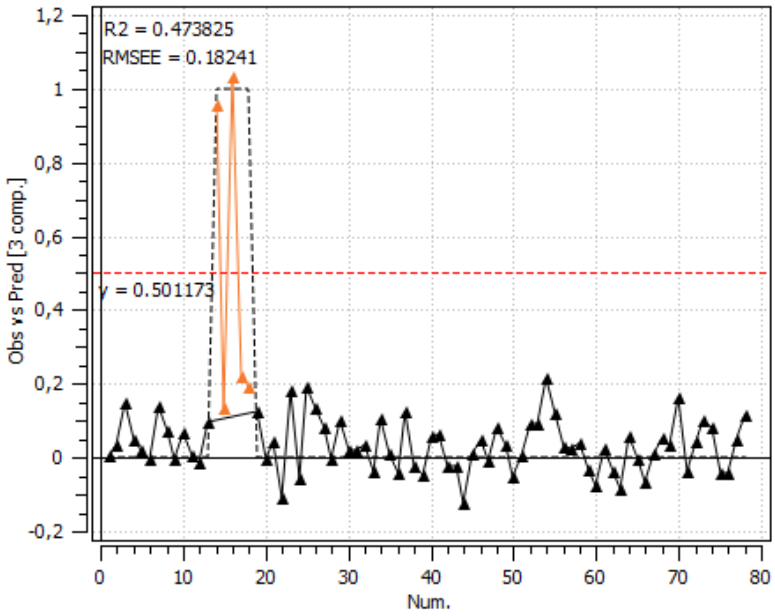

Figura 32. Valores observados frente a predichos para Narcolepsia (5) vs. No-Narcolepsia (73). Modelo PLS con 3 componentes.

Tabla 18. Matriz de confusión obtenida con umbral 0.5 para la discriminación de individuos Sanos en base a modelo PLS con 2 componentes.

Pred. / **Obs.**	**Narcolepsia**	**No-Narcolepsia**
Narcolepsia	2	3
No-Narcolepsia	0	73

Para discriminar la clase Narcolepsia de las otras clases incluidas en el conjunto, se empleó un modelo PLS con 3 componentes (Figura 31Figura 32; $R^2 \approx 0.50$, $Q^2 \approx 0.25$). El porcentaje de acierto del 40% obtenido con este modelo, no mejora el resultado obtenido con el modelo PLS realizado con todas las clases (Tabla 12). Pese a haber sido realizado con los mismo individuos, el resultado ha cambiado discriminándose mejor las clases que cuentan con más individuos, lo cual vuelve a ser una prueba de la importancia del número de individuos a la hora de generar el modelo.

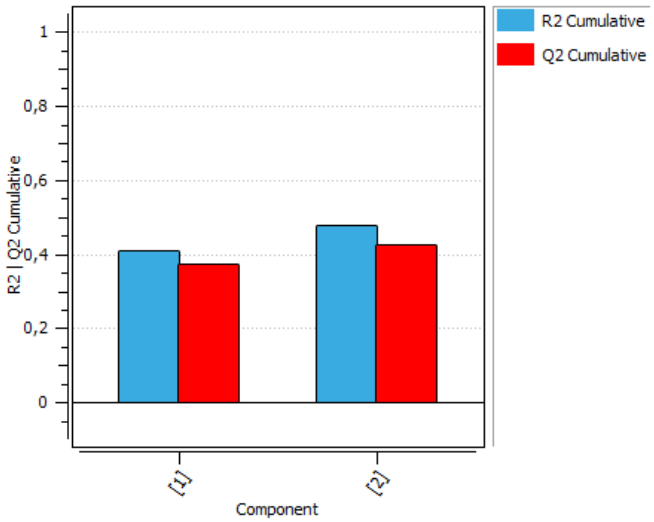

Figura 33. Suma acumulada de los coeficientes R^2 y Q^2 explicando y prediciendo la variabilidad en la matriz de observaciones X, del modelo PLS con 1 y 2 componentes para los 29/78 individuos con Epilepsia Nocturna, *VIPs* ≥ 1.

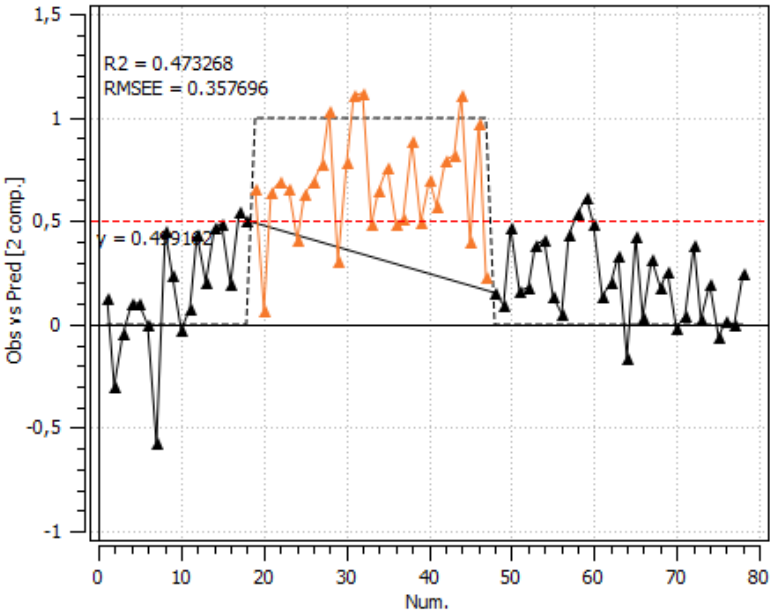

Figura 34. Valores observados frente a predichos para Epilepsia (29) vs. No-Epilepsia (49). Modelo PLS con 2 Componentes.

Tabla 19. Matriz de confusión obtenida con umbral 0.5 para la discriminación de individuos con Epilepsia Nocturna del Lóbulo Frontal en base a modelo PLS con 2 componentes.

Obs. \ Pred.	Epilepsia Nocturna	No- Epilepsia Nocturna
Epilepsia Nocturna	19	10
No- Epilepsia Nocturna	4	45

Pese a haber sido la clase mejor discriminada con el modelo PLS para las seis patologías, la clase Epilepsia Nocturna ha obtenido un peor ratio de aciertos en la clasificación de individuos de esa misma clase (65.52% frente al 100% anterior). Este hecho va unido a un aumento en la cantidad de Falsos Negativos (34.48%), así como en la de Falsos Positivos (8.16%) que antes era nula. El incremento en el número de observaciones ha introducido variabilidad en los individuos de esta clase, dificultando el ajuste del modelo y por tanto la discriminación, ya que ahora existe solape con la mayoría de clases (Figura 34).

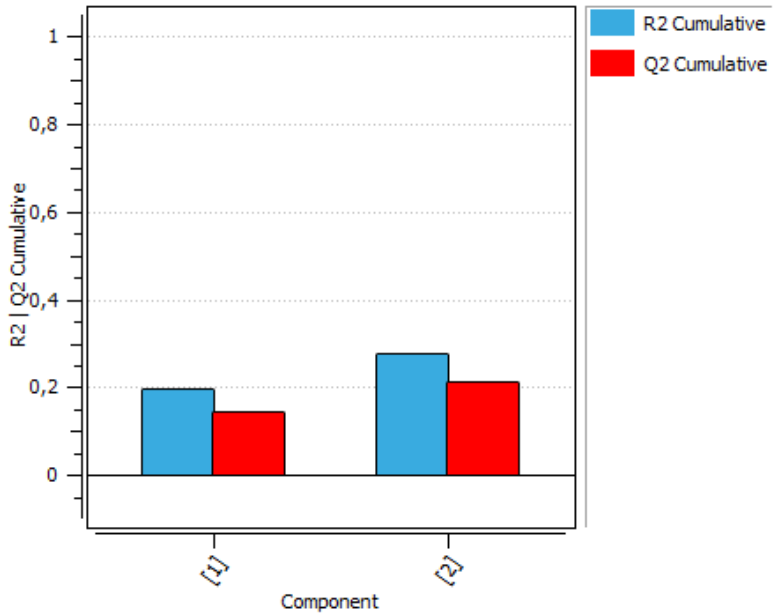

Figura 35. Suma acumulada de los coeficientes R^2 y Q^2 explicando y prediciendo la variabilidad en la matriz de observaciones X, del modelo PLS con 1 y 2 componentes para los 9/78 individuos con Movimientos Periódicos de Piernas, *VIPs* ≥ 1.

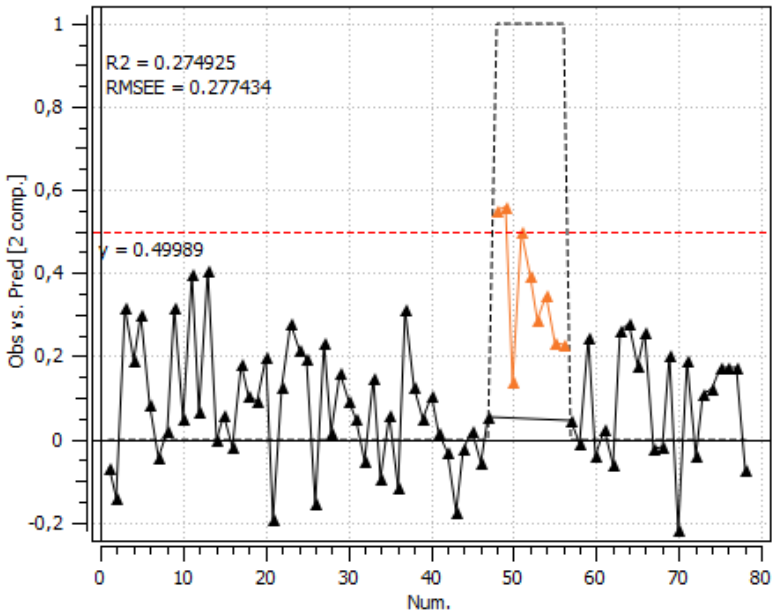

Figura 36. Valores observados frente a predichos para Movimientos Periódicos de Piernas (9) vs. No-Movimientos Periódicos de Piernas (69). Modelo PLS con 2 componentes.

Tabla 20. Matriz de confusión obtenida con umbral 0.5 para la discriminación de individuos con Movimientos Periódicos de Piernas en base a modelo PLS con 2 componentes.

Pred. / Obs.	Mov. Per. Piernas	No-Mov. Per. Piernas
Mov. Per. Piernas	2	7
No-Mov. Per. Piernas	0	69

Con un cambio similar al experimentado por los sujetos de la clase 2 (Sanos), la clase Movimiento Periódico de Piernas, ha experimentado una leve mejora en la cantidad de individuos de clase 4 clasificados como tal (28.57%), sin tener ningún Falso Positivo (Tabla 20). Sin embargo, sigue habiendo un alto porcentaje de Falsos Negativos, superando el 70%. De todos modos, una notable mejora en la discriminación es que ya no toda la case 5, se solapa por completo con la clase de individuos Sanos (Figura 36).

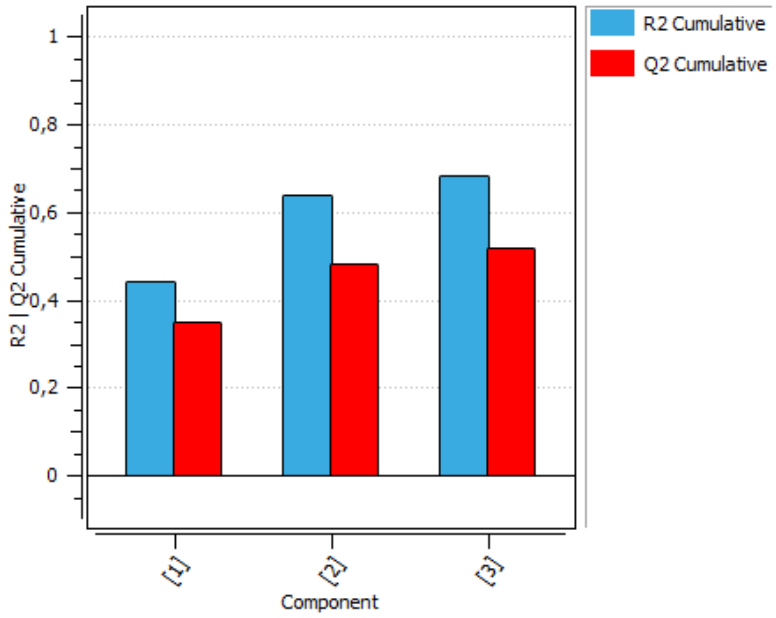

Figura 37. Suma acumulada de los coeficientes R^2 y Q^2 explicando y prediciendo la variabilidad en la matriz de observaciones X, del modelo PLS con 1, 2 y 3 componentes para los 22/78 individuos con Trastorno en la Conducta de la Fase REM, *VIPs* ≥ 1.

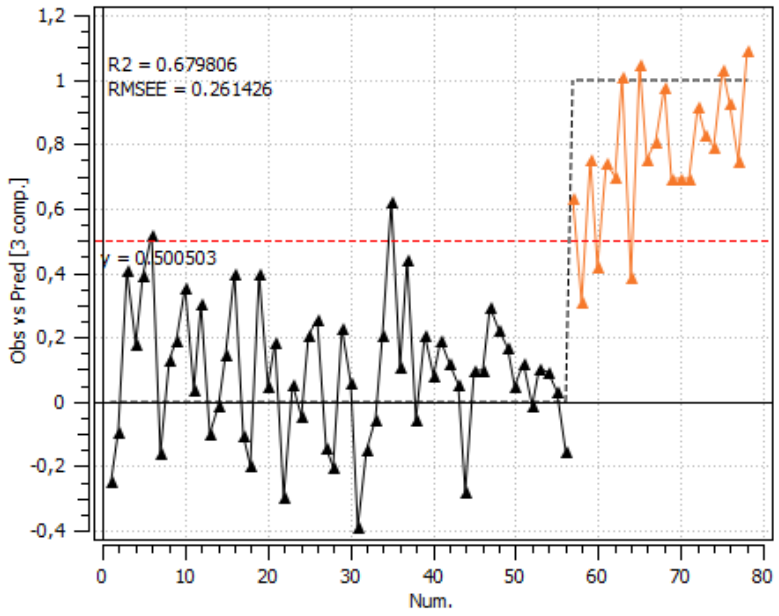

Figura 38. Valores observados frente a predichos para Trastorno en la Conducta del Sueño REM (22) vs. No-Trastorno en la Conducta del Sueño REM (56). Modelo PLS con 3 componentes.

Tabla 21. Matriz de confusión obtenida con umbral 0.5 para la discriminación de individuos con Trastorno en la Conducta del Sueño REM en base a modelo PLS con 2 componentes.

Obs. \ Pred.	Trast. Cond. REM	No-Trast. Cond. REM
Trast. Cond. REM	19	3
No-Trast. Cond. REM	2	54

En contraposición al empeoramiento de los resultados para la clase Epilepsia Nocturna al introducir más datos, con la clase Trastorno en la Conducta del Sueño REM pasa lo contrario. De no detectar ningún individuo superior al umbral de 0.5 (Tabla 15), ha pasado a tener un porcentaje de acierto del 86.36%, discriminando individuos pertenecientes a la clase 6 frente al resto como pertenecientes a dicha clase. Este aumento de sensibilidad, ha sido también a costa de perder especificidad, puesto que hay un porcentaje no nulo de Falsos Positivos (3.56%).

Finalmente, una extracción de las variables más importantes en la elaboración del modelo para cada clase, fue realizada con el fin de poder comparar qué características habrían sido útiles en mayor o menor medida, y si habría alguna relación entre estas (Tabla 7) y el conocimiento *a priori* que podría haberse aplicado a la hora de seleccionar características de interés. El orden de las variables en cada columna, es descendente según el valor R^2 estimado para cada variable por la aproximación (Tabla 8).

En general, puede decirse que las variables obtenidas se distribuyen en unos intervalos relativamente concretos (Tabla 4). Los parámetros se encuentran distribuidos según el nivel de descomposición, parámetro calculado, electrodo y si es un *score* o *loading* de las dos Componentes Principales empleadas para la construcción de la matriz.

Tabla 22. Cinco variables más importantes (mayor R^2) en cada clase.

INSOMNIO	SANOS	NARCOLEPSIA	EPILEPSIA NOCTURNA DEL LÓBULO FRONTAL	MOVIMIENTO PERIÓDICO DE PIERNAS	TRASTORNO EN LA CONDUCTA DEL SUEÑO REM
33.0	317.0	515.0	392.0	383.0	215.0
103.0	327.0	265.0	382.0	403.0	217.0
43.0	318.0	505.0	402.0	393.0	287.0
7.0	377.0	545.0	132.0	392.0	77.0
53.0	328.0	375.0	412.0	382.0	293.0

La información a la que hace referencia la variable más determinante en cada caso, prueba la conservación de la información contenida en el registro EEG referente a la actividad cerebral para ciertos casos, a lo largo del procesado de los datos (Tabla 9). Concretamente, para los registros de Insomnio, Epilepsia Nocturna y Movimiento Periódico de Piernas. El resto de clases señalan como característica más importante algún parámetro sobre bandas de frecuencia superiores, que no están vinculadas a actividad cerebral (Tabla 1). Este resultado podría ser un indicio de la necesidad de filtrar previamente los niveles de descomposición que introducir en la matriz de observaciones, estableciendo una base de conocimiento *a priori* en el análisis.

Tabla 23. Características más relevantes según modelo PLS-DA en la aproximación a cada clase.

INSOMNIO	SANOS	NARCOLEPSIA	EPILEPSIA NOCTURNA DEL LÓBULO FRONTAL	MOVIMIENTO PERIÓDICO DE PIERNAS	TRASTORNO EN LA CONDUCTA DEL SUEÑO REM
Desviación típica del 7º nivel de compresión (2 Hz), electrodo F4-C4.	Mínimo del 1er nivel de compresión (128 Hz), electrodo Fp2-F4.	Desviación típica del 1er nivel de compresión (128 Hz), electrodo C4-P4.	Mínimo del 6º nivel de compresión (4 Hz), electrodo F4-C4	Desviación típica del 7º nivel de compresión (2 Hz), electrodo F4-C4.	Entropía del 2º nivel de compresión (64 Hz), electrodo C4-P4.

5. CONCLUSIONES

La cuantificación y estandarización de los procedimientos médicos para poder evaluarlos y así mejorarlos, es un tema candente que atañe tanto al sector médico, como al campo tecnológico que está emergiendo en el área de la salud. Patologías recurrentes en la sociedad, que involucran a otros componentes de la fisiología humana y que se asocian a enfermedades de carácter crónico, son todavía estudiadas y diagnosticadas de forma no cuantitativa. Este es el caso de los trastornos en la conducta del sueño, y las consecuencias fundamentales son baja repitibilidad y reproducibilidad en los resultados, menor profundidad en el estudio realizado y por tanto, una peor calidad en la gestión de los afectados y de los recursos de los Sistemas Sanitarios.

En este trabajo se ha propuesto una metodología centrada en la compresión de la información relevante y el análisis de la información latente en las bases de datos disponibles, sin apenas introducción de conocimiento *a priori*. Así pues, la compresión de grandes volúmenes de datos, como es el caso de las Polisomnografías, ha conseguido hacerse de forma exitosa gracias al empleo de la Transformada *Wavelet* Discreta con la *Wavelet Daubechies* 4 y nueve niveles de descomposición. Además, el Análisis Multivariante realizado con Análisis de Componentes Principales, ha permitido resumir en nuevas variables generadas por el algoritmo, las variables principalmente responsables de la variabilidad en la matriz de observaciones.

Sin embargo, la clasificación de los individuos en sus respectivas clases ha tenido resultados poco concluyentes. En primer lugar, la calidad de los datos de partida es un punto clave para el correcto tratamiento de los mismos. La ausencia de registros en algunos pacientes y el desorden en el almacenamiento de la información en las cabeceras, conllevó la eliminación de varios registros. Dicha eliminación, además de suponer pérdida de información como individuos de los que extraer información para ser clasificados, también ha supuesto un déficit global en el número de individuos.

Esta ausencia de individuos ha influido muy notablemente en los modelos construidos. En primer lugar, la fase de implementación de los modelos queda restringida a únicamente la primera etapa de entrenamiento, ya que por ejemplo de las clases Insomnio, sanos y Narcolepsia había únicamente 7, 6 y 5 individuos respectivamente, teniendo en cuenta que las clases Bruxismo y Apnea fueron directamente desconsideradas porque tras el cribado inicial de pacientes, sólo quedó un individuo de cada clase.

El tamaño insuficientemente grande de la cantidad de pacientes, ha repercutido especialmente en etapas como la determinación de un umbral para el Análisis Discriminante basado el PLS, así como en la distinción de *outliers* o datos anómalos, puesto que eliminarlos puede suponer la pérdida de demasiada información y además, se desconoce qué es y qué no normal dentro de una clase si la cantidad de individuos no es lo suficientemente grande.

No obstante, los resultados tanto con KNN como con el modelo PLS-DA dan una aproximación considerable para la clase Epilepsia Nocturna (Tabla 13), aunque los resultados empeoraron al realizar un PLS de esa misma clase frente al resto (Tabla 19). Esta patología cuenta ya con múltiples referencias en el ámbito de reconocimiento de patrones, siendo un habitual caso de estudio en el Electroencefalograma. Por otro lado, las confusiones en la clasificación también pueden arrojar información sobre posibles relaciones entre patologías, como es el caso entre el Trastorno de la Conducta del Sueño REM y la Narcolepsia, analizados en los resultados del clasificador KNN.

Los resultados obtenidos con el PLS-DA realizando la comparación entre una clase y el resto, sugieren que en la estructura con la información sobre las características extraídas, hay información con cierta relevancia a la hora de discriminar una patología del resto. Sin embargo, los resultados han requerido de un filtrado previo en el número de variables de la matriz de observaciones. Esto sugiere, por un lado, que son necesarios más individuos para poder afirmar o sugerir con mayor seguridad, que las diferencias obtenidas entre las distintas clases son suficientes como para considerarlas discriminadas.

Por tanto, empleando la aproximación realizada en este trabajo para el análisis de señales cerebrales mediante *wavelets* y variables latentes, se ha obtenido un resultado alentador que invita a considerar que en el EEG hay información suficiente como para establecer una diferenciación entre clases. Por otro lado, también es una demostración del empleo de técnicas ingenieriles para el tratamiento de señales biomédicas y la problemática asociada que surge al trabajar con datos reales. Así pues, se han tratado cuestiones como la fiabilidad de la Base de Datos, la compresión de elevados volúmenes de información o la influencia del tamaño muestral en la extracción de conclusiones e información tras el análisis. Además, se han corroborado en cierto modo hechos conocidos (buena diferenciación de la epilepsia), y se han observado relaciones que pueden ser fuente o argumento para generar nuevo conocimiento en el ámbito médico (relación entre Trastorno del Sueño REM y Narcolepsia).

Por otro lado, la incertidumbre en algunos pasos del proceso, como qué características extraer o el preprocesado previo a la compresión con PCA y la clasificación, puede estar generando información no útil para establecer una clasificación. Otros parámetros del proceso, como la elección de ciertos electrodos para la detección de determinadas patologías, o los niveles exactos de descomposición *wavelet* donde se encontrase la información relevante, podrían estar detrás de la generación de ruido dentro de la matriz de observaciones. Estas incertidumbres podrían explicar la mejora considerable en la aproximación del modelo PLS-DA una vez se han eliminado las variables con importancia medida a través del VIP menor que 1.

A modo de líneas futuras, con el ánimo de suplir las carencias que hayan podido estar tras resultados poco concluyentes, se plantean nuevos escenarios. El primero y más obvio, es el crecimiento del conjunto de individuos con el que trabajar, siendo del orden mínimo de cientos en estudios similares, para así poder obtener información de forma robusta y más fiable. Un siguiente paso sería la aplicación de la misma metodología o similar, al resto de señales obtenidas en la Polisomnografía. Patologías como el Movimiento Periódico de Piernas, con incertidumbre en su clasificación por solapamiento con otras clases, podrían ser correctamente clasificadas con un segundo paso que incluyese el procesado de otro tipo de señales. Estas también son registradas a lo largo de toda la noche de sueño. Si bien pueden ser más amigables para la interpretación visual, de nuevo un análisis riguroso pasa por la cuantificación y empleo de herramientas matemáticas y estadísticas como las empleadas en este trabajo.

BIBLIOGRAFÍA

[1] C. A. Kushida, M. R. Littner, T. Morgenthaler, C. A. Alessi, D. Bailey y J. Coleman, «Practice Parameters for the Indications for Polysomnography and Related Procedures: An Update for 2005.,» *SLEEP,* vol. 28, nº 4, pp. 499 - 521, 2005.

[2] S. Tong y N. V. Thankor, Engineering in Medicine & Biology-Quantitative EEG Analysis Methods and Clinical Applications, Boston/London: Artech House, 2009.

[3] *Tema 4. Ejemplos de procesado de señales Biomédicas. Sistema nervioso.,* Valencia, Curso 2014/2015.

[4] American Sleep Association, «American Sleep Association (ASA),» ASA, 2016. [En línea]. Available: https://www.sleepassociation.org/patients-general-public/what-is-sleep/. [Último acceso: Junio 2016].

[5] AAS, «American Sleep Association,» [En línea]. Available: http://www.aasmnet.org/resources/pdf/pressroom/Telemedicine-position.pdf. [Último acceso: Junio 2016].

[6] Sociedad Española del Sueño, «ses.org,» 2014. [En línea]. Available: http://www.ses.org.es/docs/Comunicado_DMS_100314.pdf.

[7] Anxiety and Depression Association of America (ADAA), «ADAA. Understanding the facts. Related Ilnesses. Sleep Disorders.,» ADAA, 2016. [En línea]. Available: http://www.adaa.org/understanding-anxiety/related-illnesses/sleep-disorders. [Último acceso: Junio 2016].

[8] N. A. Collop, W. M. Anderson, B. Boehlecke, D. Claman, R. Goldberg y D. J. Gottlieb, «Considerations of Portable Monitoring Equipment.,» Evolve Sleep, AASM., Darien, IL, 2015.

[9] J. Singh, M. S. Badr, W. Diebert, L. Epstein, D. Hwang y V. Karres, «American Academy of Sleep Medicine (AASM) Position Paper for the Use of Telemedicine for the Diagnosis and Treatment of Sleep Disorders.,» *Journal of Clinical Sleep Medicine,* vol. 11, nº 10, pp. 1187 - 1198, 2015.

[10] Hospital Clínico San Carlos y Fundación BBVA, «El Electrocardiograma. Libro de la salud cardiovascular.,» FBBVA, Madrid, 2007.

[11] M. G. Terzano, L. Parrino, A. Sherieri, R. Chervin, S. Chokroverty y C. Guilleminault, «Atlas, rules, and recording techniques for the scoring of cyclic alternating pattern (CAP) in human sleep.,» *Sleep Medicine, Elsevier Science,* vol. 2, nº 6, pp. 537 - 553, 2001.

[12] MathWorks, «MathWorks,» MathWorks Inc., 2011. [En línea]. Available: http://es.mathworks.com/help/matlab/ref/matfile.html. [Último acceso: 02 2016].

[13] O. Faust, U. R. Acharya, H. Adeli y A. Adeli, «Wavelet-based EEG processing for computer-aided seizure detection and epilepsy diagnosis.,» *Seizure - European Journal of Epilepsy,* vol. 26, pp. 56-64, 2015.

[14] A. Folch, F. Arteaga y A. J. Ferrer, «Missing Data Imputation Toolbox for MATLAB,» *Chemometrics and Intelligent Laboratory Systems,* vol. 154, pp. 93 - 100, 2016.

[15] A. Subasi y M. I. Gursoy, «EEG signal classification using PCA, ICA, LDA and support vector machines,» *Expert Systems with Applications,* vol. 37, nº 12, p. 8659–8666, December 2010.

[16] V. Bono, S. Das, W. Jamal y K. Maharatna, «Hybrid wavelet and EMD/ICA approach for artifact suppression in pervasive EEG,» *Journal of Neuroscience Methods,* vol. 267, pp. 89-107, 19 Abril 2016.

[17] R. Bro y A. K. Smilde, «Principal component analysis,» *Analythical Methods,* vol. 6, pp. 2812-2831, 2014.

[18] J. Shlens, «A Tutorial on Principal Component Analysis,» ArXiv e-prints, Mountain View, CA, 2014.

[19] A. Khalil, M. A. Wright, M. C. Walker y S. H. Eriksson, «Loss of rapid eye movement sleep atonia in patients with REM sleep behavioral disorder, narcolepsy, and isolated loss of REM atonia.,» *Journal of Clinical Sleep Medicine,* vol. 9, nº 10, pp. 1039 - 1048, 2013.

[20] Instituto del Sueño, «Instituto del Sueño,» Instituto del Sueño , 2015. [En línea]. Available: http://www.iis.es/que-es-como-se-produce-el-sueno-fases-cuantas-horasdormir/. [Último acceso: 3 Junio 2016].

[21] G. Strang, «Appendix 1, Wavelets,» *American Scientist,* vol. 82, pp. 250-255, 1994.

[22] K. H. Esbensen y P. Geladi, «Principal Component Analysis: Concept, Geometrical Interpretation, Mathematical Backgroung, Algorithms, History, Practice,» de *Comprehensive Chemometrics*, Elsevier B.V, 2009, pp. 211-226.

[23] H. Abdi y L. J. Williams, «Principal Component Analysis,» *John Willey & Sons,* vol. 2, pp. 433 - 459, 2010.

[24] P. Geladi y B. R. Kowalski, «Partial Least-Squares Regression: A tutorial.,» *Analytica Chimica Acta,* vol. 185, pp. 1 - 17, 1986.

[25] M. Akin, «Comparison ofWavelet Transform and FFT Methods in the Analysis of EEG Signals,» *Journal of Medical Systems,* vol. 26, nº 3, pp. 241 - 247, 2002.

[26] P. S. Addison, «Wavelet transforms and the ECG: a review,» *Physiological Measurement,* vol. 26, pp. 155-199, 2005.

Anejos

ÍNDICE DE LOS ANEJOS

1. ANEJO 1

Programa ejecutado en Matlab para la carga de datos en formato .edf en una sola variable (Ejemplo con dos pacientes).

```
DATAe=matfile('DATAe','Writable',true); % Se genera la variable
matFile que
% contendrá todos los registros
% Los registros de cada paciente han sido separados previamente en
datos
% numéricos y cabecera mediante la función:
        % [hdr, record] = edfread(fname, varargin)
% Los pacientes están en ficheros diferentes
pa=matfile('brux01.mat','Writable',true);% Se abre el fichero con el
% registro de un paciente
DATAe.pa001=pa.brux1(1:4,:);% Se guardan las filas que corresponden a
los
% electrodos de interés (en este caso de la 1 a la 4, en una variable
de
% DATAe con nombre pa###, siendo ### el orden del paciente dentro de
DATAe
clear pa % Se borra la variable que almacenaba el registro del
paciente,
% para ahorrar espacio
                                        […]
% Dependiendo de cada registro, se escogen las filas del fichero
relativas
% a los electrodos de interés (en este otro caso de la 2 a la 5).
pa=matfile('ins03.mat','Writable',true);
DATAe.pa005=pa.ins3(2:5,:);
clear pa
```

2. ANEJO 2

Programa ejecutado en Matlab para la transformación al espacio *Wavelet* mediante el uso de la Transformada *Wavelet* Discreta de Daubechies 4 (db4) y cálculo de características descriptivas de dichos coeficientes.

```
%% Fragmentación en epochs de la señal.
function [W_EPOCH9,EPOCHS]
=obt_param_epochs_9levelNEW(DATA,paciente,electrodo,EPOCHS,COND)
V = who(DATA); % Lista de variables en DATA
reg = DATA.(genvarname(V{paciente})); % Extrae registro del paciente
de DATA
signal = reg(electrodo,:);  % Extrae electrodo de interés
fs = 512;% Frecuencia de muestreo del registro
epoch = 30*fs; % Ancho de cada epoch (muestras)
ini = (1:epoch:length(signal)); % Vector con posiciones de inicio de
epochs
```

```
% Podría ocurrir que la longitud del registro, no fuese múltiplo de la
% longitud de la epoch, y al calcular
multiplo = length(signal)/epoch;
if not(mod(multiplo,1)) == 1
    N_epochs = length(ini);
else
    N_epochs = length(ini)-1;
end
% Inicializa variables para más rapidez
media = zeros(N_epochs,10);
desv_tipic= zeros(N_epochs,10);
maximo = zeros(N_epochs,10);
minimo = zeros(N_epochs,10);
energia = zeros(N_epochs,10);
asimetria = zeros(N_epochs,10);
entropia = zeros(N_epochs,10);
for i=1:N_epochs % Para todas las epoch
    epN = signal(ini(i):(ini(i)+epoch-1)); % Epoch enventanada
    [coeff,long] = wavedec(epN,9,'db4'); % Calcula los coeficientes de
la TW de la epoch
    posicion = [long(1);sum(long(1:2));sum(long(1:3));sum(long(1:4));

sum(long(1:5));sum(long(1:6));sum(long(1:7));sum(long(1:8));
                sum(long(1:9));sum(long(1:10))];
    L9 = coeff(1:posicion(1));
    H9 = coeff(posicion(1)+1:posicion(2));
    H8 = coeff(posicion(2)+1:posicion(3));
    H7 = coeff(posicion(3)+1:posicion(4));
    H6 = coeff(posicion(4)+1:posicion(5));
    H5 = coeff(posicion(5)+1:posicion(6));
    H4 = coeff(posicion(6)+1:posicion(7));
    H3 = coeff(posicion(7)+1:posicion(8));
    H2 = coeff(posicion(8)+1:posicion(9));
    H1 = coeff(posicion(9)+1:posicion(10));
    W_EPOCH9{i} = {H1,H2,H3,H4,H5,H6,H7,H8,H9,L9}; % Cell con series
de coeficientes
    % Una fila / epoch
    % Una columna / nivel de descomposición
end
%% Obtención de parámetros de interés: media, varianza, asimetría y
kurtosis
% Para cada nivel de compresión de cada una de las epoch:
for i=1:N_epochs
    for j = 1:10
    media(i,j) = mean(W_EPOCH9{i}{1,j}); % Media de las series de la
epoch
    desv_tipic(i,j) = std(W_EPOCH9{i}{1,j}); % Varianza de las series
de la epoch
    maximo(i,j) = max(W_EPOCH9{i}{1,j}); % Maximo de las series de la
epoch
    minimo(i,j) = min(W_EPOCH9{i}{1,j}); % Mínimo de las series de la
epoch
    energia(i,j) = sum((W_EPOCH9{i}{1,j}).^2); % Potencia de cada
subbanda de la epoch
    asimetria(i,j) = skewness(W_EPOCH9{i}{1,j}); % Asimetría de los
coeficientes
    entropia(i,j) = entropy(W_EPOCH9{i}{1,j}); % Entropía de los
coeficientes
    end
end
```

```
% Se almacenan los parámetros para cada electrodo de cada paciente.
EPOCHS{paciente,electrodo} =
[media,desv_tipic,maximo,minimo,energia,asimetria,entropia];
% Están en matrices de cada parámetro: epoch x nivel de compresión
end
```

3. ANEJO 3

Función Matlab empleada para el preprocesado de los datos, obteniendo el resultado del Autoescalado así como del Escalado por bloques, para una estructura de entrada formada por tantos cell arrays como individuos, representados por matrices de tantas observaciones como *epoch*s y características como parámetros calculados previamente.

```
%% Preprocesado para datos

function [X_autoescalada,X_escalada_bloques] =
PRIMER_preprocesado(EEPP)

%% Forma 1. Autoescalado normal
% Número de epochs = filas de cada una de las matrices del conjunto
input (EEPP)
for paciente = 1:size(EEPP,1)
    N_epochs(paciente,1) = size(EEPP{paciente,1},1);
        X_cent{paciente,1} = EEPP{paciente}-
(ones(N_epochs(paciente),1)*mean(EEPP{paciente}));
%       CENTRADO: Resta la media de cada una de las columnas (=
parámetro) a los
%       elementos de esa misma columna
        dt=std(EEPP{paciente});
%       ESCALADO: Calcula la desviación típica de ese parámetro y
divide por ella
        INV_DESV_TIP = diag(1./dt);
        X_autoescalada{paciente,1} = X_cent{paciente}*INV_DESV_TIP; %
Corpus con el preprocesado simple
end
%%
%% Forma 2. Escalado por bloques (necesita la matriz centrada de la
Forma 1).
for paciente =1:size(EEPP,1)
    iter = (size(X_cent{paciente,1},2))/10; % Número de bloques
(=parámetros)
    for i = 1:iter;
        BLOQUE = X_cent{paciente}(:,(10*i-9):10*i); % Cada bloque
equivale a
    % la información de un parámetro para todos los niveles de
compresión
    % Se reorganiza todo el bloque en un sólo vector columna B del que
se
    % calcula la desviación típica (sd)
        B = reshape(BLOQUE,size(BLOQUE,1)*size(BLOQUE,2),1);
        sd = std(B);
    % Se almacena en la nueva matriz DT2 la división del bloque entre
su
```

```
    % varianza
        Bloq_Escalado{i}= (1/sd)*BLOQUE;
        BLOQUE=[];B=[];sd=[];
    end
    X_escalada_bloques{paciente,1} = [Bloq_Escalado{1:iter}];
end
end
```

4. ANEJO 4

Código empleado para la extracción de Componentes Principales aplicando PCA en cada paciente, pasando así de la estructura *cell array*, a una matriz con tantas observaciones como individuos y características como parámetros de los *scores,* junto con los *loadings,* variabilidades explicadas y sin explicar, y la información personal de cada paciente (duración de registro, edad y género).

```
function [MATRIX_primerafase,analisis_PCA] =
matrizPCA7_2(EEPP,X_primer_prepro,COND)
for i = 1:size(X_primer_prepro,1) % Para cada paciente
   [loadings{i}, scores{i}, latent{i}, tsq{i}, expl{i}] =
pca(X_primer_prepro{i});
end
analisis_PCA = {loadings,scores,latent,tsq,expl};
% Estructura de la nueva matriz
%% Guarda los parámetros de los scores de las 2 primeras componentes
for i = 1:size(scores,2)
    T{i,1} = scores{i}(:,1:3);
end
clear i
for i = 1:size(T,1)
        M(i,1:3) = (median(T{i},1));
        DT(i,1:3) = (std(T{i},1));
        A(i,1:3) = (skewness(T{i},1));
        Mi(i,1:3) = (min(T{i}));
        Ma(i,1:3) = (max(T{i}));
        Pow(i,1:3) =
[sum(T{i}(:,1).^2),sum(T{i}(:,2).^2),sum(T{i}(:,3).^2)];
        S(i,1:3) =
[(entropy(T{i}(:,1))),(entropy(T{i}(:,2))),(entropy(T{i}(:,3)))];
end
parametros_T = [M,DT,A,Mi,Ma,Pow,S];
clear i M DT A Mi Ma Pow S
%% Los loadings de 2 PCs
for i = 1:size(loadings,2)
        pesos_P(i,:) = [loadings{1,i}(:,1)'];
end
pesos_P;
clear i
%% Calcula la variabilidad explicada por ellas, y la que deja fuera
for i=1:length(expl)
        explicada(i) = sum(expl{i}(1:3));
        no_expl(i) = 100-explicada(i);
end
```

```
    clear i
variabilidad = [explicada',no_expl'];
%% Se añade la información de tiempo, edad y sexo
    % Tiempo en horas del registro.
for paciente = 1:size(EEPP,1)
        N_epochs(paciente,1) = size(EEPP{paciente,1},1);
end
    TIEMPO = (N_epochs*30/3600);
    % Edad del sujeto
    EDAD = COND{3};
    % Género del sujeto almacenado como Variables dummy
    GENERO = COND{2};
for i=1:size(EEPP,1)
    if GENERO{i}=='M'
            sexo(i)=0; % Hombre (Male) = 0
    else sexo(i)=1; % Mujer = 1
    end
end
sexo=sexo';
persona = [TIEMPO,EDAD, sexo];
MATRIX_primerafase = [parametros_T,pesos_P,variabilidad,persona];
end
```

Presupuesto

1. NECESIDAD DEL PRESUPUESTO.

Como en todo proyecto, la realización del trabajo descrito en la Memoria conllevó el uso de materiales, así como otros factores que deben considerarse a la hora de estimar el presupuesto necesario para realizar un proyecto como el presente.

2. ESTRUCTURACIÓN DEL PRESUPUESTO.

Las necesidades a tener en cuenta para elaborar el presupuesto se centran en la realización de la metodología descrita en el Capítulo 3. Se han tenido en cuenta dos grandes subunidades de material necesarias dentro de la obtención de resultados: hardware y software. Además habrá que considerar la labor realizada por parte del personal implicado.

Nº. Orden	Ud. Descripción	Medición	Precio (€/u. medición)	Importe (€)
Cap. 3. *MATERIALES Y MÉTODOS.*	UD. APLICACIÓN DE METODOLOGÍA PARA OBTENCIÓN DE RESULTADOS			
	UD. HARDWARE PARA ALMACENAMIENTO DE DATOS Y CÁLCULOS			
	Ordenador con procesador i7.	1 U.	635	635
	Memoria externa con 200 GB.	1 U.	74,24	74,24
	UD. SOFTWARE PARA COMPUTACIÓN DE CÁLCULOS			
	Software de cálculo Matlab 2013 – 2015.	1 U.	35	35
3.1. *Estructuración de los datos*.	Software ofimático Microsoft Office 2013.	1 U.	68.10	68.10
3.6. *Modelos de clasificación*.	Software para análisis estadístico de los datos.	1 U.	214.66	214.66
	H. INGENIERO JUNIOR	300 H.	20	6000
	H. INGENIERO SENIOR 1	150 H.	50	7500
	H. INGENIERO SENIOR 2	150 H.	50	7500
	% COSTES DIRECTOS COMPLEMENTARIOS	0.02	22027	4405.54
			Costes directos	22467.54
			Coste total	22467.54

Asciende el presupuesto de Ejecución Material a la expresa cantidad de Euros:

VEINTIDÓS MIL CUATROCIENTOS SESENTA Y SIETE EUROS CON CINCUENTA Y CUATRO CÉNTIMOS.

TOTAL EJCUCIÓN POR CONTRATA	22467.54	(€)
21 % IVA	4718.18	(€)

PRESUPUESTO TOTAL: 27185.72 (€)

Asciende el presupuesto base de licitación a la expresa cantidad:

VEINTISIETE MIL CIENTO OCHENTA Y CINCO CON SETENTA Y DOS CÉNTIMOS.