UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA MATEMÁTICA

APLICACIONES DE APRENDIZAJE DE MÁQUINAS Y EEG PARA SALUD MENTAL

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL MATEMÁTICO

TOMÁS ALONSO VALDIVIA HENNIG

PROFESOR GUÍA:
JUAN VELÁSQUEZ SILVA

MIEMBROS DE LA COMISIÓN:
FELIPE TOBAR HENRÍQUEZ
JUAN PEYPOUQUET URBANEJA

Este trabajo ha sido parcialmente financiado por el proyecto CMM Conicyt PIA AFB170001

SANTIAGO DE CHILE
2018

RESUMEN DE MEMORIA PARA OPTAR AL TÍTULO DE: Ingeniero Civil Matemático
POR: Tomás Alonso Valdivia Hennig
FECHA: 05/07/2018
PROFESOR GUÍA: Juan Velásquez Silva

RESUMEN

APLICACIONES DE APRENDIZAJE DE MÁQUINAS Y EEG PARA SALUD MENTAL

La ciencia de los datos y el aprendizaje de máquinas son un conjunto de técnicas y herramientas que han permitido gran desarrollo en diversas áreas de la técnica y las ciencias naturales. Las técnicas utilizadas en general no son necesariamente nuevas, sin embargo, el aumento explosivo en la capacidad de almacenamiento de datos y cómputos han hecho que estas herramientas hayan sido más aplicadas que nunca y su creciente fama puede deberse a la efectividad que tienen. Una de las áreas que se ha visto influenciada por este tipo de aplicaciones es la biomedicina.

Un desafío importante en neuropsiquiatría es tener exámenes objetivos que indiquen presencia de trastornos psiquiátricos por sus anormalidades a nivel fisiológico. Se han realizado diversos esfuerzos en esta línea, muchos con buenos resultados, aunque esto todavía no se emplea a nivel clínico.

En la presente memoria se evalúa la capacidad discriminativa de las mediciones de electroencefalografías para diferenciar poblaciones clasificadas previamente por expertos. Este trabajo busca ahondar en esta línea, mostrando que no sólo es posible para adultos, sino que también para niños y adolescentes. Junto a esto, se estudia el caso de la diferenciación utilizando grabaciones realizadas con un sistema de adquisición de bajo costo.

Para responder a estas interrogantes se diseña una metodología basada en ciencia de los datos y aprendizaje de máquinas para evaluar la capacidad de discriminación del EEG en reposo en trastornos psiquiátricos relacionados con la esquizofrenia. Se conducen dos experimentos, el primero con grabaciones de EEG clínico de adolescentes rusos sin tratamiento farmacológico, con síntomas asociados a esquizofrenia y sin síntomas; y el segundo con grabaciones de EEG tomadas con un hardware comercial de bajo costo a adultos chilenos en tratamiento farmacológico diagnosticados con esquizofrenia u otro trastorno psiquiátrico. Esta discriminación se plantea como un problema de clasificación.

La evaluación se realiza a través de un test estadístico de diferencia de medianas no paramétrico a los puntajes asignados por los clasificadores. Los resultados muestran que los clasificadores diferencian a ambas poblaciones de adolescentes con una significancia estadística del 0.1% con todos los métodos estudiados, se determinan además familias de características relevantes. El segundo caso logra una diferenciación con una significancia estadística del 5% para uno de los métodos empleados. Se estima que esto se puede mejorar con más datos, ajuste de los parámetros de los métodos y con la toma de datos de más datos.

A Adriana.

AGRADECIMIENTOS

Quiero agradecer a todas las personas que me acompañaron en este proceso (que fue más largo de lo que debió ser). Primero a mi familia. A mi nona, Adriana; a mi mamá, Vicky; a mi hermano, Nico; a mi hermana, Nely; a mis hermanitos, Magda y Maxi; y a mi padrastro, Pato.

Segundo, a mis amigos cercanos (por orden de aparición en el proceso), Gato, Matías, Rodrigo, Renzo, Nacho, Caro y Vale. Gracias por todo lo que me han enseñado, por su apoyo y preocupación. Al CRI que fue mi primer grupo ñoño en Santiago, al grupo B104 y amigos de la tía, a mis compañeros del DIM, amigos que hice a lo largo de los cursos, a mis amigos de Punta Arenas.

Tercero quisiera agradecer al grupo del WIC y al profesor Juan Velásquez, que me recibieron como memorista y me apoyaron en el camino para sacar adelante esta investigación. A Felipe Tobar por motivarme a ver el aprendizaje de máquinas desde una perspectiva más formal y matemática. A Pablo Guerrero por las conversaciones y por haberme brindado su confianza y apoyo al integrarme al laboratorio ALGES. A los miembros de ALGES.

Agradecer también a Karen Hernández y Natacha Astromujoff, quienes se preocuparon por hacer que este trabajo concluyera antes de que se siguiera alargando innecesariamente.

Quiero agradecer también a Catherine, quien ha sido una gran compañera y un gran apoyo durante la finalización de este trabajo; y a Igor. Finalmente quiero extender mi gratitud a todas las personas que fueron parte de mi proceso que por descuido y no por maldad, no he nombrado explícitamente.

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

Capítulo 1: Introducción

En este capítulo se presenta el contexto en el que se enmarca el desarrollo del trabajo de título, se aborda el contexto, el problema general, los objetivos tanto generales como específicos del mismo y el alcance del mismo.

1.1 Contexto

La ciencia de los datos y el aprendizaje de máquinas son herramientas que han permitido generar nuevo conocimiento y tecnologías en diversas áreas como la neurociencia, las ciencias sociales y la ingeniería. Presenta ventajas de utilizar herramientas tanto específicas del campo al que se aplica como agnósticas a este, la metodología es llevada por los datos más que por la intuición o experiencias personales de los investigadores, lo que puede traducirse en una ventaja metodológica al enfrentarse a un problema científico. Esto no quiere decir que sea una metodología sin sesgos ya que estos pueden provenir de los datos y su proceso de recolección (Crawford, 2016). Las aplicaciones de estas técnicas en medicina han ido en aumento, desde la utilización de modelos simples para establecer relaciones iniciales entre distintos factores para una posterior investigación hasta tener máquinas capaces de colaborar en diagnósticos (Dilsizian & Siegel, 2014).

Se estima que la prevalencia de la esquizofrenia es entre el 0.5 y el 1% y su incidencia se considera independiente del trasfondo racial, étnico o de género (Bhugra, 2005). Es una enfermedad de difícil diagnóstico pues presenta síntomas compartido con otros trastornos psiquiátricos y se trata siempre de diagnósticos clínicos que no utilizan información fisiológica más que para descartar una eventual epilepsia. Existen estudios que presentan buenos resultados a la hora de diferenciar esquizofrenia de otros trastornos psiquiátricos mediante el uso de electroencefalogramas (EEG) y técnicas de aprendizaje de máquinas, este trabajo busca mostrar que no sólo se puede hacer en el caso de pacientes con esquizofrenia y sin síntomas psiquiátricos, sino que además existe la posibilidad de diferenciar distintos trastornos con un sistema de adquisición de datos de bajo costo.

1.2 Problemática

El desafío que se plantea en el presente trabajo consiste en aplicar, desarrollar, implementar y evaluar técnicas propias de la ciencia de los datos y el aprendizaje de máquinas para la clasificación de sujetos según su diagnóstico psiquiátrico a partir únicamente de mediciones de potenciales eléctricos a nivel cortical en estado de reposo. Esto se basa en que algunos trastornos psiquiátricos presentan variaciones similares en lugares anatómicos cercanos a nivel de la corteza cerebral, más aún, que estas variaciones son perceptibles mediante un hardware de mediciones de EEG extracraneal cuyo diseño, en uno de los casos a analizar, está pensado como interfaz cerebro computador para neurorretroalimentación y videojuegos.

Para esto se utilizan dos conjuntos de datos distintos, uno que perteneces a un laboratorio de investigación ruso, y consta de mediciones hechas a niños y adolescentes, en el cual hay sujetos con y sin síntomas de esquizofrenia y el otro de fabricación propia utilizando un neuroheadset

EPOC a través del Centro de Inteligencia Web (WIC) de la Facultad de Ciencias Físicas y Matemáticas (FCFM) de la Universidad de Chile, en conjunto con el Hospital de Día del Servicio de Psiquiatría del Hospital Barros Luco-Trudeau, que consta de mediciones hechas a personas en tratamiento psiquiátrico por distintos trastornos, principalmente de esquizofrenia.

Para poder utilizar las mediciones de EEG se emplean técnicas de ciencias de los datos e ingeniería de características basadas en análisis lineal y no lineal de series de tiempo. Estas herramientas están presentes en la literatura científica y en general con buenos resultados, sin embargo, dentro de la información que el autor ha recabado, no existen estudios que utilicen este tipo de hardware en este tipo de problemas de clasificación.

1.3 Objetivos

La orientación del trabajo tiene que ver con diseñar una metodología desde la ciencia de los datos para enfrentarse al problema de utilizar EEG para contribuir a la decisión diagnóstica en el contexto de trastornos psiquiátricos. Los objetivos irán en esta línea, la de mostrar una significancia estadística en el poder de clasificación y etiquetamiento de los métodos de aprendizaje de máquinas con un esquema basado en ciencias de los datos.

1.3.1 Objetivo general

Evaluar la capacidad de la metodología de ciencia de los datos para diferenciar poblaciones de sujetos mediante electroencefalografías en el contexto de diagnósticos psiquiátricos relacionados a la esquizofrenia.

1.3.2 Objetivos específicos

a) Diseñar una metodología basada en ciencias de los datos para evaluar las hipótesis de investigación.

b) Implementar un esquema de análisis y construcción de características para los EEG basado en análisis lineal y no lineal de series de tiempo.

c) Entrenar y evaluar distintos métodos de clasificación y extraer grupos de características relevantes para posterior análisis.

d) Determinar la validez estadística de las hipótesis de investigación.

1.4 Alcance y contribuciones del estudio

El trabajo realizado se circunscribe en:

- Obtención, selección y transformación de datos para poder responder a las hipótesis planteadas.

- Diseño e implementación de rutinas para construir características a partir de las grabaciones de EEG.

- Análisis de los resultados de clasificación para evaluar estadísticamente si la metodología de clasificación permite concluir rechazando la hipótesis nula.

Es importante aclarar que los resultados obtenidos son, en principio, solo aplicables a poblaciones similares a las que se utilizan en el estudio, aunque las herramientas generadas y las técnicas implementadas son potencialmente útiles en otros contextos de análisis de EEG. Del resultado se espera encontrar evidencia de la expresividad del EEG tomado en reposo para diferenciar distintas poblaciones con trastornos psiquiátricos en los contextos mencionados.

1.5 Estructura de la memoria

La memoria se divide en cinco capítulos descritos a continuación.

El capítulo 1 sirve de presentación del contexto y problemas a abordar, así como los objetivos esperados de la realización de trabajo. Posteriormente se delimitan los alcances y las contribuciones esperadas.

El capítulo 2 se constituye como presentación del marco teórico que da sustento técnico y científico al trabajo realizado. Se subdivide en cuatro partes: actividad bioeléctrica; esquizofrenia; ciencia de los datos, aprendizaje de máquinas y estadística; y revisión bibliográfica.

El capítulo 3 contiene la estructuración de la metodología empleada en la investigación, a modo de permitir la replicabilidad de los resultados. En este se explicitan las hipótesis de investigación y cómo se procede para evaluarlas.

El capítulo 4 muestra los resultados obtenidos y su análisis correspondiente al utilizar el esquema descrito en el capítulo anterior, concluyendo con una discusión en torno a estos.

Finalmente, en el capítulo 5 se concluye en base a los objetivos inicialmente planteados y los resultados obtenidos. Se presentan recomendaciones para posibles trabajos futuros.

Capítulo 2: Marco Teórico

En este capítulo se presentan las bases y herramientas teóricas que permiten dar sentido al trabajo realizado. El capítulo se divide fundamentalmente en cinco partes. La primera trata sobre las bases neurobiológicas que dan origen al fenómeno físico a medir, la segunda es una breve descripción del trastorno de esquizofrenia más algunas consideraciones al respecto para así dar contexto al problema posterior de clasificación, la tercera es sobre las bases técnicas de ciencias de los datos y aprendizaje de máquinas, luego se provee una breve descripción del test de hipótesis a considerar para finalmente concluir con una discusión en torno al estado del arte.

2.1 Actividad bioeléctrica.

La actividad bioeléctrica corresponde a los fenómenos eléctricos que ocurren en seres vivos. El principal agente de esto en el cuerpo humano son las neuronas En lo respecta a este trabajo, la actividad bioeléctrica que se busca medir es la cerebral, sin embargo, existen otras fuentes como las asociadas a movimientos musculares que pueden entorpecer esta tarea.

2.1.1 De la neurona al EEG

La neurona es una célula altamente especializada y es la unidad funcional del sistema nervioso. Todos los procesos neurológicos dependen de interacciones complejas célula a célula de neuronas individuales y/o de grupos relacionados de éstas.

Las neuronas se polarizan con respecto al medio en el que están mediante proteínas embebidas en las membranas citoplasmáticas llamadas canales iónicos. Estas proteínas les permiten tener una diferencia de potencial acumulado que pueden liberar bajo ciertas circunstancias. La liberación de este se denomina liberación del potencial de acción. Las neuronas se dividen morfológicamente en tres zonas: Soma, dendritas y axón. El Soma es el centro metabólico de la neurona y es donde se ubica el núcleo junto a la mayoría de organelos. De este pueden ramificarse un número variable de dendritas y un axón. Las dendritas reciben señales de otras neuronas generando un potencial se señales agregadas. Si este potencial alcanza cierto valor umbral al llegar al soma, la neurona liberará su potencial de acción por el axón. El axón puede estar asociado a distintos receptores como por ejemplo otras neuronas.

Las neuronas se asocian con otras neuronas a través de sinapsis. Las sinapsis son aproximaciones funcionales entre una neurona y otra célula especializada. La función de esta aproximación es transmitir una señal desde la neurona presináptica a la célula postsináptica.

Cuando muchas neuronas se asocian pueden formar redes. En estas redes se tienen muchas neuronas liberando potenciales de acción de manera recurrente, generando asociaciones y estructuras particulares. Estas asociaciones, liberaciones de potencial y estructura crea oscilaciones en el campo eléctrico, son estas variaciones agregadas del potencial que pueden medirse

extracranealmente con un electroencefalógrafo[1]. Es esta variación promedio la que es capaz de medir el sistema. Un EEG entonces es una medición del potencial que causan grupos de neuronas sobre alguna parte de la zona encefálica.

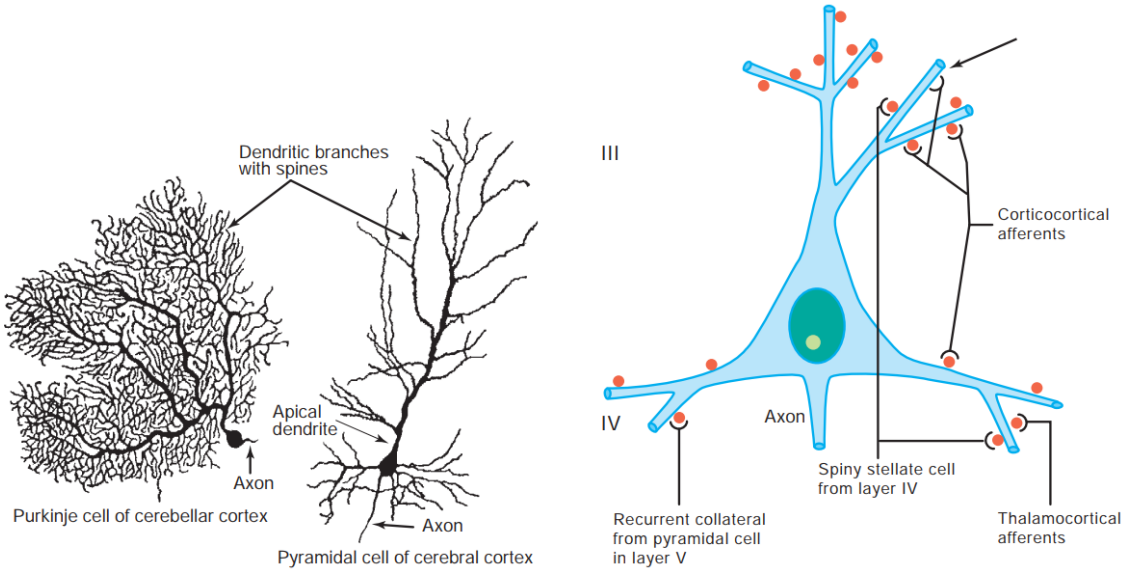


Figura 1: Esquema de neuronas (Lac, Squire, Bloom, & Berg, 2008).

Los grupos de neuronas pueden producir distintos tipos de sincronías, estas se dividen tradicionalmente en ritmos de frecuencias dominantes, esta división se hace en ritmos delta ($< 4\ Hz$), theta ($4 - 8\ Hz$), alfa ($8 - 13\ Hz$), beta ($13 - 30\ Hz$), gamma ($30 - 100\ Hz$) y de alta frecuencia (>100 Hz), estos ritmos además están asociados a funcionalidades distintas y compiten una con otras (Treviño & Gutiérrez, 2007).

Existen distintos tipos de EEG, desde externos y no invasivos hasta intracraneales. El estándar que se usa para garantizar replicabilidad en los estudios y toma de EEG se denomina sistema internacional 10-20. Este sistema estandariza la posición de los electros relativos a una protuberancia y

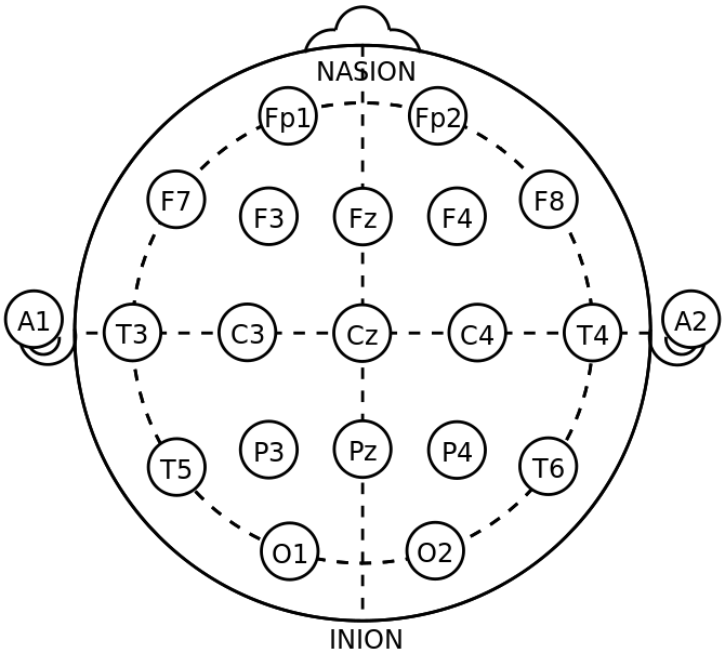


Figura 2: Esquema de posición de electrodos para el sistema internacional 10-20. Los pares (impares) corresponden al lado derecho (izquierdo resp.). (Wikimedia Commons, 2015)

[1] También existen sistemas de adquisición de EEG subdurales (intracraneales) invasivos que no serán considerados por su poca utilidad para un paciente con el tipo de trastorno considerado. Estos pueden resultar útiles para personas que sufran de epilepsia.

una depresión ósea. Usa una nomenclatura para los electrodos asignando primero una letra según en qué región está, F para frontal, T para temporal, C para central, P para Parietal y O para occipital, números para la distancia desde el centro en la que los pares están asignadas al hemisferio derecho y con impares al izquierdo. La letra z (de cero en inglés) indica que está en la línea divisoria. La distancia entre electrodos se mide siempre como porcentaje de la distancia entre las referencias óseas. Además de esto, es necesario identificar el voltaje de referencia contra el que se está midiendo la variación de potencial. Para esto existen diversos métodos, como utilizar el lóbulo una o ambas orejas, o referenciarlo a un electrodo en particular para luego sustraer el promedio.

2.1.2 Emotiv EPOC

El hardware utilizado para la toma de EEG es un Neuroheadset EPOC, electroencefalógrafo inalámbrico comercializado por la empresa Emotiv. Su diseño está inicialmente pensado como periférico para interfaz cerebro-computador, sin embargo, se ha utilizado en diversos estudios científicos por ser un hardware simple de usar y de bajo. El sistema no ha estado exento de críticas, pero la evidencia apunta a que es utilizable para investigación, a pesar de ser muy ruidoso (Duvinage, y otros, 2013). Graba 14 canales de actividad eléctrica referenciados a otros dos y su frecuencia de muestreo es de 256 Hz para guardar los datos a 128Hz reduciendo problemas de aliasing.

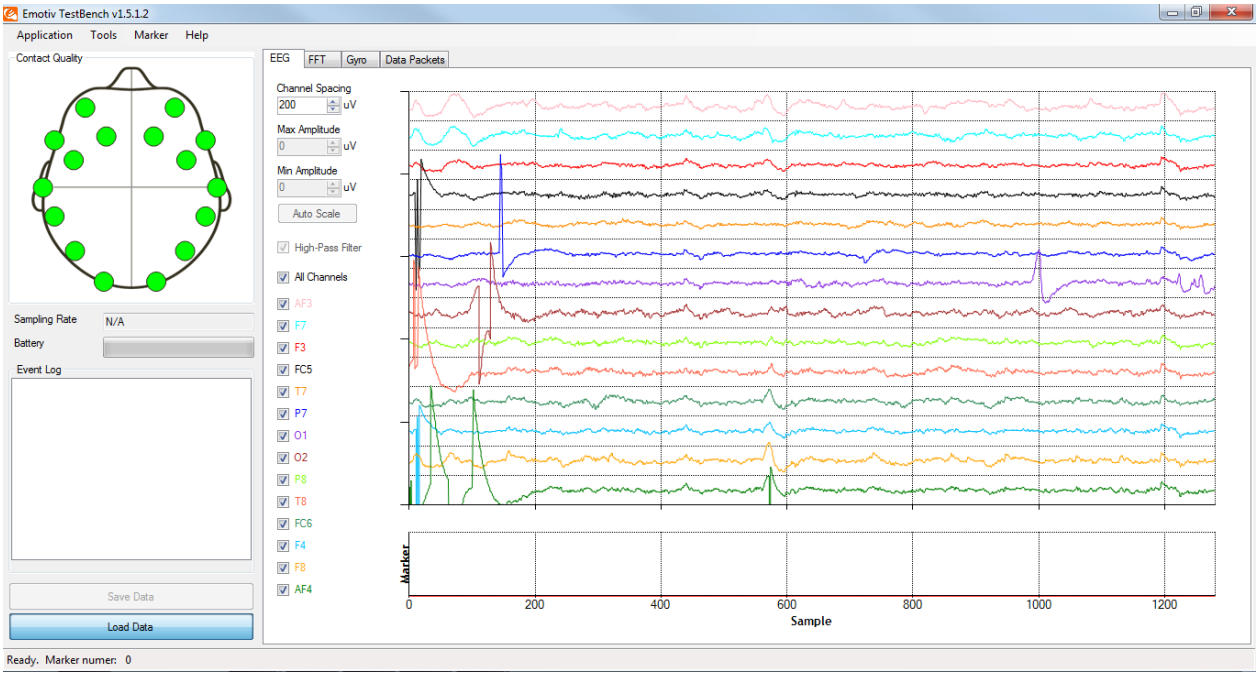


Figura 3: Captura de pantalla de software TestBench para la adquisición de datos de EEG.

Los canales están asociado a las divisiones tradicionales de un EEG, sin embargo, al tener un tamaño estándar su colocación no coincide exactamente con lo que sería hacer estas tomas con la precisión de un examen médico.

2.2 Esquizofrenia

La esquizofrenia es un trastorno mental crónico y severo, caracterizado por una profunda alteración del pensamiento, que afecta el lenguaje, la percepción y el sentido de sí mismo (OMS, WHO, 2016). Un análisis de los síntomas sugiere que pueden ser agrupados en cinco categorías: (i) Psicosis (Síntoma positivo); (ii) Alteración en los impulsos y voluntad (Síntomas negativos); (iii) Alteraciones neurocognitivas; (iv, v) Desregulación afectiva depresiva y maniaca (van Os, 2009). En este momento no existe una cura para la esquizofrenia, sin embargo, existen tratamientos que controlan los síntomas positivos del trastorno.

La prevalencia de vida de la enfermedad a nivel mundial se estimas de entre el 0.5 al 1%, es decir que entre cinco y diez de cada mil personas tiene o tendrá esquizofrenia (Bhugra, 2005). La evidencia científica dice que el diagnóstico y tratamiento temprano de esquizofrenia lleva a mejores resultados a largo plazo, esto acompañado a que, también, se estima que el tratamiento es costo efectivo los esfuerzos por detectar y empezar un tratamiento lo antes posible genera bienestar no sólo para quién sufre de la enfermedad, sino que para la sociedad toda (Agius, Butler, & Holt, 2011).

Para el diagnóstico de esta enfermedad no se utilizan exámenes fisiológicos o anatómicos, aun cuando estos se han reportado, en parte, porque no se justifica la utilización del recurso para establecer un diagnóstico. Sin embargo, la falta de marcadores biológicos a la hora de establecer un diagnóstico en psiquiatría ha sido fuente de debate y polémica durante décadas (Rosenhan, 1973).

Hoy se cuenta con un corpus grande de evidencia científica que indica la presencia de alteraciones neurológicas en pacientes con esquizofrenia y otros trastornos psiquiátricos, más sobre esto en el apartado de 2.5 de revisión bibliográfica.

La guía clínica AUGE para el tratamiento de personas desde el primer episodio de esquizofrenia recomienda considerar según las circunstancias la toma de un EEG estándar en el caso de pacientes con primer brote psicótico, esto como medida para confirmar o descartar posibles patologías neurológicas-médicas (MINISTERIO DE SALUD, 2016). En caso no encontrarse evidencia que sugiera este último tipo de patologías, el examen no es utilizado.

2.3 Ciencia de los datos, aprendizaje de máquinas y estadística

La ciencia de los datos es una disciplina que involucra el uso de métodos para analizar datos con el fin de extraer conocimiento de estos (Cielen, Meysman, & Ali, 2016). En muchos sentidos es una extensión de la estadística que toma elementos de ciencias de la computación para su aplicación. El proceso de extracción de conocimiento o información mediante el análisis de datos ha sido explorado y desarrollado de distintas maneras. Algunos estándares del proceso de ciencia de los datos son *KDD (Knowledge Discovery in Databases)*, *SEMMA (Sample, Explore, Modify, Model and Assets)* y *CRISP-DM (Cross-industry Standard Process for Data Mining)*, que se diferencian en dónde ponen los énfasis, pero que finalmente son muy similares (Azevedo & Santos, 2008). Una forma de dividir el proceso es:

i. Seleccionar el objetivo de la investigación
ii. Recuperación de datos
iii. Preparación de datos
iv. Exploración de los datos
v. Modelamiento de los datos
vi. Presentación y automatización

Es importante entender que la enumeración no implica necesariamente la realización lineal del proceso, algo que el estándar *CRISP-DM* enfatiza en su metodología, pero que aplica a cualquier proyecto aplicado de ciencia de los datos.

La etapa de *Seleccionar el objetivo de la investigación* hace referencia a definir claramente lo que se quiere lograr al finalizar el proceso, involucra además responder a las preguntas *Qué*, *Cómo* y *Por qué*. La etapa de *Recuperación de datos* involucra la obtención de estos, ya sea a través de datos disponibles o al diseño para tomar los mismos. La *Preparación de datos* tiene relación con los preprocesos a los que se deben someter los datos para ser utilizados, se incluye la limpieza, transformación y combinación. En la *Exploración de los datos* se busca obtener una noción más completa de la estructura de los datos, una de las herramientas más potentes de esta parte es la visualización de datos, aunque no es la única. Otros elementos utilizados pueden ser reducciones de dimensionalidad con fines exploratorios y gráficos, o la confección de modelos simples. Los modelos más complejos y la construcción de características están considerados dentro del paso *Modelamientos de los datos*. Además de estas tareas, también se considera la evaluación y comparación de modelos. Finalmente se pasa a la *Presentación y automatización*, que consiste en la exposición de resultados y automatización de los procesos para la reproducibilidad. En el proceso de ciencia de los datos, las técnicas utilizadas principalmente en el quinto paso provienen de la rama del aprendizaje de máquinas.

El aprendizaje de máquinas es un conjunto de técnicas estadísticas y algorítmicas que permiten detectar patrones en datos de manera automática, para luego utilizar esos patrones en la tarea de predecir datos futuros, o para ejecutar otro tipo de decisiones bajo incertidumbre. Los métodos tradicionalmente se dividen en supervisado, no supervisado y reforzado (Murphy, 2012).

Los métodos supervisados buscan ser predictivos y se basan en conjuntos de datos en los que cada observación tiene un valor asignado. En este contexto, cada dato tiene asociada una etiqueta que corresponde al valor de la función que se ha de aprender. El problema puede verse formalmente como, dado un conjunto de entrenamiento $D = \{(x_i, y_i)\}_{i=1}^{N}$, con $x \in X$ e $y \in Y$ conjuntos apropiados, y se quiere encontrar una función $f: X \to Y$ tal que sea capaz de generalizar la relación existente entre las características x y las respuestas y. Un ejemplo clásico es la regresión lineal por mínimos cuadrados, es decir, encontrar la recta que minimiza la suma del cuadrado de las diferencias.

Los métodos no supervisados buscan estructuras que pueden en principio no ser fácilmente observables en los datos, son técnicas descriptivas y representa un problema no tan bien definido como el caso anterior. En este contexto se cuenta con un conjunto de datos $D = \{x_i \in X\}$ y se quiere encontrar una función $f: X \to Y$ que tenga utilidad en el reconocimiento de patrones. Esto puede ser para agrupar observaciones similares (clústering o aglomeración), determinar distribución de los datos (estimadores de densidad) o encontrar factores latentes que generan a los

datos. Un ejemplo de esto último es reducir la dimensionalidad de las características para, entre otras cosas, poder visualizar posibles estructuras.

El aprendizaje reforzado, por otro lado, toma inspiración en los modelos de psicología conductista y se busca aprender una política de acción en base a prueba y error. El modelo puede verse como un proceso de decisión de Markov en el que no se conocen a priori las distribuciones de transición ni la función de pagos en su totalidad.

2.3.1 Clasificadores

En aprendizaje de máquinas se denomina clasificador a algoritmos supervisados sobre datos cuyas etiquetas son categóricas. Así, se busca encontrar una forma de decidir la categoría de un nuevo dato, a partir de los datos observados. El problema más simple de este tipo es el caso de clasificación binaria, en el que la etiqueta y se restringe a valores en $\{0,1\}$. Las formas más comunes de enfrentar el caso multiclase es a través de múltiples clasificadores binarios, esto se puede hacer de más de una forma y la solución dependerá del problema en particular. A continuación, se describen algunos de los métodos más utilizados para la tarea de clasificación binaria.

2.3.1.1 Regresión logística

La regresión logística es un método de clasificación binario que puede ser considerado como una extensión de la regresión lineal. En este caso se quiere estimar la probabilidad de que un dato x pertenezca a la clase y. El modelo utilizado entonces es

$$\mathbb{P}(y|x,w) = Ber\big(y\big|\mu_w(x)\big),$$

esto es, que y se distribuye como una variable aleatoria Bernoulli de parámetro $\mu_w(x)$, con

$$\mu_w(x) = sigm(w^T x)\,.$$

En este caso $sigm$ corresponde a una función sigmoidea conocida como función logística definida como

$$sigm(\eta) \coloneqq \frac{1}{1-e^{-\eta}} = \frac{e^{\eta}}{e^{\eta}+1}.$$

Para construir el clasificador, el método se basa en encontrar el mejor vector de pesos w acorde a alguna regla de optimalidad. Típicamente, para aprender los parámetros se utiliza el método de máxima verosimilitud en el que se maximiza

$$\mathbb{P}(y|x,w) = \prod_{i\in D_e} \mu_w(x_i)^{y_i}\,(1-\mu_w(x_i))^{1-y_i}$$

con respecto a la variable w. En la práctica, se minimiza el logaritmo negativo de esta cantidad (NLL) más un factor de regularización. La elección de utilizar el NLL en vez de la expresión original viene motivada por la estabilidad numérica que esto genera al problema.

El factor de regularización más común es una penalización sobre el cuadrado de la norma del vector de coeficientes a aprender, como además este factor es estrictamente convexo, mantiene la unicidad de la solución. La expresión por minimizar en este caso es

$$F_{X,y}(w) = \sum_{i=1}^{N} \log(1 + \exp(-y'_i w^T x_i)) + \lambda \|w\|_2^2 ,$$

donde y' es un reetiquetamiento de y, en el que la clase 0 es reemplazada por -1. En esta expresión, la sumatoria representa la NLL del modelo probabilístico y la norma al cuadrado por una constante el factor de regularización, sin embargo, esto también puede verse como una estimación de máximo a posteriori (MAP) con una distribución a priori gaussiana sobre los pesos.

Otra técnica de regularización muy utilizada emplea $\|w\|_1$ en vez de $\|w\|_2^2$. Esta técnica llamada operador de selección y contracción mínima absoluta (LASSO por su nombre en inglés) o regularización ℓ_1 fue introducida inicialmente por Thibshirani 1996 para problemas de regresión de mínimos cuadrados. El nombre original se debe a que las soluciones de este problema tienden a aprender vectores ralos, es decir, con muchos coeficientes con valor 0. Este factor también tiene una interpretación como MAP, en el que la distribución a priori de los pesos corresponde a una laplace.

2.3.1.2 SVM

Las máquinas de soporte vectorial (SVM por su sigla en inglés) son un tipo de clasificador introducido en la década de los 90 por Cortes y Vapnik (Cortes & Vapnik, 1995). La idea básica proviene de encontrar en el espacio de las características un hiper-plano separador entre dos clases. En este sentido, múltiples clasificadores pueden ser interpretados de esta manera incluida la regresión logística. La diferencia que se hace en las SVM es buscar aprender el hiper-plano que maximice el margen de separación entre los datos de las clases. Formalmente, el hiper-plano afín que se busca queda descrito por dos parámetros w y b, y es el lugar geométrico que satisface

$$w \cdot x + b = 0 , \quad x \in \mathbb{R}^d,$$

en el que estos dos parámetros son la solución al problema de optimización

$$\begin{aligned} &\min \ \frac{1}{2} \|w\|^2 \\ &s.a: \ y'_i(w \cdot x_i + b) \geq 1 , \forall i \in D \end{aligned}$$

donde y' es un reetiquetamiento de y, en el que la clase 0 es reemplazada por -1. Este problema tiene solución sólo si los datos son separables, algo que suele no suceder naturalmente, es por eso que se extiende esta idea a conjuntos de datos no separables, mediante la adición de variables de holgura no negativas ξ. La nueva formulación entonces queda como

$$\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i$$
$$s.a:\ y'_i(w \cdot x_i + b) \geq 1 - \xi_i\ , \forall i \in D$$
$$\xi_i \geq 0, \qquad \forall i \in D$$

Este problema cumple con propiedades de regularidad y convexidad que permiten asegurar que tiene solución, su dual también y cumplen con las condiciones de Karush-Kuhn-Tucker (KKT). Típicamente se plantea y se resuelve el problema dual de este. Para esto se toma el lagrangeano del problema

$$L(w, b, \xi, \alpha, \beta) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i(1 - \xi_i - y'_i(w \cdot x_i + b)) - \sum_{i=1}^{n} \beta_i \xi_i\ ,$$

que da origen al problema dual

$$\max \sum_{i=1}^{n} \alpha_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y'_i y'_j (x_i \cdot x_j)$$
$$s.a:\ 0 \leq \alpha_i \leq C\ \ , \forall i \in D$$

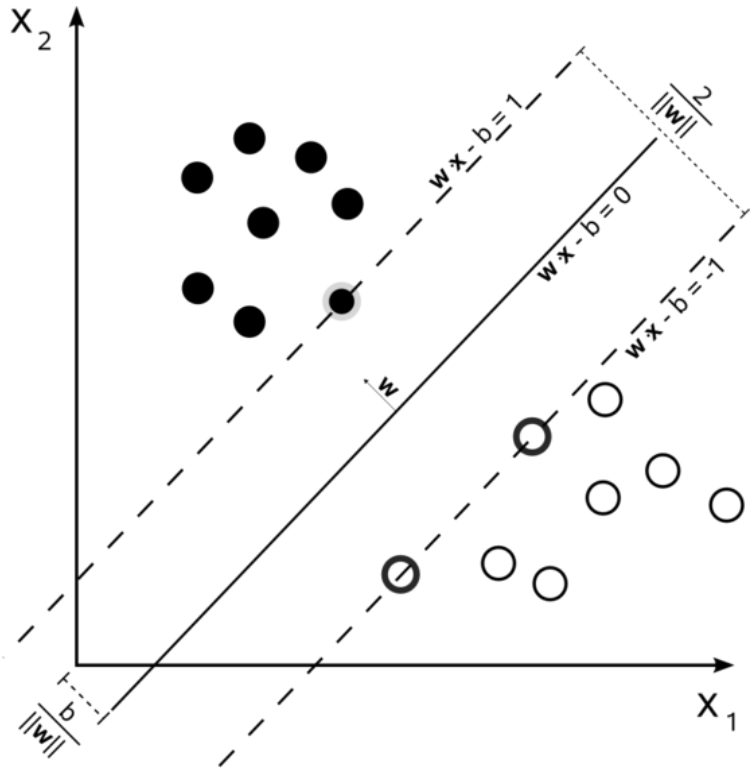


Figura 4: Ejemplo de hiperplano separador para SVM. (Wikimedia Commons, 2016)

$$\sum_{i=1}^{n} \alpha_i y_i' = 0 \, .$$

Al resolver este problema se determinan los valores óptimos de las variables primales utilizando las condiciones de KKT. De esto se desprende que $w = \sum_i \alpha_i y_i' x_i$, para los k tal que $\alpha_k > 0$, se satisface que $y_k'(w \cdot x_k + b) = 1 - \xi_k$, además si $\alpha_k < C$ se tiene que $\xi_k = 0$, de donde se puede calcular el parámetro b y posteriormente los ξ_i. Si bien todo esto es factible, en la práctica no es necesario calcular explícitamente w, pues para clasificar el dato k se necesita únicamente el signo de la expresión $(w \cdot x_k + b) = (\sum_i \alpha_i y_i'(x_i \cdot x_k) + b)$. Esta forma se prefiere pues permite la utilización de métodos basados en kernel (para más detalles, ver 2.7).

2.3.1.3 Naive Bayes

El clasificador Naive Bayes (o Bayes ingenuo) es un método relativamente simple, usado frecuentemente en diversas aplicaciones o como base de comparación. Se basa en suponer una distribución de datos condicionalmente independiente y según esta realizar la clasificación. Para esto se modelan los datos como provenientes de dos distribuciones distintas, cada una con una distribución sobre las características de la forma

$$\mathbb{P}(x|y = c, \theta) = \prod_{j=1}^{d} \mathbb{P}(x_j|y = c, \theta_{jc}) \, .$$

En este contexto $c \in \{0,1\}$ y es la clase a la que pertenece la observación x, θ son los parámetros que definen la distribución multivariada y los θ_{jc} los parámetros que definen las distribuciones marginales de los x_j cuando vienen de la clase c.

Este modelo supone que las distribuciones de las características son independientes condicionalmente a la clase a la que pertenecen (de ahí el nombre "ingenuo"), lo que en general no es esperable, sin embargo, es considerado un clasificador efectivo que, aun cuando el modelo no sea realmente bueno para estimar la probabilidad, clasifica exitosamente (Rish, 2001).

Las distribuciones utilizadas para modelar las marginales $\mathbb{P}(x_j|y = c, \theta_{jc})$ dependen de la naturaleza de los datos. Para datos continuos es común utilizar marginales gaussianas cuyos parámetros se pueden aprender mediante el método de máxima verosimilitud.

2.3.1.4 Árbol de decisión

Los modelos de aprendizaje basados en árboles de decisión (DT por sus siglas en inglés), también conocidos como árboles de clasificación y regresión (CART por sus siglas en inglés) son un tipo de modelo predictivo que utiliza la estructura de árbol de decisión para realizar la clasificación o regresión. El modelo en cuestión se basa en un árbol con raíz en el que cada nodo no hoja del árbol representa una separación que se realiza por una única característica, por ejemplo $x_1 \leq \theta_0$ da origen a dos nodos descendientes; según esta separación se procede al siguiente nodo, si es una hoja, se asigna el valor de dicha hoja, si no, se continua con la separación.

El aprendizaje que se realiza en este procedimiento son las preguntas en los nodos internos y los resultados en los nodos hoja. Encontrar el particionamiento óptimo es un problema NP-completo, pero existen algoritmos basados en enfoques avaros que producen resultados suficientemente buenos. Para estos enfoques es necesario escoger alguna métrica de costo de clasificación. Algunas de las más usadas en clasificación son la tasa de clasificación errónea:

$$\frac{1}{n}\sum_{i=1}^{|D|} \mathbb{1}_{\{y_i \neq y^*\}},$$

en el que y^* es la clase más representada en el nodo a evaluar. La entropía:

$$-\sum_{y'=0}^{1} \frac{|\{d \in D \mid y_d = y'\}|}{n} \log\left(\frac{|\{d \in D \mid y_d = y'\}|}{n}\right).$$

El índice de Gini:

$$\sum_{y'=0}^{1} \frac{|\{d \in D \mid y_d = y'\}|}{n} \left(1 - \frac{|\{d \in D \mid y_d = y'\}|}{n}\right).$$

En este contexto, D es el conjunto de datos que quedan en el nodo a evaluar y las fórmulas están descritas para clasificación binaria, el caso multiclase no difiere mucho y puede revisarse en (Murphy, 2012), para más detalle y otras métricas en (Lior & others, 2014). Es importante notar además que se puede restringir la cantidad máxima de características elegidas limitando la profundidad del árbol. Notando que la cantidad de preguntas corresponde los nodos internos y la cantidad de nodos internos es igual a $2^p - 1$ en caso de un árbol binario, donde p es la profundidad.

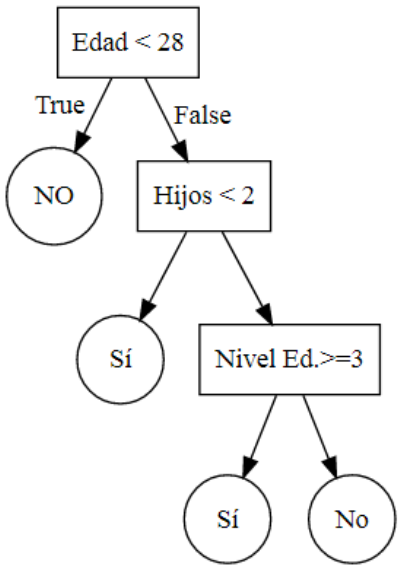


Figura 5: Ejemplo de árbol de decisión. Fabricación propia.

2.3.1.5 Bosque aleatorio

Los bosques aleatorios (RF por sus siglas en inglés) son un método de conjuntos de clasificadores que utilizan varios DTs entrenados en los datos. Si bien, los algoritmos de aprendizaje pueden obtener resultados distintos ante las mismas entradas debido a heurísticas aleatorizadas, estos generalmente obtienen resultados altamente correlacionados. Para descorrelacionar los DTs se utilizan dos métodos de aleatorización de la entrada, el primero es remuestrear los datos y el segundo es elegir un conjunto aleatorio de las características a utilizar para el aprendizaje. Con esto se obtienen DTs menos exactos que al no aleatorizar, sin embargo, al considerarlos todos y promediar el resultado, resulta en un mejor clasificador (Lior & others, 2014).

2.3.2 Ingeniería de características

Uno de los pasos más importantes en los problemas aplicados de aprendizaje de máquinas es encontrar descriptores relevantes para la tarea en particular. En principio, los datos se pueden utilizar tal como son extraídos, pero esto dificulta la interpretación y, por la naturaleza de las series de tiempo estacionarias, no hay un marco claro sobre cuándo se está empezando la señal a procesar en cuestión, por lo que la forma de proceder es extraer características que sean capaces de resumir información relevante de las secuencias.

2.3.2.1 Matrix de covarianza (Cov)

La matriz de covarianza de una variable aleatoria X se define como

$$\Sigma = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T].$$

Dado un conjunto de vectores de datos $\{x_i\}_{i=1}^n$ i.i.d., un estimador insesgado de la matriz de covarianza es

$$Q = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T,$$

donde $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Esta matriz es simétrica semidefinida positiva y si hay al menos d vectores linealmente afines, entonces la matriz es definida positiva. Como es una matriz simétrica, los datos no redundantes no son d^2, sino $\frac{d(d+1)}{2}$, donde d es la dimensión de cada observación.

2.3.2.2 Modelo Autoregresivo (AR8)

Los modelos autoregresivos son una familia de modelos que permiten representar series de tiempo estacionarias. Estos suponen que el valor que toma es una combinación lineal de los valores previos y un término aleatorio. Formalmente, un modelo autoregresivo X de orden p queda definido por una constante c, un vector de coeficientes ϕ de dimensión p y un proceso aleatorio $\{Z_t\}$ que se distribuya como ruido blanco, es decir, que tenga media cero y $cov(Z_t, Z_s) = \sigma^2 \mathbb{1}_{\{t=s\}}$ y cumple que

$$X_t = c + \sum_{i=1}^{p} X_{t-i}\phi_i + Z_t.$$

Los coeficientes calculados de modelos autorregresivos para EEG han sido utilizado por varios autores para clasificar trastornos psiquiátricos y los resultados que han tenido muestran que para mediciones a 128Hz el orden adecuado a escoger es 8, por lo que se toma en cuenta esta recomendación (Green, 2012).

2.3.2.3 Complejidad de Lempel-Ziv (LZC)

La complejidad de Lempel-Ziv es una estimación conservadora de la complejidad de una secuencia binaria finita y permite evaluar su aleatoriedad. Esta estimación se hace a partir de contar las diferentes subcadenas que se encuentran al leer la secuencia en orden (Lempel & Ziv, 1976). Al ser una medida de calculada sobre secuencias binarias no es directamente aplicable a los datos provenientes de un EEG, para realizar la codificación a secuencia binaria se hace sobre una nueva serie $\bar{S}$ definida como

$$\bar{S}(t) = \begin{cases} 1 & si\ S(t) > T \\ 0 & si\ S(t) \leq T. \end{cases}$$

2.3.2.4 Dimensión fractal de Katz (Kat)

La dimensión fractal de Katz es una aproximación a la dimensión fractal de una secuencia (Esteller, Vachtsevanos, Echauz, & Litt, 2001). Ha sido ampliamente utilizada en análisis de series de tiempo en biomedicina y se calcula como

$$D = \frac{\log L}{\log d},$$

con $L = \sum_{t=1}^{n-1} |X_t - X_{t+1}|$ y $d = \max_{t} |X_1 - X_t|$.

2.3.2.5 Entropía Espectral (SpE)

La entropía de información es una medida de incertidumbre de una variable aleatoria discreta. Dada una variable aleatoria discreta $X: \Omega \to I$, se define su entropía como

$$H(X) = -\sum_{i \in I} p(i) \log\big(p(i)\big).$$

Usando esta función, se calcula la entropía del espectro de frecuencias de una señal. Para esto se toma la transformada de Fourier discreta de la señal, con la cual se obtiene la densidad de energía discreta. Una vez obtenido este resultado, se normaliza para tener una densidad de probabilidad discreta a la que se le calcula la entropía. Este cálculo se puede realizar sobre el espectro completo de la señal o sobre bandas del mismo. En este caso se toman las entropías espectrales de la señal completa y, por separado, de las bandas asociadas a los ritmos delta, theta, alfa, beta y gamma.

2.3.2.6 Entropía de la densidad del período de recurrencia (RPDE)

La RPDE es una característica que se basa en sistemas dinámicos y se utiliza como medida de la repetitividad de una señal. Para esto, entonces, se realiza inicialmente un encaje lineal de la serie de tiempo en un espacio de mayor dimensión. Sobre la señal encajada se calcula la distribución empírica del período de recurrencia, y es a esta distribución a la que se le calcula la entropía.

El encaje que se utiliza es una transformación lineal que toma la secuencia unidimensional $S = (x_0, x_1, \dots, x_n)$ y la transforma en una secuencia de vectores d dimensionales $X = \{X_i\}_{i=0}^{n-d}$, donde $X_i = [x_i, x_{i+1}, \dots, x_{i+d-1}]$. Como la distribución es continua sobre los reales, los tiempos de retorno se calculan para ε-vecindades, con ε fijo. Para este trabajo se usa un encaje 11-dimensional tomando en consideración los trabajos de (Green, 2012) y (Jeong, 1998). La constante ε se fijó de manera heurística en 1, por tratarse de una serie normalizada.

2.3.2.7 Entropía Aproximada (ApEn)

La Entropía Aproximada (ApEn) es una medida que cuantifica la regularidad de una secuencia de datos, fue propuesta originalmente en (Pincus, Gladstone, & Ehrenkranz, 1991) y su motivación eran series de tiempo de mediciones fisiológicas relativamente cortas y ruidosas. La medida asigna un valor no negativo a las secuencias. Valores cercanos a 0 conllevan una mayor regularidad que aquellos más altos.

Para el cálculo de la ApEn es necesario fijar dos parámetros, un parámetro m que corresponde al largo de las subsecuencias a comparar y un parámetro r que corresponde a la tolerancia de similitud entre estas secuencias, estos parámetros juegan roles similares a los parámetros d y ε de la RPDE, respectivamente. Una vez fijados estos parámetros se procede de la siguiente manera:

i. Se define $C_r^m(i) = \frac{1}{N-d+1}\sum_{j=0}^{N-d} \mathbb{1}_{\{||X(i)-X(j)||_\infty < r\}}$
ii. Se define $\Phi_r^m = \frac{1}{N-d+1}\sum_{i=0}^{N-d} \ln C_r^m(i)$
iii. $ApEn_r^m = \Phi_r^m - \Phi_r^{m+1}$.

Las constantes elegidas son las mismas que para el caso de RPDE.

2.3.3 Reducción de dimensionalidad

Las técnicas de reducción de dimensionalidad son herramientas que se utilizan para disminuir el número efectivo de variables a considerar de los datos, esto puede ser a través de extracción de características o selección de características.

2.3.3.1 Extracción de características

Los métodos de extracción de características suponen que las características iniciales presentan una estructura latente de menor dimensionalidad y que parte de la variabilidad se debe a ruido o algún otro proceso no relevante para el estudio. El método más usado para esta tarea es el análisis

de componentes principales (PCA) que realiza una rotación en el espacio de las características buscando encontrar aquellas direcciones en las que hay más varianza para descartar las con menos.

La metodología de PCA permite encontrar una base ortogonal para los datos centrados con un orden en particular, cuyo primer vector es la dirección en que las características tienen mayor varianza, el segundo es el con mayor varianza en el espacio ortogonal al espacio generado por el anterior, y así sucesivamente. Formalmente, bajo el supuesto de $\sum_i x_i^{(d)} = 0$, la primera componente satisface

$$w_{(1)} = \arg\max \frac{w^T X^T X w}{w^T w},$$

y definiendo $X_k^* = X - \sum_{s=1}^{k-1} X w_{(s)} w_{(s)}^T$, la componente k-ésima satisface la relación

$$w_{(k)} = \arg\max \frac{w^T {X_k^*}^T X_k^* w}{w^T w}.$$

El resultado de esto es una lista de vectores $\{w_{(s)}\}_{s=1}^{d}$ que se corresponden con los vectores propios de la matriz de covarianza empírica

$$C = \frac{1}{n} X^T X,$$

con sus respectivos valores propios. Los valores propios cuantifican la varianza representada en esa dirección. Convencionalmente se utiliza esta matriz y no la de máxima verosimilitud de la sección 2.3.2.1, esto no cambia la base encontrada ni el orden, solo produce un reescalamiento de los valores propios.

Una extensión del método de PCA es el PCA basado en kernel (k-PCA). Para esto, el problema de encontrar la descomposición en valores y vectores propios

$$\lambda v = C v$$

es reformulada como

$$\lambda x_i^T v = x_i^T C v, \quad \forall i \in [1, n].$$

Mediante este planteamiento se pueden aprovechar las operaciones basadas en productos internos para utilizar el método del kernel y encontrar las proyecciones en las componentes principales, aunque no necesariamente las componentes en sí.

2.3.3.2 Selección de características

La selección de características es, en principio, más simple que la extracción de características. Se basa en desechar características, esto puede ser, entre otros motivos, porque el problema tiene

demasiadas dimensiones y esto produce problemas por la maldición de la dimensión o porque se quieren seleccionar sólo las características más relevantes para posterior análisis. El segundo es el caso de este trabajo.

La selección puede realizarse con tres enfoques distintos no excluyentes, a saber: filtrado, wrapper y embebido. En el filtrado se eliminan características que se correlacionan pobremente con el valor o clase objetivo. Para esto, típicamente, se realiza algún test estadístico o se calcula alguna medida de información eliminando las características que no pasan cierto umbral o se mantienen solo un número fijo de características.

El enfoque de wrapper se basa en utilizar varios conjuntos de características para entrenar con un subconjunto de los datos y evaluarlo con otro, es importante que ninguno de estos conjuntos se utilice para la evaluación final del clasificador (para más detalle ver la sección 2.4). Al evaluarlo se puede comparar tanto el tamaño como el puntaje asociado para elegir el subconjunto de características final. Este método es computacionalmente más intensivo, por lo que se prefiere en el caso de clasificadores rápidos de entrenar, por otra parte, se obtienen mejores resultados que solo filtrando.

El tercer caso es el enfoque embebido o incrustado, esto porque la selección se hace a la par con el entrenamiento del clasificador. Ejemplo de esto son los clasificadores con representaciones ralas como los DTs de profundidad acotada o la regresión logística con regularización tipo LASSO.

2.3.4 Métricas de evaluación

Los clasificadores presentados son solo algunos de los métodos más utilizados en problemas de clasificación, cada uno con ventajas y desventajas, sin embargo, es necesario poder evaluar y compararlos según su éxito al momento de clasificar datos con los que no ha sido entrenado. Para esto se utilizan algunas de las medidas más comunes utilizadas tanto en problemas generales de clasificación como en problemas aplicados a ciencias de la salud. Para esto es útil introducir el concepto de matriz de confusión.

La matriz de confusión es una herramienta gráfica que permite analizar los resultados de un proceso de clasificación en la que se conocen las etiquetas previamente. Para esto se realiza una separación en cuatro segmentos cruzando los datos pertenecientes a una categoría y su etiquetado por el sistema de clasificación. Para mayor detalle ver figura 6.

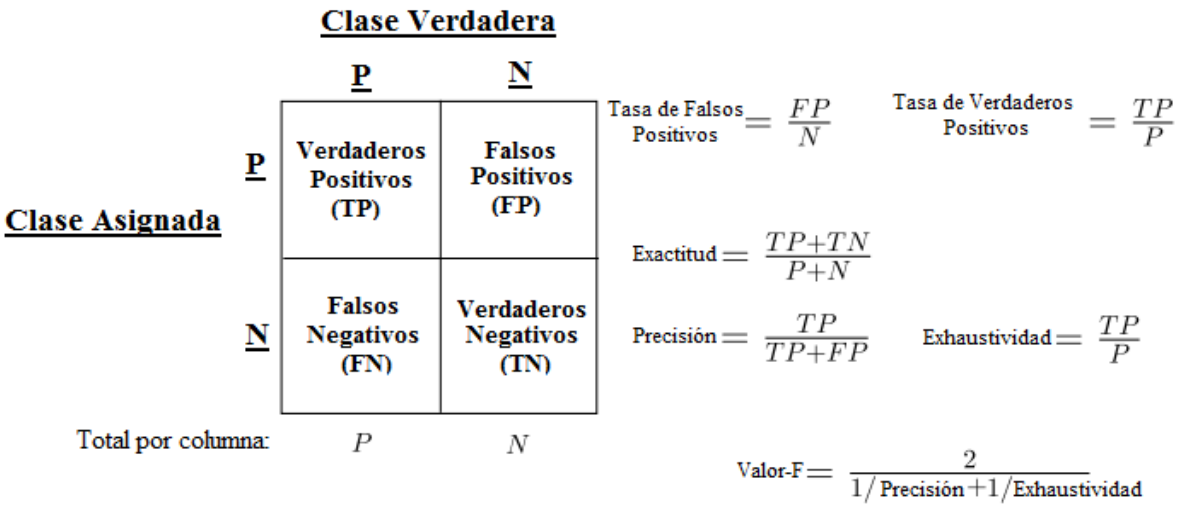


Figura 6: Esquema de matriz de confusión con algunas métricas. Traducción propia de (Fawcett, 2006)

2.3.4.1 Exactitud

La primera de las métricas es la exactitud (*accuracy* en inglés) que hace referencia a qué tan cercano a la realidad es la herramienta para clasificar. Esta se define como el porcentaje de datos clasificados correctamente, es decir

$$accuracy = \frac{TP + TN}{TP + FN + TN + FP}.$$

La exactitud es considerada una buena primera aproximación para evaluar modelos, sin embargo, es poco apropiada en múltiples contextos, especialmente cuando los conjuntos de datos etiquetados están desbalanceados.

2.3.4.2 Precisión, Exhaustividad y Valor-F (F1)

La precisión y exhaustividad son dos medidas que sirven para evaluar un clasificador y da información relevante respecto a la proporción de errores de clasificación hecho. Particularmente la precisión (*precision* en inglés) se define como

$$precision = \frac{TP}{TP + FP},$$

es decir, es el cociente entre los verdaderos positivos y todos los clasificados como positivos. Una alta precisión asegura que cuando se clasifica como positivo, es verdad con seguridad. Por otro lado, está la exhaustividad (*recall* en inglés) definida como

$$recall = \frac{TP}{TP + FN},$$

o sea, la proporción de positivos que fuero etiquetados como tal. En general, existe una solución de compromiso entre estas dos medidas en las que al mejorar una, la otra empeora. Por esta naturaleza complementaria de las medidas, se construye una medida en base a la media armónica de ambas, llamada Valor-F (también F1 o F-Score). Esta se calcula como

$$F1 = \frac{2\,precision * recall}{precision + recall}.$$

Para ilustrar estas métricas se puede considerar el siguiente ejemplo de juguete. Se tienen dos clasificadores binarios entrenados para determinar si una persona es más que un valor determinado. Los clasificadores se prueban en una población de 100 personas, en la que 30 tienen clasificación de alto obteniendo los siguientes resultados:

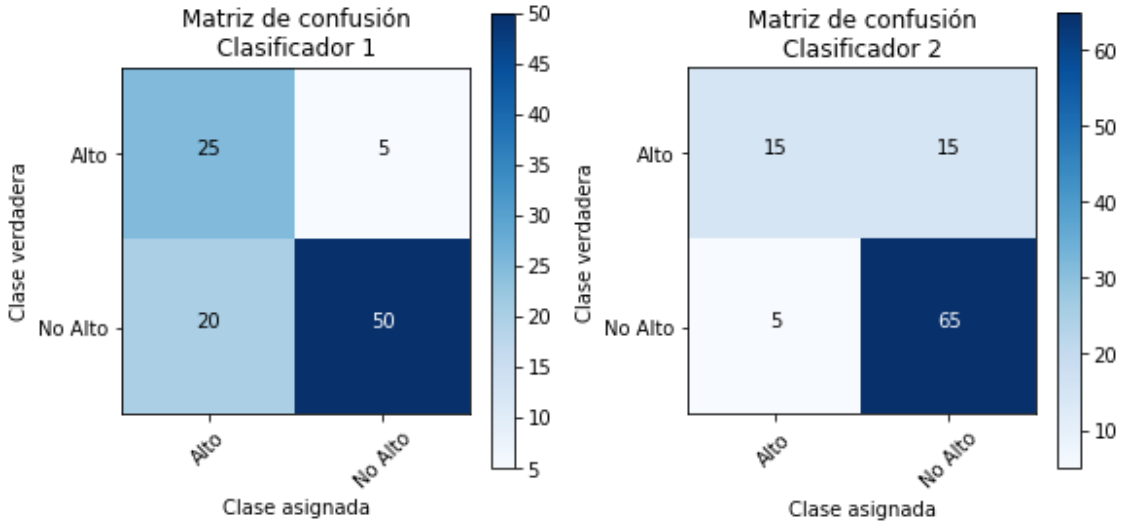


Figura 7: Matrices de confusión para un ejemplo de juguete

En este caso, el clasificador 2 obtiene una mayor precisión, pues de un 75% de los asignados como altos son efectivamente altos, frente a un 55% para el clasificador 1. Por otro lado, el clasificador logra encontrar a sobre el 80% de los altos de la población, mientras que el clasificador 2 tiene una exhaustividad del 50%. De estos valores se calcula el Valor-F de cada clasificador, expresado en la siguiente tabla.

	C 1	**C 2**
Exactitud	75%	80%
Exhaustividad	83.3%	50%
Precisión	55%	75%
F1	66.6%	60%

Tabla 1: Resumen de métricas para un ejemplo de juguete

La tabla refleja la complejidad inherente al comparar clasificadores en poblaciones desbalanceadas. Por un lado, se tiene un clasificador con mayor exactitud y precisión, pero que es

poco exhaustivo, por otro lado, se tiene un clasificador un poco menos exacto y preciso, pero que goza de una alta exhaustividad y con una mejor relación entre este y la precisión, reflejado en un mayor Valor-F. La decisión de cuál clasificador es mejor que el otro dependerá del contexto específico del problema.

2.3.4.3 Curva ROC y AUC

La curva ROC (por *receiver operating characteristic curve*) es una herramienta gráfica que se basa en comparar la tasa de verdaderos positivos contra la tasa de falsos positivos para un clasificador cuya salida entrega una forma de ranking asociado a la decisión, por ejemplo, la probabilidad en el método Naive Bayes induce un ranking. Esto también ocurre con la regresión logística, o en el caso de la SVM se tiene la distancia y dirección de esta al hiperplano separador.

De este gráfico se calcula el área bajo la curva ROC (AUC por sus siglas en inglés). Esta métrica es una forma resumida del poder discriminativo que tiene el clasificador. Una interpretación de esta métrica es que corresponde a la probabilidad de clasificar bien dos muestras sabiendo que son de clases distintas, formalmente

$$AUC = \mathbb{P}(f(i) = y(i), f(j) = y(j) \mid y(i) \neq y(j)) \ ,$$

donde $y(i)$ e $y(j)$ son las etiquetas reales de las muestras i y j respectivamente, y $f(i)$ y $f(j)$ son las etiquetas con las que se clasifican sabiendo que tienen etiquetas distintas. En este sentido, una línea recta entre el origen y el fin del gráfico daría un AUC de 0.5, es decir que la probabilidad de realizar bien la discriminación es análoga a decidir mediante una moneda equilibrada.

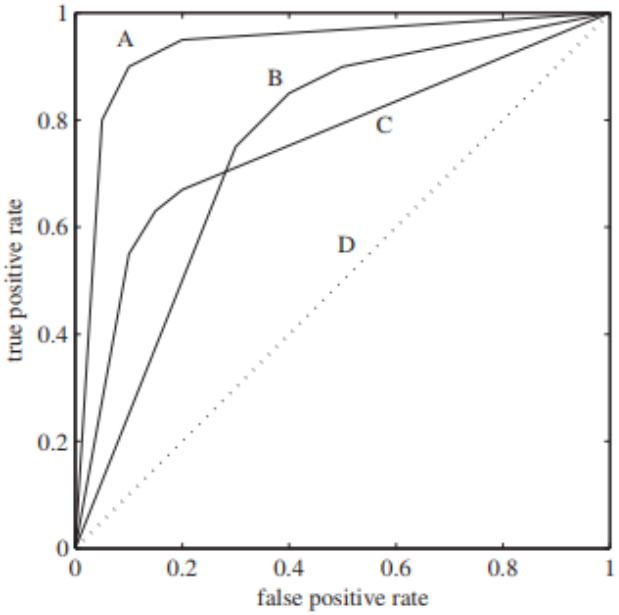


Figura 8: Curvas ROC. Obtenidas de (Majnik & Bosnic, 2013)

2.3.4.4 Curva de precisión-exhaustividad

La curva de precisión-exhaustividad es una herramienta gráfica que permite comprender a mayor profundidad un clasificador al comprarlo para distintos valores de precisión y exhaustividad. Para su construcción se requiero lo mismo que para la de la curva ROC, decir, la salida de un clasificador en forma de ranking. Según (Saito & Rehmsmeier, 2015) el análisis gráfico de esta curva es más informativo que la curva ROC en el caso de datos desbalanceados.

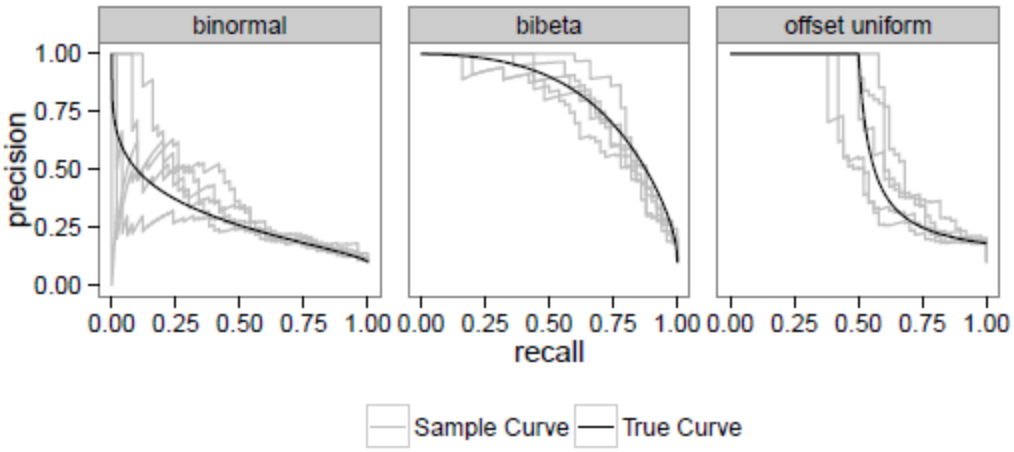


Figura 9: Curvas precisión-exhaustividad. Obtenidas de (Boyd, Eng, & Page, 2013)

2.3.5 Validación cruzada

Para evaluar modelos en aprendizaje de máquinas es necesario garantizar la independencia de los conjuntos de entrenamiento y test, sin embargo, realizar una separación dura puede comprometer la capacidad de generalizar adecuadamente. Las técnicas de validación cruzada son estrategias que pueden ser utilizadas tanto para ajustar (hiper-)parámetros de los modelos, como para garantizar independencia de los valores reportados en el desempeño de los clasificadores. La idea tras esto es que al promediar distintas evaluaciones de desempeño en realizaciones independientes se tiene una mejor estimación del desempeño teórico del clasificador. La idea principal es particionar los datos en dos conjuntos múltiples veces, entrenar con uno y evaluar con el otro. Estos desempeños luego se promedian para entregar una medida metodológicamente independiente sobre el clasificador.

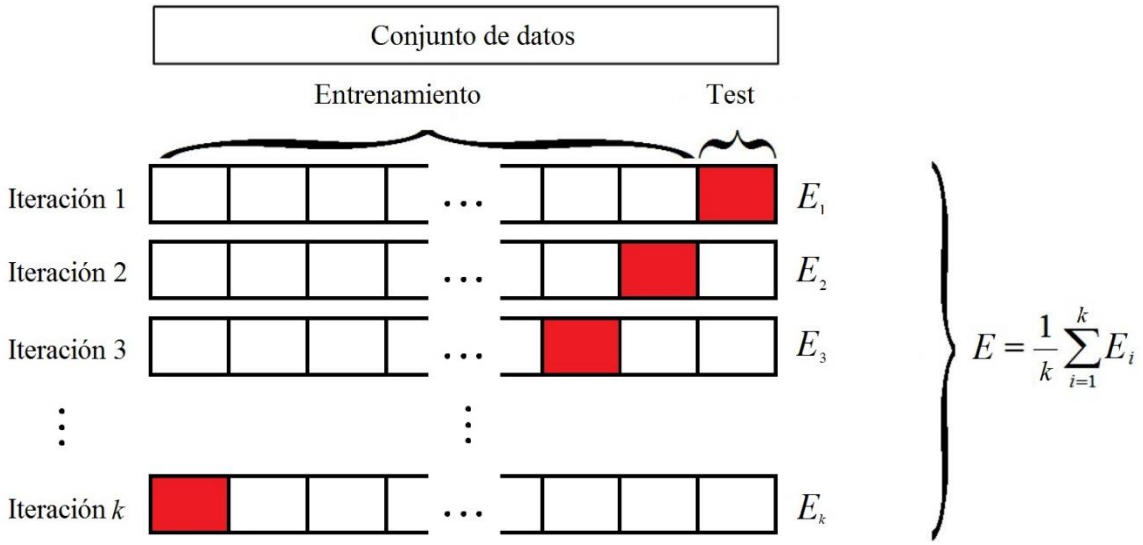


Figura 10: Esquema K-Fold para el cálculo de una métrica E.

El tipo de validación cruzada más comúnmente usado es el de K-iteraciones o K-fold. Este método particiona el conjunto de datos en K subconjuntos equilibrados, y en K iteraciones entrena con $K-1$ subconjuntos y evalúa con el restante. Un caso extremo de este método es cuando K es igual al número de datos. Este caso se denomina validación cruzada dejando uno fuera (LOO por sus siglas en inglés). En el caso LOO se itera n veces, donde n es el número de datos, y se evalúa la clasificación solo en el elemento dejado fuera. Esto puede utilizarse tanto en contextos de clasificación dura (ejemplo, SVM) como en clasificación blanda o basada en puntaje (ejemplo, NB).

2.3.6 Señal estacionaria

Para justificar la extracción de características realizado es necesario referirse a las señales desde un punto de vista más general. En el contexto de señales, estas pueden verse como un proceso estocástico indexado por el tiempo $\{X_t\}_{t\in I}$. Este proceso estocástico se denomina estacionario si es invariante en el tiempo, formalmente, si denotamos $F_X(x_{t_1},\ x_{t_2},\dots,x_{t_k})$ como la distribución conjunta de X_t en los tiempos $t_1, t_2, \dots, t_k$, la el proceso estocástico $\{X_t\}_{t\in I}$ se dirá estacionario si

$$F_X\left(x_{t_1},\ x_{t_2},\dots,x_{t_k}\right) = F_X\left(x_{t_1+\tau},\ x_{t_2+\tau},\dots,x_{t_k+\tau}\right)\ \ (\forall(\tau+t_i)\in I)(\forall i\in\{1,\dots,k\}).$$

Esto es relevante ya que al estimar las características de los EEG se trabaja bajo la suposición de estacionariedad de las señales cortadas en ventanas, lo que permite utilizar medidas centrales como estimadores de la característica en cuestión.

2.3.7 Medidas de tendencia central como estimador

Se debe buscar un descriptor relevante que resuma la información contenida en múltiples ventanas a las que se les calculan las mismas características. Al considerar las señales como

estacionarias en breves periodos de tiempo, podemos tomar medidas de tendencia central como estimados del valor real. Así, para cada sujeto en uno de los casos de estudio, para todas las características calculadas en todas las ventanas, se toma la media muestral y se considerará ese valor como la característica. Se hace una diferencia con las matrices de covarianzas, a las que se calcula la media geométrica riemmaniana de las matrices semidefinidas positivas. Se elige esta media en particular pues en las matrices de covarianza, el determinante es una medida de dispersión (o volumen generalizado), y como el determinante de la media geométrica es la media geométrica de los determinantes se prefiere sobre la media muestral.

Los elementos fuera de rango (también conocidos como outliers) son valores numéricos anormales que se presentan en los datos. Estos pueden generar grandes errores al ser considerados. En el esquema propuesto de media como estimador se está suponiendo que los datos no presentan estas anomalías, sin embargo, para uno de los casos de estudio esta suposición no es correcta, por lo que se prefiere la mediana, al ser un estimador de tendencia central robusto a elementos fuera de rango.

2.3.8 Método del kernel

Dado un conjunto no vacío $\mathcal{X}$, una función simétrica $K: X \times \mathcal{X} \to \mathbb{R}$ se llamará kernel si $\forall n \in \mathbb{N}, \forall x \in \mathcal{X}^n, \forall c \in \mathbb{R}^n$:

$$\sum_{i}^{n}\sum_{j}^{n} c_i c_j K(x_i, x_j) \geq 0.$$

Este tipo de funciones es considerado una generalización de matrices, pues otra forma de verlo es que $\forall x \in X^n$, la matriz M definida por $M_{i,j} = K(x_i, x_j)$ es semidefinida positiva.

Una forma de construir un kernel es a través de un mapeo $\Phi: X \to \mathcal{H}$, donde $\mathcal{H}$ es un espacio de Hilbert en el que los funcionales de evaluación son continuos. Bajo estas hipótesis, basta definir la función $K(x, y) =< \Phi(x), \Phi(y) >_{\mathcal{H}}$, con $< \cdot, \cdot >_{\mathcal{H}}$ el producto bilineal de $\mathcal{H}$, y se tiene un kernel válido. Pero, tal vez, más importante es que también se cumple una forma recíproca de esto. Dado un kernel K, este induce un espacio de Hilbert $\mathcal{H}$ en el que los funcionales de evaluación son continuos y un mapeo de $\Phi: \mathcal{X} \to \mathcal{H}$, tal que $K(x, y) =< \Phi(x), \Phi(y) >_{\mathcal{H}}$.

La utilidad que se le da a este tipo de funciones en aprendizaje de máquinas tiene que ver con los métodos en los que las soluciones quedan expresadas en formas de productos internos. Si se tiene una solución que queda expresada de esta forma, como en el caso de las SVM, basta ver que la solución es idéntica tanto si se hace para las características originales x, o para la imagen de estas $\Phi(x)$. De esta forma, si se tiene un kernel K, se pueden sustituir todos estos productos internos por evaluaciones de la función y obtener el resultado como si se hubiera hecho en el espacio inducido por K.

Esto tiene dos ventajas, una tiene que ver con disminuir el número de operaciones realizados al evitar el cálculo explícito del mapeo $x \mapsto \Phi(x)$. La segunda es que se pueden utilizar mapeos a espacios de dimensión infinita que son imposibles de implementar computacionalmente, pero que

sí tienen un producto interno de forma cerrada que se puede calcular. El ejemplo más común de esta segunda ventaja es el kernel gaussiano.

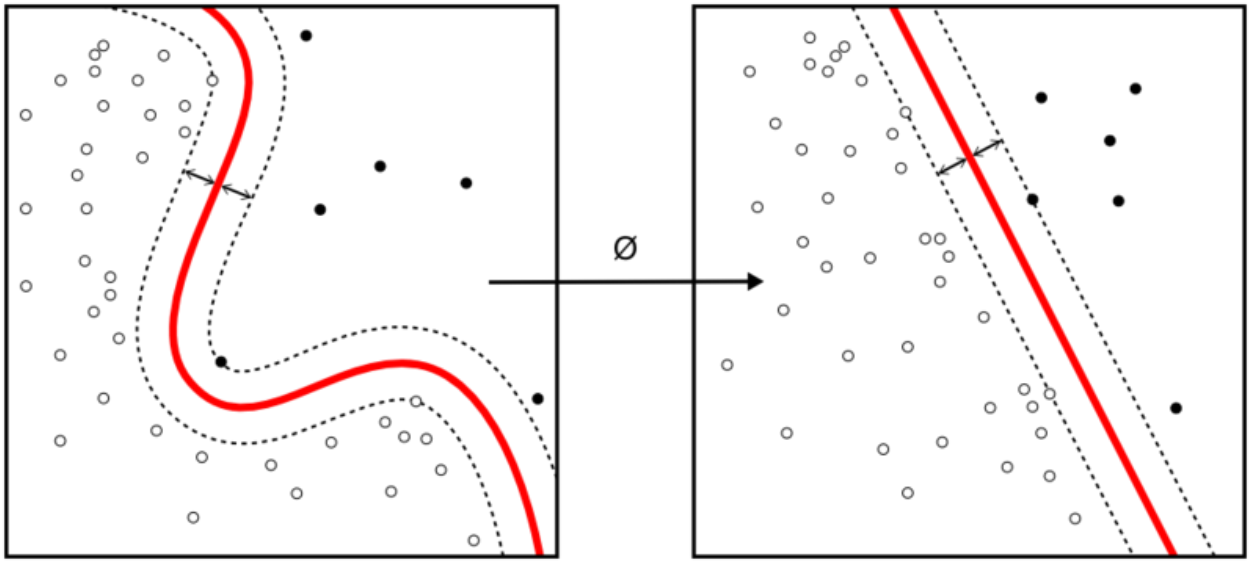


Figura 11: Separación de margen máximo utilizando un mapeo en el que los datos son linealmente separables. (Wikimedia Commons, 2016)

2.3.9 Test de hipótesis

Los test de hipótesis son herramientas de la inferencia estadística que buscan dar respuesta a la pregunta por alguna característica de una muestra o población, al modelarlas como variables aleatorias. Una de las preguntas más comunes es si dos muestras son estadísticamente similares o si, por el contrario, es más probable que vengan de poblaciones distintas.

Comúnmente, para la realización del test es necesario especificar las hipótesis, para esto se tiene la hipótesis nula H_0, que suele expresar la forma más simple del caso, por ejemplo, que dos muestras sean de la misma población, pues es más complejo que exista más de una. La hipótesis alternativa H_1 expresa lo que se quiere mostrar como estadísticamente más factible. Es importante enfatizar que estos procedimientos en ningún momento demuestras que H_0 ni H_1 sean ciertas, y es solo una forma de comparar dos hipótesis.

Para rechazar la hipótesis nula es necesario aplicar algún test adecuado al contexto, es decir, que el contexto en el que se implementa cumpla con las hipótesis para que el test sea válido. Del test se extrae típicamente un estadístico asociado a la prueba y su valor de p. El valor de p corresponde a la probabilidad de haber obtenido resultados como los observados o resultados más extremos (en contra de H_0). El significado de más extremo depende del test que se esté practicando. Si el valor de p es menor a cierto nivel de significancia α, entonces se dirá que se ha conseguido significancia estadística a nivel α.

2.3.9.1 Test U de Mann-Whitney

El test U de Mann-Whitney es un test estadístico no paramétrico que se utiliza para evaluar si una de dos poblaciones es estocásticamente mayor que otra. La hipótesis nula (H_0) del test corresponde a que $\mathbb{P}(X > Y) = \mathbb{P}(X < Y)$, en donde X, Y son variables aleatorias provenientes de la primera y la segunda población respectivamente. La hipótesis alternativa puede tomar dos formas según si se plantea como una prueba de dos colas o unilateral. Para el caso unilateral la hipótesis alternativa (H_1) es que $\mathbb{P}(X > Y) < \mathbb{P}(X < Y)$, o, en otros términos, que Y es estadísticamente mayor que X.

Los supuestos que se requieren para el test son que:

- Las observaciones de ambas muestras deben ser independientes.
- Las variables observadas deben ser de naturaleza continua u ordinal.

Para realizar el contraste de hipótesis se calcula es estadístico U, definido como

$$U = \sum_{n=1}^{N} \sum_{m=1}^{M} \mathbb{1}_{\{X_n < Y_n\}} + \frac{\mathbb{1}_{\{X_n = Y_n\}}}{2},$$

donde N y M son los tamaños de la primera y la segunda población respectivamente. Una vez calculado U, se compara el valor obtenido con la distribución del estadístico bajo la hipótesis nula para calcular el valor de p.

El estadístico U tiene una relación funcional con la métrica de evaluación AUC que toma la siguiente forma:

$$AUC = \frac{U}{NM}\ .$$

2.4 Revisión bibliográfica

Las técnicas de electroencefalografía cuantitativa tienen poca aceptación por el mundo clínico a excepción de algunas aplicaciones en epilepsia, a pesar de esto han sido utilizadas para la investigación de detección y clasificación de diversas condiciones, entre ellas trastornos psiquiátricos.

En la literatura científica se encuentran múltiples estudios que buscan utilizar mediciones de EEG para clasificar psicopatologías psiquiátricas estos tienen distintos grados de desarrollo y éxito. Uno de los más completos se puede encontrar en (Green, 2012), que realiza un estudio con pacientes con esquizofrenia y sanos como control. Realiza un estudio para seleccionar las características más relevantes e introduce el uso de RPDE para clasificación de esquizofrenia descrito en el apartado de ingeniería de características. La población que utiliza es únicamente de adultos, que pueden estar con tratamiento estable o sin tratamiento farmacológico, los datos son balanceados entre pacientes con esquizofrenia y adultos sanos y en total son 74 individuos. El estudio se centra en la métrica de exactitud, aunque también se reportan la sensibilidad (o exhaustividad) y la especificidad (o tasa de verdaderos negativos). Emplea el método de k-SVM

para realizar la clasificación. Dentro de los resultados se tiene que logra una clasificación con exactitud del 86%, con sensibilidad del 81% y especificidad del 91%. Las características más relevantes encontradas corresponden a coeficientes autoregresivos, RPDE, SpE y KZ, las que son calculadas en ventanas de 4 segundos y luego promediadas.

Otro como en (Khodayari-Rostamabad, Reilly, Hasey, MacCrimmon, & others, 2010) que estudian el caso de distinción entre cuatro clases: pacientes con esquizofrenia, trastorno depresivo mayor, bipolaridad y población sana de control. Si bien no realiza una prueba con las cuatro clases a la vez, en los resultados de a pares logran una exactitud mayor al 85%, sin embargo, para esto realizan un paso metodológicamente erróneo, que es seleccionar las características a utilizar antes de dividir el conjunto de datos en datos de entrenamiento y test, pudiendo comprometer así sus resultados. En este caso es lamentable que no se presenten cuáles fueron las características seleccionadas para la clasificación.

Por otro lado, hay estudios como (Li & Fan, 2006) que de características utilizan la energía asociada a los ritmos alfa, beta, delta y theta para 16 electrodos, hacen una reducción de dimensionalidad con PCA también antes de separar conjuntos de entrenamiento y test para luego entrenar una red neuronal artificial prealimentada (feed-forward) y un mapa auto-organizante (self-organizing map), un tipo de red neuronal competitiva. Los resultados son engañosos, en la mayoría de los casos tienen una precisión del 60%, que calculan al evaluar con cinco muestras.

Por otro lado, (Alimardani, Boostani, & Taghavi, 2015) proponen un método que utiliza la media geométrica riemanniana sobre las matrices de covarianzas utilizando una métrica en este espacio para determinar ventanas ruidosas y emplea esta distancia para dos variaciones del método de vecino más cercano. En este caso se busca diferenciar 27 pacientes con trastorno bipolar y 26 con trastorno de esquizofrenia, se utilizaron ventanas de uno y dos segundos para calcular las matrices de covarianza. Los resultados son muy buenos, sobre el 90% de exactitud empleando validación cruzada LOO.

En (Hasey, 2013) se realiza una revisión de la literatura resiente en técnicas electroencefalográficas para estudiar distintos ámbitos de la esquizofrenia, entre ellos menciona algunos trabajos realizados en estados de conciencia y relajo. Entre los resultados destacados se encuentran que, para los ritmos alfa y beta, existe mayor correlación cruzada para tiempos de desfase desde 5 segundos a los 50. Otro estudio según esta revisión encuentra ritmos gamma más altos en pacientes, algo que no se presenta en los sujetos de control ni en familiares, también encuentran ritmos más altos entre las theta y alfas, pero esto es tanto para los pacientes como para sus familiares, por lo que puede provenir de una expresión genética que predisponga a la esquizofrenia. En otros resultados no asociados a trabajos realizados en conciencia y relajo, se destaca uno en el que se logra clasificar con un 85% de exactitud a pacientes según si responden bien o mal a la clozapina.

Capítulo 3: Metodología

La metodología se basa en el proceso de ciencia de los datos descrito en el capítulo anterior y que lo divide en:

i. Seleccionar el objetivo de la investigación
ii. Recuperación de datos
iii. Preparación de datos
iv. Exploración de los datos
v. Modelamiento de los datos
vi. Presentación y automatización

Los experimentos llevados a cabo son de dos trasfondos distintos, pero con una base en común: poder diferenciar poblaciones utilizando muestras de EEG tomadas en reposo.

El primer experimento se trata de poder diferenciar dos grupos de adolescentes rusos según si presentan o no síntomas de esquizofrenia. Esta base de datos fue compilada por investigadores del laboratorio de neurofisiología e interfaces neuro-computador (NNCI) de la Universidad Estatal M. V. Lomonósov de Moscú, y hecha pública a través de su página web (NNCI, s.f.).

El segundo experimento es de diseño del autor, y se basa en usar técnicas similares para diferenciar adultos diagnosticados de esquizofrenia y adultos con diagnóstico de otros trastornos psiquiátricos. Esta base fue compilada por el autor en el marco de un estudio realizado durante el año 2017 en el Hospital del día (HdD) del Hospital Barros Luco-Trudeau, dependiente del Servicio de Salud Metropolitano Sur, Región Metropolitana, Chile.

3.1 Hipótesis

En el contexto del proceso de ciencia de los datos, la formulación de la hipótesis representa el "*¿Qué?*" de *Seleccionar el objetivo de la investigación*. Las hipótesis de trabajo son que las herramientas de aprendizaje de máquinas permiten encontrar diferencias en ciertas poblaciones de sujetos a partir de grabaciones de EEG en reposo. En este caso, para el primer experimento son:

H_0^A: Las herramientas de aprendizaje de máquinas no permiten encontrar una diferencia estadística en dos grupos de adolescentes rusos según si presentan o no síntomas de esquizofrenia, a partir de grabaciones de EEG en reposo.

H_1^A: Las herramientas de aprendizaje de máquinas permiten encontrar una diferencia estadística en dos grupos de adolescentes rusos según si presentan o no síntomas de esquizofrenia, a partir de grabaciones de EEG en reposo.

Para el segundo experimento las hipótesis declaradas son:

H_0^B: Las herramientas de aprendizaje de máquinas no permiten encontrar una diferencia estadística en dos grupos de adultos chilenos en tratamiento farmacológico según si presentan esquizofrenia o si presentan otro trastorno psiquiátrico, a partir de grabaciones de EEG en reposo.

H_1^B: Las herramientas de aprendizaje de máquinas permiten encontrar una diferencia estadística en dos grupos de adultos chilenos en tratamiento farmacológico según si presentan esquizofrenia o si presentan otro trastorno psiquiátrico, a partir de grabaciones de EEG en reposo.

3.2 Diseño experimental

Para la contrastación de las hipótesis es necesario diseñar, implementar y obtener los datos y herramientas necesarias para llevar adelante el experimento. A continuación, se describe el diseño experimental.

3.3 Recursos

La obtención y procesamiento de datos requiere instrumental especializado para esto. Para la reproducibilidad del experimento es importante establecer cuáles son los recursos utilizados para llevarlo a cabo.

3.3.1 Datos del NNCI

El primero son los datos obtenidos por el NNCI, descargados desde la página web del instituto. Los datos corresponden a dos grupos de archivos electrónicos, el primero, que involucra muestras de EEG de 45 niños con diagnóstico de trastornos esquizofrénicos de síntomas similares (esquizofrenia infantil, trastorno esquizotípico y esquizoafectivo); y otra con muestras de EEG de 39 niños sanos. Las edades van de los diez años ocho meses a los 14 años y el promedio de edad es de doce años y tres meses en ambos casos. Ninguno de los pacientes recibía terapia farmacológica durante el período de toma de datos ni examinación por parte de especialistas (Borisov, Kaplan, Gorbachevskaya, & Kozlova, 2005).

Los EEG fueron tomados en estado de conciencia, relajo y con los ojos cerrados. Las grabaciones corresponden a potenciales eléctricos medidos a través de 16 canales, ubicados según el sistema internacional 10-20 en las posiciones $O_1, O_2, P_3, P_4, P_z, T_5, T_6, C_3, C_4, C_z,\ T_3, T_4, F_3, F_4, F_7$ y F_8, y con referencia monopolar a ambos lóbulos de las orejas. La frecuencia de muestreo es de 128 [Hz], la duración es de un minuto de grabación y se le ha realizado una selección de tramos libres de artefactos. Esto último aligera mucho el preprocesamiento de los datos, pues la remoción de artefactos no es tarea sencilla y es un tópico que da para tesis completas. No se tienen más detalles sobre la toma de estos datos.

3.3.2 Datos del HdD

Para la construcción de la segunda base de datos se utilizan las dependencias del HdD. Los datos fueron obtenidos mediante el hardware Emotiv Epoc descrito en el capítulo de marco teórico y con un computador portátil Samsung Q470C/500P4C para la grabación de los EEG. En este caso, se cuenta con 39 participantes, de los cuales 30 tiene diagnóstico de esquizofrenia y 9 otro diagnóstico

psiquiátrico. Todos mayores de edad y ambos sexos están representados las poblaciones. Los sujetos estaban en tratamiento farmacológico al momento de la toma de datos.

Para este estudio no se utilizan sujetos sin diagnóstico psiquiátrico por las siguientes consideraciones metodológicas. Para realizar una toma adecuada de datos es necesario que todos los sujetos estén en condiciones suficientemente similares al momento de la realización de las grabaciones. Una de las restricciones de realizar la prueba es que esta debía tomarse en dependencias del HdD, lo que condiciona fuertemente la toma de datos a personas que no estuvieran familiarizadas con el entorno, pues el estrés modifica, al menos, la potencia relativa de las ondas β (Seo & Lee, 2010) por lo que una distinción entre estas poblaciones podría deberse a esto y no a algún otro proceso endógeno.

Los EEG fueron tomados en estado de conciencia, relajo, y se tomaron muestras con ojos cerrados y con ojos abiertos a todos los participantes. Las grabaciones corresponden a potenciales eléctricos medidos a través de 14 canales, ubicados de manera aproximada según el sistema internacional 10-20 en las posiciones en las posiciones O_1, O_2, P_7, P_8, T_7, T_8, FC_5, FC_6, F_3, F_4, F_7, F_8, AF_3 y AF_4, y con referencia monopolar a las posiciones P_3 y P_4. La frecuencia de muestreo es de 128Hz y las grabaciones de tres minutos de duración, a las que se les desecha el primer minuto.

El trabajo con estas muestras presenta un desafío metodológico por tres razones. La primera es que los electrodos están en posiciones aproximadas, pues el hardware de adquisición tiene un tamaño estándar, por lo que no se respeta la proporción del sistema internacional 10-20. La siguiente dificultad proviene de la baja calidad de las muestras, estas presentan muchos artefactos que complejizan la ingeniería de características. Finalmente, el trabajar con un conjunto de datos de tamaño pequeño, desbalanceado y con muchas dimensiones resulta en un desafío técnico desde la perspectiva de la ciencia de los datos.

Se utiliza el software TestBench para la toma experimental de datos, Python para el preprocesamiento, ingeniería de características, exploración de datos, clasificación y selección de características, mientras que para el análisis estadístico se utiliza el software R.

3.4 Preprocesamiento

El preprocesamiento o la *Preparación de los datos* se realiza en varias etapas. La primera es seleccionar los segmentos de grabación con la menor cantidad de artefactos. Para la base de datos del NNCI no fue necesario realizar esto, pues los datos ya venían seleccionados y libres de artefactos. Para la selección en el caso de los datos del HdD, se seleccionaron los tramos menos ruidosos de al menos 1 segundo de duración y se seleccionan los canales que hay en común entre ambas bases de datos, quedando los canales O_1, O_2, F_3, F_4, F_7 y F_8. Esto se hizo mediante inspección visual.

Una vez seleccionados los segmentos y canales a utilizar, se utiliza una técnica de corte en ventanas. Se tomaron ventanas no disjuntas de 1 segundo de duración con una superposición de 0.5 segundos entre ventanas cuando fuera posible. En el caso de las grabaciones del NNCI, esto quiere decir que por sujeto se consiguen 119 grabaciones de un segundo. En el caso de los datos provenientes del HdD se utilizó esta estrategia en cortes continuos, a modo de ejemplo, si para un sujeto se tienen cortes seleccionados de 30, 20 y 40 segundos de duración, del primero se extraen

59, del segundo 39 y del tercero 79 cortes de un segundo. En la práctica, por simplicidad y estandarización, se tomaron los primeros 119 cortes de 1 segundo de grabación, aunque hubiera más disponibles.

A cada uno de estos segmentos se le estandariza -es decir, se deja con media cero y varianza 1-, y se filtran las frecuencias mayores o iguales a 50Hz mediante un filtro lineal utilizando la transformada discreta de Fourier. Finalmente, como prepocesamiento extra para la base de datos del HdD se utilizó una técnica de remosión de tendencias considerando los puntos inicial, medio y final.

Al finalizar el preprocesamiento, cada sujeto de las bases de datos tiene asociado 119 muestras de un segundo de duración, cada una con 128 mediciones de potencial eléctrico por cada uno de los seis canales seleccionados.

3.5 Exploración de los datos

La exploración de los datos inicia en la selección de los datos, pues para esto es necesaria la exploración visual y continúa en la de modelamiento de los datos con un análisis de reducción de dimensionalidad exploratorio para el caso de los datos del NNCI. Este se realiza con la metodología de k-PCA.

3.6 Modelamiento de los datos

3.6.1 Ingeniería de características

El modelamiento de los datos se inicia por la construcción de las características. Para esto, se toman las ventanas de 1 segundo y se calculan las características descritas en el marco teórico, estas son, matriz de covarianza (21 características), coeficientes autoregresivos (54 cararacterísticas), número de Lempel-Ziv (6 características), dimensión fractal de Katz (6 características), Entropía Espectral (36 características), RPDE (6 características), y ApEn (6 características). Una vez calculadas las 135 características para cada ventana se utiliza algún estadístico de tendencia central para regularizar sobre las observaciones. En el caso de las matrices de covarianzas, se toma el promedio geométrico riemmaniano sobre las matrices definidas positivas, y para el resto se utiliza el promedio aritmético en el caso de la base de datos NNCI y la mediana en el caso del HdD. La mediana se toma en el segundo caso debido a que es una medida poco sensible a elementos fuera de rango.

Característica Implementada	Cantidad total de características	Descripción
Media geométrica de matrices de covarianza (Cov)	21	Generalización de la media de matrices definidas positivas
Coeficientes autoregresivos de orden 8 (AR8)	54	Modelo para series estacionarias

Complejidad Lempel-Ziv (Lz)	6	Complejidad de compresión de series binarias
Dimensión fractal de Katz (Kz)	6	Aproximación de la dimensión fractal de la serie de tiempo
Entropía espectral multibanda (SpE)	36	Entropía de espectro de frecuencias completo y por bandas alfa, beta, gama, delta y theta
Entropía de la densidad del periodo de recurrencia (RPDE)	6	Medida distribución de recurrencia de la serie de tiempo
Entropía aproximada (ApEn)	6	Medida de repetitividad de la serie de tiempo

Tabla 2: Características implementadas con número total y breve descripción.

Debido a la baja cantidad de datos y desbalanceo de las muestras provenientes del HdD, se utiliza un esquema de reducción de dimensionalidad y extracción de características mediante k-PCA. La reducción de dimensionalidad se hace encontrando primero las familias más relevantes para el caso del NNCI y sobre estar realizar la extracción de características. La cantidad de dimensiones de decide por recuperación de varianza mínima al 95%.

3.7 Modelos de clasificación

El modelamiento de clasificación se realiza con los métodos de k-SVM, Bayes ingenuo con distribuciones gaussianas, Regresión logística y bosques aleatorios. A partir de estos clasificados se obtendrán los resultados para contrastar la hipótesis de investigación.

Como producto paralelo se buscan las características más relevantes mediante métodos de selección embebidos. Para esto se utilizan los modelos de regresión logística con regularización tipo LASSO y modelos de árboles de decisión con profundidad máxima de tres. Dado que la cantidad de características máximas posibles a escoger por un árbol de decisión de profundidad tres son $2^3 - 1 = 7$, se afina el modelo tipo LASSO para que mantenga una cantidad similar de características. Los métodos de selección de características se emplean solo en los datos del NNCI.

Finalmente, el diagnóstico de los modelos se realiza mediante la comparación de cinco medidas y mediante el estadístico U y su p-valor. Las cinco medias utilizadas son exactitud, precisión, exhaustividad, valor-F y AUC. De estas, todas a excepción del AUC son dependientes del umbral elegido para clasificar, por lo que se ha decidido comparar los valores para cuando se maximiza el valor-F como medida central entre la precisión y la exhaustividad.

El test de hipótesis se realiza sobre el puntaje o probabilidad que asocian los métodos. Dada esta medida continua, se puede realizar un test de hipótesis U de Mann-Whitney para dos poblaciones. Para esto se realiza el cálculo a través del paquete *stats* del software R, pues la implementación permite calcular el p-valor exacto.

3.8 Otras consideraciones

La validación de esta metodología se hace mediante validación cruzada tipo LOO. Para poder tener una descripción más compleja, y validar estadísticamente la separación, se busca que los clasificadores generen un puntaje con valores continuos, de los clasificadores escogidos el único que no realiza esto en su forma clásica es la SVM, sin embargo, se usa la función de decisión para este fin. (SKLEAR [dot] com / SVC Notes probability).

Un último comentario con respecto a la metodología es que muchas de las técnicas empleadas requieren de la selección de parámetros, ya sea el coeficiente de regularización en el caso del método de LR o los hiperparámetros del kernel en los métodos basados en éstos. A menos que se indique lo contrario, los parámetros utilizados son los que emplea por defecto la implementación del método correspondiente en el paquete scikit-learn 0.19.1 para Python 3. Estos parámetros podrían ser afinados para obtener mejores resultados en los datos con los que se trabaja, sin embargo, esto hubiera involucrado necesariamente la disminución de datos de entrenamiento para poder realizar la búsqueda de manera independiente y que los resultados fueran metodológicamente válidos.

Capítulo 4: Resultados y Análisis

En este capítulo se exponen los resultados obtenidos tras seguir la metodología de ciencia de los datos. El orden sigue el descrito en el capítulo de metodología. Se inicia con la exploración gráfica de los datos del NNCI, para luego dar paso a los resultados de clasificación y selección de características. Posteriormente se exponen los resultados del proceso realizado en el caso de los datos del HdD tomando como punto de partida los aprendizajes del primer caso.

4.1 Exploración de datos

El primer resultado corresponde a la visualización obtenida mediante la metodología de k-PCA con kernel gaussiano. Para esto se utilizan las características calculadas de la base de datos del NNCI y se proyecta sobre las primeras dos componentes principales, es decir, se busca resumir en dos dimensiones la variabilidad de las 135 características calculadas para los 84 sujetos. De la figura (4.1) se puede inferir inmediatamente que métodos no lineales sobre este espacio podrían lograr una clasificación relativamente buena y da luces de cierta estructura subyacente de los datos, que pareciera separar naturalmente las dos poblaciones.

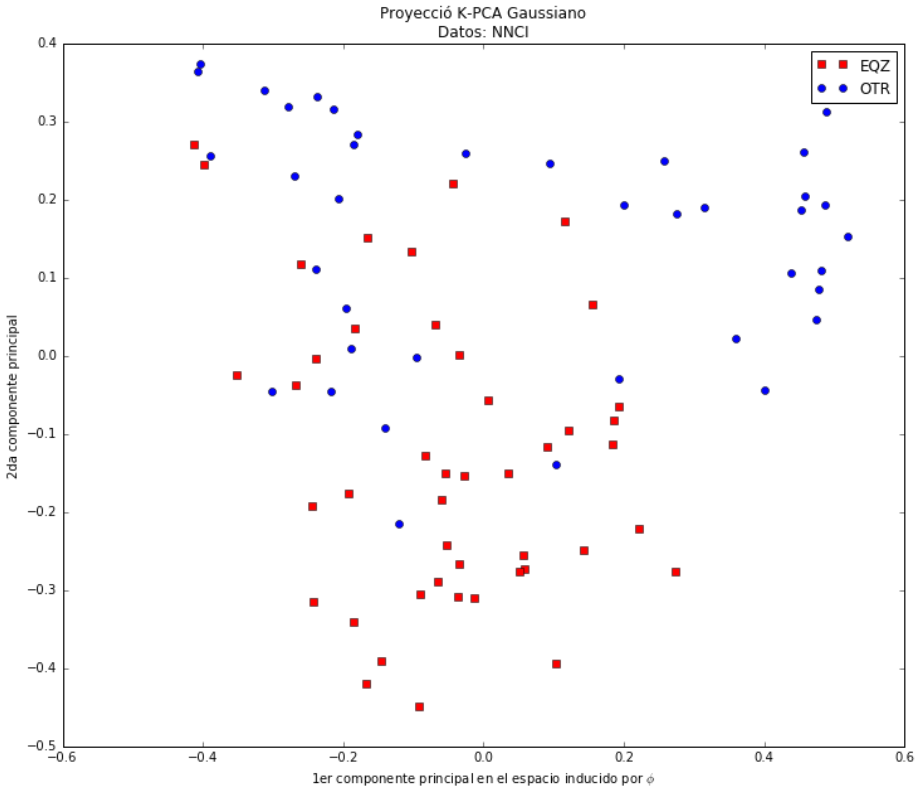


Figura 12: K-PCA (NNCI)

Es importante remarcar que este método es no supervisado, por lo que no utiliza información alguna sobre la etiqueta de cada muestra. En la figura 12, del porcentaje de varianza explicada para las primeras 20 componentes, se observa que, si bien las primeras involucran proporcionalmente más varianza, se requieren varias, de hecho, al llegar a las 20 componentes todavía no se ha recuperado ni el 60% de la varianza de los datos. Con esto se puede inferir que, o la estructura de los datos es en general bastante compleja o las características contienen mucho ruido y por eso se recupera de a tan poca varianza.

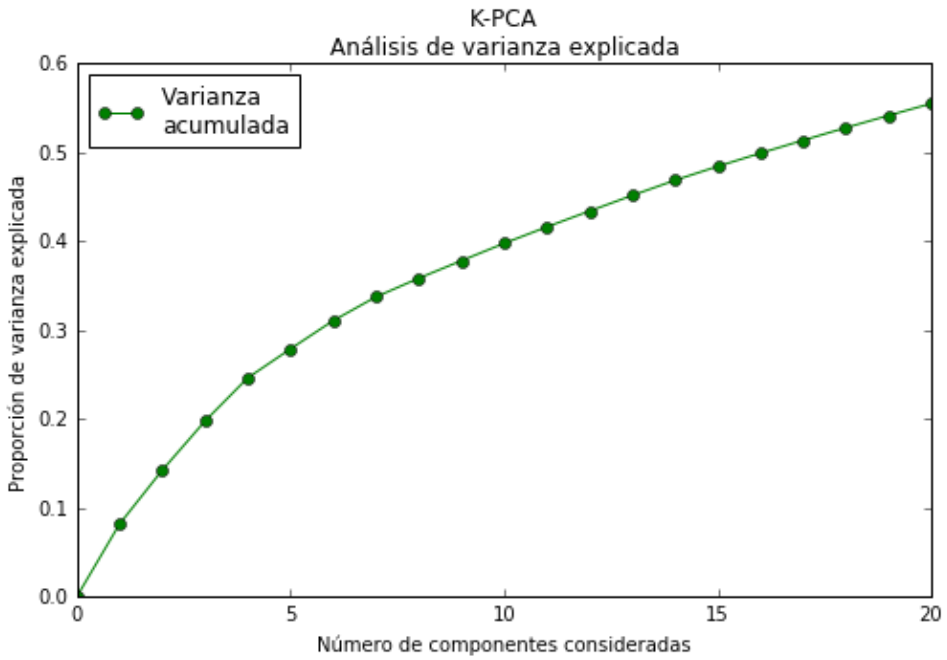


Figura 13: Varianza explicada por las primeras 20 componentes del K-PCA (NNCI)

4.2 Clasificación NNCI

La clasificación en el caso de los datos del NNCI se hace tomando como etiqueta positiva la observación compatible con esquizofrenia y negativo el grupo de control. La matriz de datos contempla las 135 características descritas en la metodología. Los modelos de clasificación utilizados son SVM, NB, LR y RF.

Para la SVM se utiliza un kernel de tipo gaussiano. En el caso de NB la única decisión de modelación es la forma que tienen las probabilidades marginales, en este caso se modelan todas como gaussianamente distribuidas. Para el caso del método de LR no se realiza ninguna consideración extra. Finalmente, para el método de RF se emplea la ganancia de información como función para encontrar los cortes en cada nodo. Para tener medidas más complejas sobre los clasificadores se utilizan herramientas que permitan no sólo dar una predicción, sino, además, una probabilidad o puntaje a esta. De los métodos utilizados, todos excepto SVM permiten tener esto

en forma de probabilidad, por lo que para este clasificador se utiliza el valor de la función de decisión como puntaje asociado.

Los resultados de clasificación para los distintos modelos se reportan en la tabla (4.1). De acuerdo con lo señalado en el capítulo de metodología, las métricas son calculadas según los valores que maximizan el valor F. De esta tabla se puede apreciar que los cuatro clasificadores obtienen resultados prometedores como métodos de clasificación, con los métodos de SVM y LR dominando en cuatro de las cinco métricas consideradas. Los resultados para estos dos clasificadores ejemplifican bien la solución de compromiso que existe entre la precisión y la exhaustividad, en este caso la SVM prioriza la precisión y la LR la Exhaustividad para maximizar la métrica F1.

	SVM	NB	LR	RF
Exactitud	89.28	84.52	88.09	85.71
AUC	90.02	86.32	92.36	87.80
F1	89.88	86.86	89.36	87.75
Precisión	90.90	79.62	85.71	81.13
Exhaustividad	88.88	95.55	93.33	95.55

Tabla 3: Métricas de evaluación (NNCI)

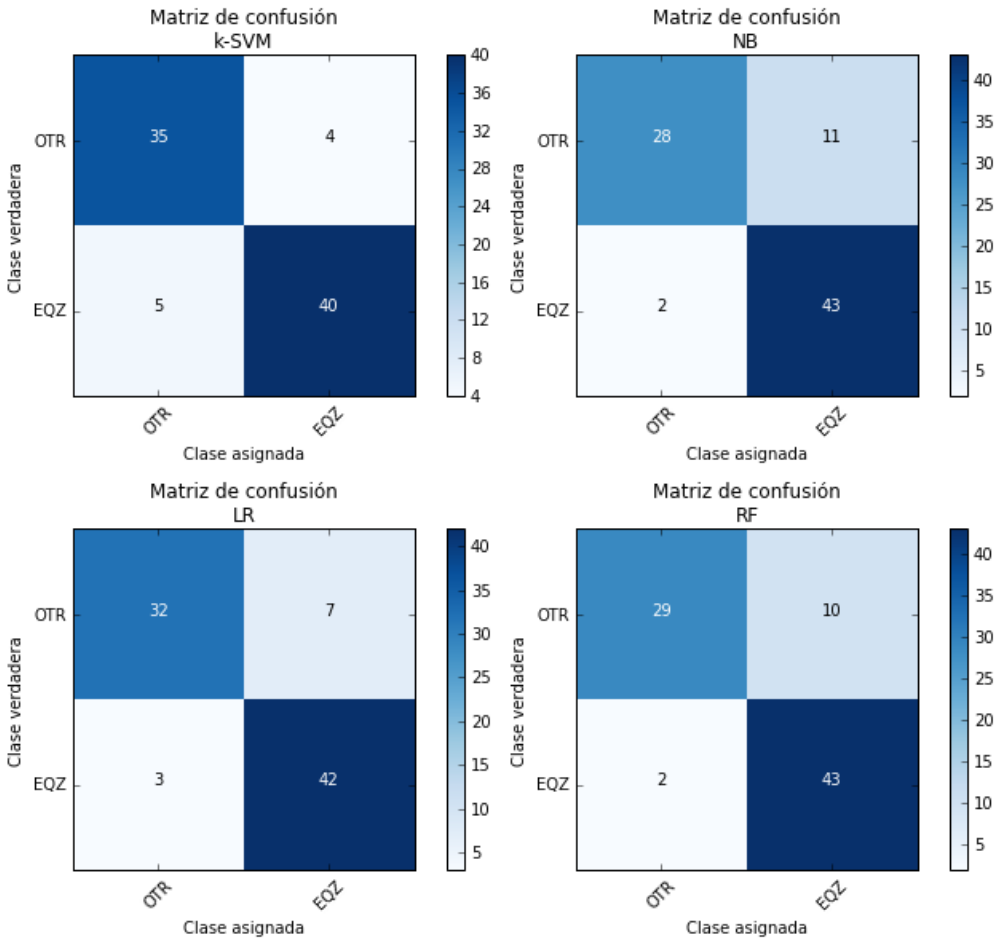


Figura 14: Matrices de confusión para los cuatro clasificadores considerados. (NNCI)

Las matrices de confusión permiten ir más a fondo en lo que respecta a este tipo de análisis. De estas, se desprende que todos los clasificadores, a excepción del SVM, cometen más errores de tipo II, es decir, incurren en más falsos positivos que falsos negativos, mientras que los otros parecieran privilegiar la exhaustividad. Este fenómeno puede darse por el ligero desbalanceo existente entre las clases.

Un análisis más profundo de los clasificadores se obtiene al comparar las curvas ROC y precisión-exhaustividad para los distintos clasificadores. Estas herramientas gráficas permiten diagnosticar regiones en las que algunos clasificadores pueden ser mejores que otros.

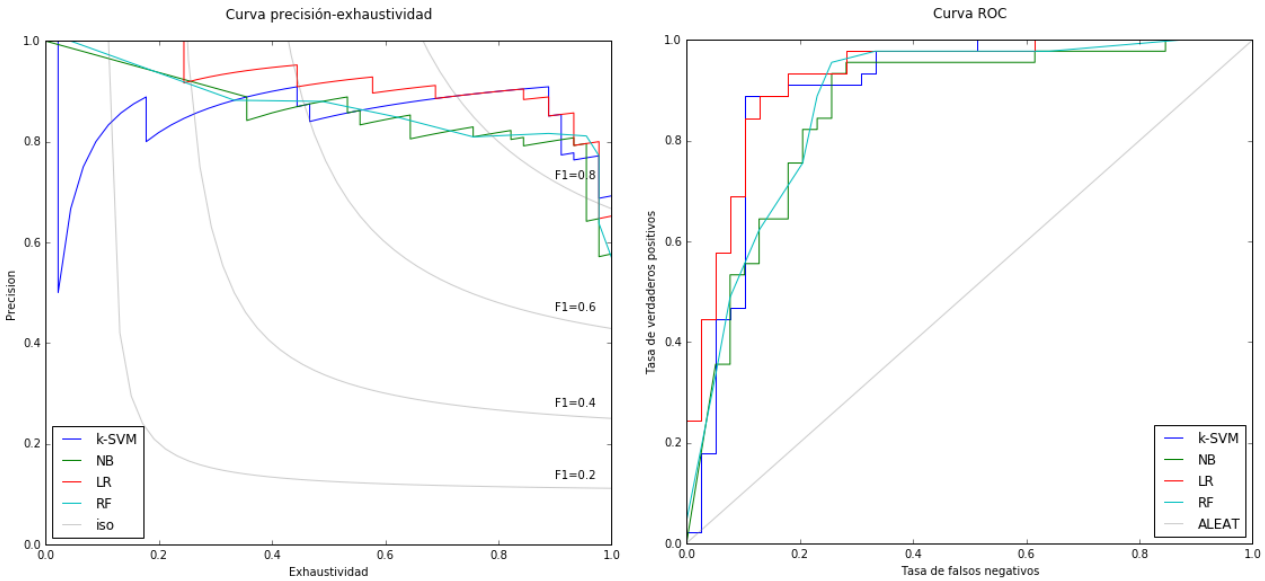


Figura 15: Izq: Curvas de precisión-exhaustividad para los cuatro clasificadores considerados. Der: Curvas ROC para los clasificadores en el caso NNCI.

De los gráficos podemos ver que el método de LR es dominante en casi toda región, salvo lugares específicos en los que es superado brevemente por la SVM. Esta característica, juntos a la interpretabilidad que tiene el vector de pesos que se aprende hacen que la LR sea una opción mucho más atractiva frente a la SVM, especialmente en contextos de salud, en que es relevante poder verificar cómo se están sopesando las características para tomar una decisión.

4.3 Validación estadística

La validación de los métodos se hace a través de un test de hipótesis bajo la prueba U de Mann-Whitney de un lado. Para esto se busca ver que la distribución de los valores asignados por los clasificadores no solo es una distribución distinta, sino que además toma valores mayores para el caso positivo.

La significancia estadística se alcanza al 0.1% para todas las metodologías de clasificación, lo que indica que hay evidencia fuerte para rechazar la hipótesis nula en los cuatro casos y que estos métodos efectivamente permiten discernir los dos tipos de poblaciones.

	SVM	NB	LR	RF
U	1580	1515	1621	1541
Valor de p	5.3 e-12	4.6 e-09	5.2 e-14	1.0 e-09

Tabla 4: Resultados del test de hipótesis para los clasificadores (NNCI)

4.4 Selección de características

La selección de características se hace realizando el mismo esquema de clasificación con validación cruzada tipo LOO, pero en este caso se busca no sólo validar el método de clasificación, sino que encontrar características relevantes para la tarea en cuestión. Para esto se utilizan clasificadores con representaciones ralas sobre las características, es decir, que están diseñados para utilizar pocas de las características a disposición. Para hacer el análisis de las características más relevantes, se cuentan cuántas veces se eligió cada una. En el análisis se canales relevantes se excluyen las covarianzas en el análisis pues no utilizan un único canal.

Los métodos utilizados son DT y LASSO. En este caso se fijan parámetros para lograr resultados similares. En este caso se fija en 3 la profundidad del árbol a aprender y por inspección se elige un valor de 0.13 para el inverso del coeficiente de regularización, esta decisión viene motivada para que la regularización tome alrededor de seis características distintas en promedio por entrenamiento.

Antes de presentar los resultados de selección, es importante presentar la validez de estos eventuales resultados. Para estos, se presentan las métricas antes propuestas para evaluar estos clasificadores y se realiza el mismo test de hipótesis sobre las salidas. Con esto se busca dar una base de confiabilidad a las familias de características escogidas.

	DT	LASSO
Exactitud	86.90	88.09
AUC	85.24	89.23
F1	88.17	89.13
Precisión	85.41	87.23
Exhaustividad	91.11	91.11
U	1496	1566
Valor de p	7.5 e-10	8.9 e-12

Tabla 5: Métricas de evaluación y resultados del test de hipótesis para los clasificadores con selección de características embebido.

Los resultados obtenidos por estos métodos de clasificación permiten dar confianza respecto a que las características seleccionadas son buenas discriminadoras para esta tarea. Como se ve en la tabla (4.4) los valores no son tan alejados de los obtenidos por los clasificadores al utilizar todas las características. Esto resulta particularmente potente cuando se toma en consideración que la

cantidad promedio de características seleccionadas por los métodos DT y LR son de 3.98 y de 6.04 respectivamente.

En general, las características seleccionadas por los métodos difieren, a excepción de una. Esto da para suponer que hay información redundante entre ellas. Metodológicamente se considera con no es la característica la relevante al ser escogida, sino la familia a la que pertenece. Esto puede ser considerado de dos maneras, la primera es que depende de cuál característica viene, es decir si es una covarianza o un coeficiente de modelos AR, o de cuál canal es calculada. Estas estadísticas se resumen en la figura 15. De estos gráficos se desprende que tanto los coeficientes AR como los valores de ApEn calculados resultan ser muy informativos. Por otro lado, la covarianza pareciera ser también importante, pero redundante tal vez con la SpE. Esto porque el método de DT elige un 97% de las veces al menos una característica de covarianza, cuando la LR sólo un 38%, mientras que en el caso de la SpE los porcentajes son 4% y 27% respectivamente. Las otras familias no parecieran ser relevantes en este caso.

En relación con el caso de la relevancia de los canales, resulta poco claro poder extraer una conclusión relevante además de que el canal 8 pareciera no ser relevante. Esto puede darse porque el resto ya es suficientemente informativo, sin embargo, no se descarta su utilización a priori. Es importante remarcar que como la estadística de los canales no considera las variables de covarianza, algunos canales pueden quedar infrarrepresentados.

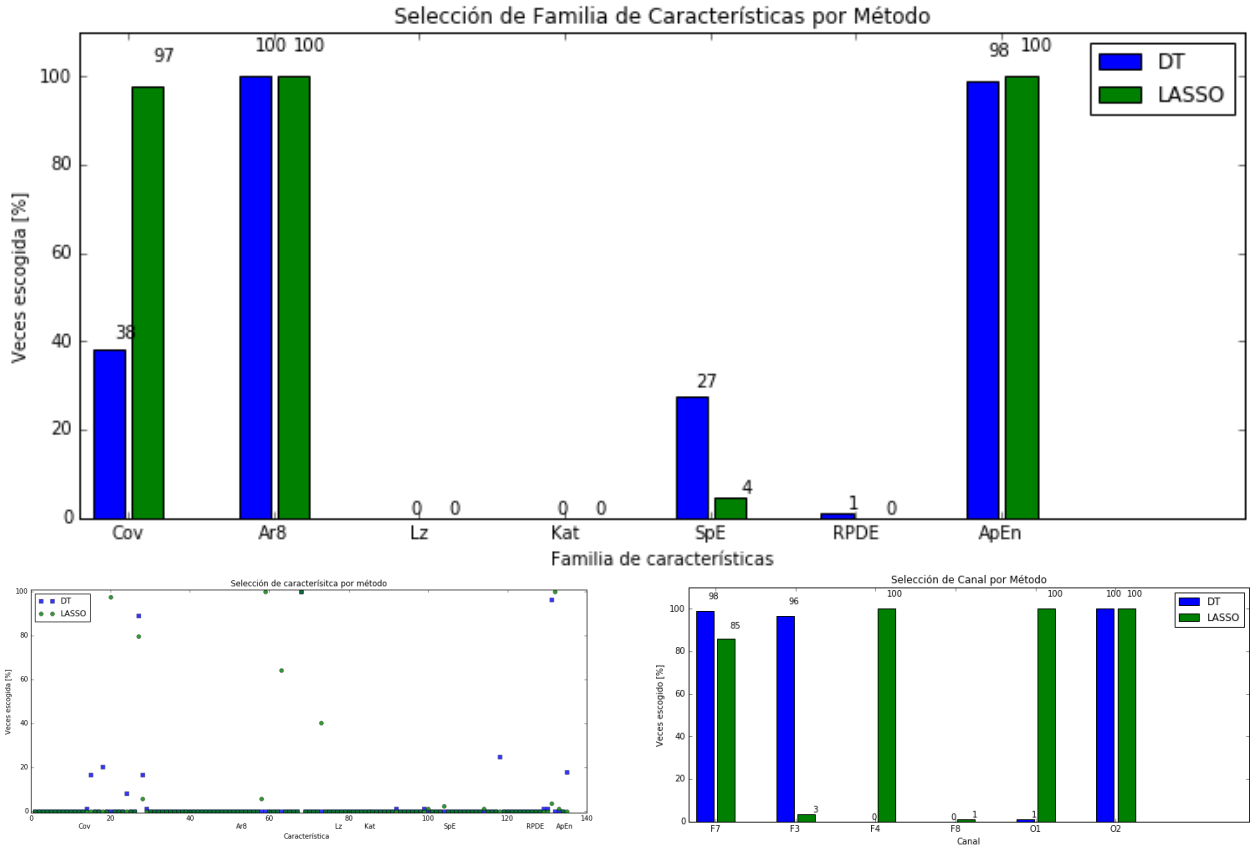


Figura 16: Arriba: Familias de características más seleccionadas por los métodos. Abajo izq.: Características individuales más seleccionadas. Abajo der.: Electrodos más seleccionados por las características individuales.

4.5 Clasificación HdD

El esquema de clasificación en el caso de los datos del HdD es diferente al hecho previamente. Las modificaciones realizadas vienen de dos partes, la primera es que el desbalance de los datos genera un desafío no suscitado en el caso anterior. La segunda es que se utiliza información de los resultados de la selección de características realizado para los datos del NNCI.

Para afrontar estos desafíos se seleccionan las características relevantes para el caso anterior, es decir, Cov, AR8, SpE y ApEn. Después de esto, la matriz de datos se compone de 39 entradas, cada una con 117 características, por lo que se decide realizar un esquema de extracción de características con la metodología de k-PCA, manteniendo un mínimo del 95% de la varianza explicada. Para conservar la independencia de cada clasificación, la extracción se realiza para cada iteración de la validación cruzada, y los datos a evaluar se proyectan sobre las componentes principales encontradas. A posteriori se tiene que la cantidad de variables que se extrajo fue de 21 en todos los casos, por lo que la matriz de datos efectiva tiene 39 datos, con 21 variables cada dato.

Los modelos de clasificación utilizados son los mismos que en el caso del NNCI y se toman las mismas consideraciones al reportar las métricas.

	SVM	**NB**	**LR**	**RF**
Exactitud	23.07	74.35	66.66	53.84
AUC	23.70	67.40	71.11	60.37
F1	37.50	50.00	58.06	40.00
Precisión	23.07	45.45	40.90	28.57
Exhaustividad	100.0	55.55	100.0	66.66
U	206	88	78	107
Valor de p	0.94	0.06	0.02	0.17

Tabla 6: Resultados de métricas de evaluación y test de hipótesis (HdD)

Los resultados en este caso son muy diferentes a los obtenidos con la base del NNCI. Las métricas de desempeño no son conclusivas respecto si se logra generalizar a partir de los datos. Para evaluar la significancia que tiene el método, se utiliza el test U de Mann-Whitney, del que se concluye que solo el método de LR es significativo al 5% en generar un etiquetamiento en el que la población de clase positiva tiene puntajes más altos.

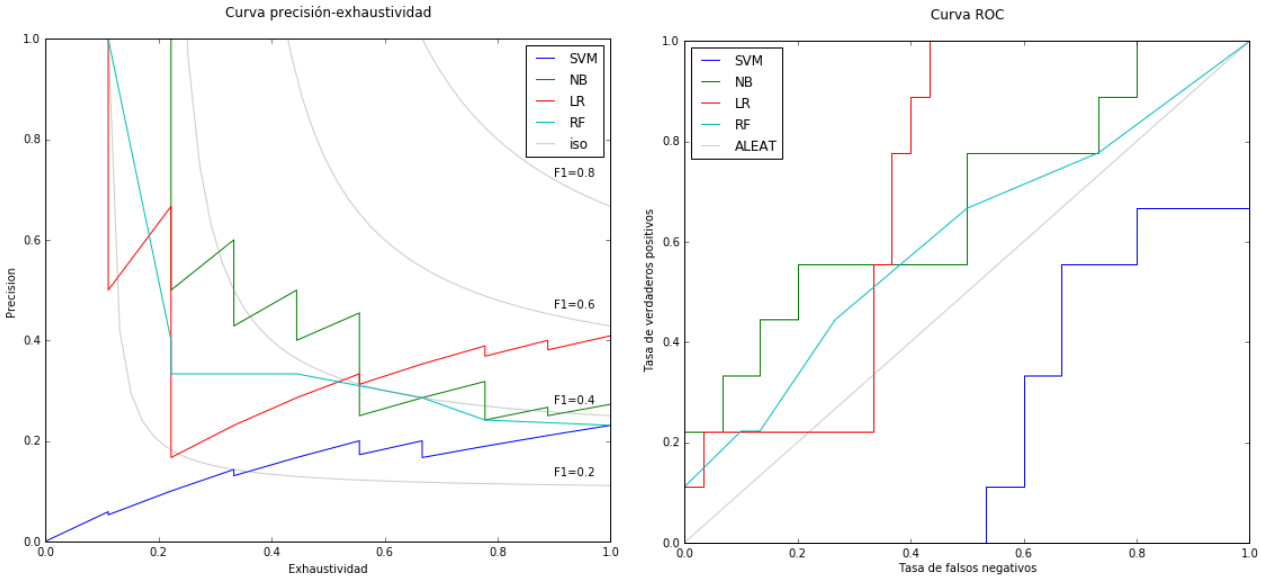


Figura 17: Izq: Curvas de precisión-exhaustividad para los cuatro clasificadores considerados. Der: Curvas ROC para los clasificadores en el caso HdD.

Tomando los resultados tanto de las métricas de desempeño, las curvas ROC y precisión-exhaustividad y las matrices de confusión resulta claro que el modelo de SVM no logra generalizar bien la estructura de los datos. Tanto es así, que el resultado que genera es incluso peor que tirar una moneda equilibrada para realizar las clasificaciones. Por otro lado, se ve que los modelos NB, LR y RF logran generalizar, incluso sin afinar los hiperparámetros, aun cuando esto no sea estadísticamente significativo para los casos de NB y RF.

Se puede ver en lo gráficos de las curvas ROC y precisión-exhaustividad que, si bien el método de LR obtiene mejores estadísticas globales, hay una región en la que el método de NB tiene mejor desempeño.

4.6 Discusión

La exploración visual de los datos es relevante pues permite inmediatamente descartar comportamientos extraños en el conjunto de datos. En este caso permite además conjeturar que una buena clasificación es posible a priori, aunque esto no se puede asegurar pues el aprendizaje de esta representación bidimensional se hace utilizando todos los datos, por lo que no es metodológicamente válido para evaluar.

Los resultados de clasificación en la base de datos del NNCI muestran estadísticamente con una significancia del 0.1% que los clasificadores logran ordenar y por lo tanto diferenciar las distintas poblaciones de niños y adolescentes rusos según si presentan o no síntomas de

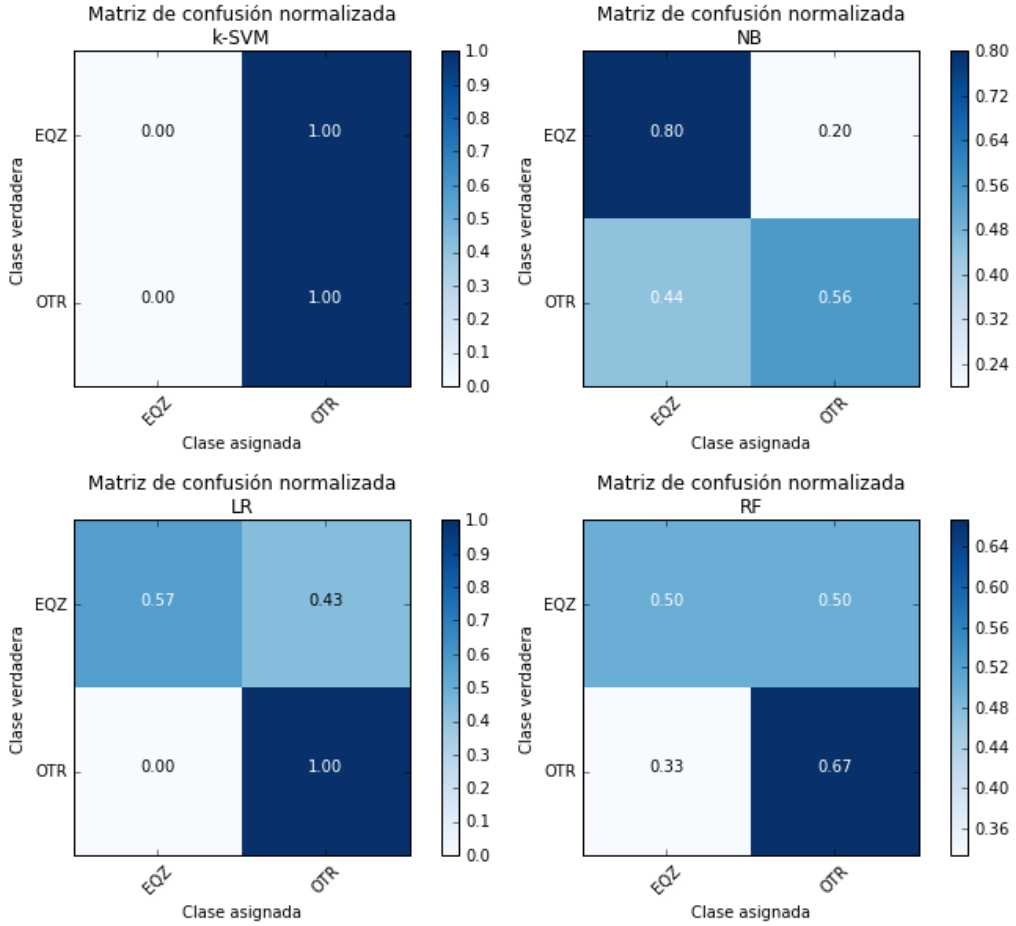


Figura 18: Matrices de confusión normalizadas por clase verdadera para los cuatro clasificadores (HdD)

esquizofrenia. Con esto se rechaza la hipótesis nula y se concluye que hay fuerte evidencia para aceptar la hipótesis alternativa. Las métricas con las que se comparan son enfáticas al mostrar que los clasificadores logran generalizar a partir los datos, manteniendo un ordenamiento consistente según se puede interpretar del AUC y además con buen balance entre precisión y exhaustividad.

La selección de características resulta particularmente importante pues por un lado permite realizar inferencias sobre la naturaleza de la distinción y por otro resulta útil como consideración para la reducción de dimensionalidad hecha en el caso de los datos del HdD. Es importante enfatizar que la separación realizada utilizando clasificadores ralos entrega resultados casi tan buenos como al utilizar todas las características, y se desprende del test de hipótesis que estos también logran ordenar de manera satisfactoria los sujetos según la población de la que provienen.

Las características seleccionadas por los métodos embebidos de selección entregan información relevante con respecto a la naturaleza del fenómeno a estudiar. La primera observación es que tanto los AR8 como la ApEn son muy informativos para la clasificación y por tanto tiene un comportamiento distinto en ambas poblaciones, esto permite inferir que el fenómeno subyacente tiene características no lineales aun cuando se pueda diferenciar también mediante descriptores

lineales. Esto porque la ApEn es una característica no lineal de la serie de tiempo, a diferencia de las AR8. Otro elemento que se destaca es la importancia que se le da a la matriz de covarianza de las señales. De esto se desprende que el método logra extraer que las diferencias no sólo se dan por un comportamiento distinto en ciertas zonas del cerebro, sino que también depende de cómo distintas zonas se relacionan entre ellas. Una última mención se merece la entropía espectral, estas características se relacionan con el grado de información hay en el espectro de frecuencias de una señal. Esto parece mostrar que las diferencias, aunque sutiles e imperceptibles al ojo humano, son multidimensionales, de naturaleza no lineal y medibles calculables mediante este tipo de metodología. Estos resultados van en la misma línea de otras investigaciones que han tenido buenos resultados en este tipo de problemas de clasificación usando medidas de coeficientes autoregresivos, medidas de covarianza y características del análisis no lineal.

El caso estudiado para el HdD es menos prometedor. En este caso la hipótesis nula se rechaza únicamente para el método de LR con una significancia estadística del 5%, lo que permite aceptar la hipótesis de investigación. Es importante enfatizar que los resultados de este experimento se acotan a los tipos de poblaciones consideradas, es decir, personas con diagnóstico psiquiátrico bajo tratamiento farmacológico. Los trastornos considerados corresponden a esquizofrenia, bipolaridad, depresión mayor y trastorno dual con esquizofrenia. Esto implica que los resultados obtenidos no son directamente extrapolables a poblaciones que no estén en tratamiento, sin embargo, da luces de la eventual factibilidad clínica utilizando un mejor sistema de adquisición de datos que no sea de tamaño estándar y poder así comparar más adecuadamente las señales, aunque hay evidencia que señala que algunas diferencias en ritmos de oscilación neuronal se presentan con o sin tratamiento farmacológico.

La diferencia de los resultados obtenidos en el caso de los datos del HdD frente a los obtenidos del NNCI puede deberse a tres factores. El primero es el tamaño de la muestra, mientras el tamaño total del cado del NNCI es de 84 sujetos, el total de sujetos considerados en el otro caso son sólo 39. Un segundo factor viene dado por el desbalance de los datos, en el que en el caso favorables una etiqueta representa el 54% de los datos y 46% la otra, que dista mucho del 23% de la clase menos representada de los datos de HdD. Esto proporción podría no ser un inconveniente tan grande en caso de tener más. El último punto tiene relación con el equipo de adquisición de EEG cuya diferencia se ve en la calidad de señales obtenidas. En este sentido, pareciera ser que un hardware como el Emotiv Epoc no es suficientemente bueno como para obtener mejores resultados, sin embargo, es relevante que estos resultados consideran únicamente las características calculadas a partir de seis de los canales disponibles.

Capítulo 5: Conclusiones

En este trabajo se analiza la implementación de una metodología basada en ciencia de los datos que permite distinguir sujetos según un posible diagnóstico psiquiátrico utilizando únicamente exámenes de EEG tomados en reposo con ojos cerrados para dos poblaciones. Una tomada desde el NNCI de niños y adolescentes con y sin síntomas de esquizofrenia y otra tomada del HdD mediante un hardware de bajo costo. Los resultados son mixtos, pero prometedores. Se identificaron familias de características relevantes y se realizaron las pruebas correspondientes para evaluar la metodología.

5.1 Sobre los objetivos

El objetivo general, establecido en el capítulo de introducción de la forma "evaluar la capacidad de la metodología de ciencia de los datos para diferenciar poblaciones de sujetos mediante EEG en el contexto de diagnósticos psiquiátricos relacionados a la esquizofrenia", se considera cumplido. Esta evaluación resulta de la culminación del proceso de investigación del trabajo, más aun, la evaluación demuestra una buena capacidad de discriminación con los datos del NNCI, esto es, entre una población con síntomas de esquizofrenia frente a una sin síntomas a través de la metodología propuesta. Por otro lado, en el caso de los datos adquiridos con el hardware de Emotiv se obtienen resultas más pobres, aunque no por eso inutilizables, de hecho, para uno de los métodos este sí obtiene relevancia estadística, aunque con un nivel de significancia menor. Por lo que se concluye que la metodología de ciencia de los datos y las herramientas de aprendizaje de máquinas permiten diferenciar las poblaciones en ambos casos con distintos niveles de éxito.

Los objetivos específicos del trabajo son:

a) Diseñar una metodología basada en ciencias de los datos para evaluar las hipótesis de investigación.

b) Implementar un esquema de análisis y construcción de características para las EEG basada en análisis lineal y no lineal de series de tiempo.

c) Entrenar y evaluar distintos métodos de clasificación y extraer grupos de características relevantes para posterior análisis.

d) Determinar la validez estadística de las hipótesis de investigación.

Los objetivos específicos se realizaron en distintas etapas del trabajo. El caso del objetivo a) se cumple a lo largo del capítulo de Metodología. Los objetivos b) y c) se realizaron para poder presentar el capítulo de resultados, las implementaciones se pueden revisar en los anexos. Estas fueron realizadas en Python 3 utilizando las librerías de NumPy y SciPy para computación científica, Scikit-learn para aprendizaje de máquinas y Matplotlib para gráficos. El test de hipótesis se hizo con el software estadístico R y el paquete stats. El objetivo d) se obtiene del resultado del

test U de Mann-Whitney y se discute en el capítulo Resultados y Análisis. De esto se desprende que todos los objetivos específicos fueron cumplidos en el desarrollo de este trabajo.

5.2 Sobre los resultados

La conclusión más importante que se desprende de los resultados del trabajo realizado es que los EEG en reposo son una fuente suficientemente informativa para evaluar la presencia de trastornos psiquiátricos ligados a la esquizofrenia. Esta conclusión toma, además, dos formas específicas al hablar de cada uno de los problemas particulares. Una de estas es que estos datos son informativos al comparar poblaciones de adolescentes, aspecto poco estudiado en la literatura referente a la esquizofrenia. La otra forma específica es que la conclusión se mantiene incluso al utilizar un hardware de adquisición de bajo costo, resultado novedoso también en este ámbito.

5.3 Recomendaciones para trabajos futuros

El punto más importante para trabajos futuros es priorizar la adquisición de datos, tanto en número como en calidad. La posibilidad de tener acceso a datos de mejor calidad resulta en menos tiempo limpiando datos y medidas más precisas de las características calculadas, más aun, tener más datos permite utilizar esquemas de entrenamiento y validación para ajustar los hiperparámetros de los métodos utilizados.

De lo aprendido surgen múltiples posibilidades y preguntas que responder, a continuación, se enumeran algunas como propuestas de investigación futura:

- Estudio de clasificación utilizando muestras de pacientes que no están con tratamiento farmacológico, pues de esta manera se puede apuntar a implementar una herramienta con repercusiones reales que sea de utilidad en los servicios de salud.

- Estudio para evaluar la respuesta a distintos tipos de fármacos a través del EEG mediante técnicas de ciencias de los datos y aprendizaje de máquinas. Según lo mostrado en la revisión bibliográfica, es posible para algunos casos, pero el contexto canadiense puede ser muy distinto al chileno.

- Evaluación de la metodología de ciencias de los datos y aprendizaje de máquinas para el contexto de detección temprana de demencias a través de EEG. Si bien este problema es de otra naturaleza visto desde las ciencias de la salud, son de naturaleza similar al verlas desde el enfoque de ciencias de los datos.

Capítulo 6: Bibliografía

Agius, M., Butler, S., & Holt, C. (2011). Does early diagnosis and treatment of schizophrenia lead to improved long-term outcomes? *Neuropsychiatry, 1*, 553-565.

Alimardani, F., Boostani, R., & Taghavi, M. (2015). Classification of BMD and schizophrenic patients using geometrical analysis of their EEG signal covariance matrices. *Telecommunications and Signal Processing (TSP), 2015 38th International Conference on*, (págs. 1-5).

Azevedo, A. I., & Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: a parallel overview. *IADS-DM*.

Bermúdez, A. N., Spinelli, E. M., & Muravchik, C. M. (2011). Detección de eventos en señales de EEG mediante Entropía Espectral. *XVIII Congreso Argentino de Bioingeniería SABI.*

Bhugra, D. (2005). The global prevalence of schizophrenia. *PLoS medicine, 2*, e151.

Bishop, C. M. (2006). *Pattern recognition and Machine Learning.* Springer Science+Business Media.

Borisov, S. V., Kaplan, A. Y., Gorbachevskaya, N. L., & Kozlova, I. A. (2005). Analysis of EEG structural synchrony in adolescents with schizophrenic disorders. *Human Physiology, 31*, 255-261.

Boutros, N. N., Arfken, C., Galderisi, S., Warrick, J., Pratt, G., & Iacono, W. (2008). The status of spectral EEG abnormality as a diagnostic test for schizophrenia. *Schizophrenia research, 99*, 225-237.

Boyd, K., Eng, K. H., & Page, C. D. (2013). Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals. En H. Blockeel, K. Kersting, S. Nijssen, & F. Zelezny (Ed.), *Machine Learning and Knowledge Discovery in Databases* (págs. 451-466). Berlin: Springer Berlin Heidelberg.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery, 2*, 121-167.

Cielen, D., Meysman, A., & Ali, M. (2016). *Introducing data science: big data, machine learning, and more, using Python tools.* Manning Publications Co.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 273--297.

Crawford, K. (2016). Artificial intelligence's white guy problem. *The New York Times*.

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd international conference on Machine learning*, (págs. 233-240).

Dhindsa, J. (2017). *Generalized Methods for User-Centered Brain-Computer Interfacing.* Ph.D. dissertation.

Dilsizian, S. E., & Siegel, E. L. (2014). Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. *Current cardiology reports, 16*, 441.

Dou, Y. (2017). *Artifact Analysis and Removal of Electroencephalographic (EEG) Recordings.* Ph.D. dissertation, Concordia University.

Duvinage, M., Castermans, T., Petieau, M., Hoellinger, T., Cheron, G., & Dutoit, T. (2013). Performance of the Emotiv Epoc headset for P300-based applications. *Biomedical engineering online, 12*, 56.

El Gohary, M. I., Al Zohairy, T. A., Eissa, A. M., El Deghaidy, S., & Hussein, H. M. (s.f.). An intelligent System for Diagnosis of Schizophrenia and Bipolar Diseases using Support Vector Machine with Different Kernels.

Elorza Pérez-Tejada, H., & Medina Sandoval, J. C. (2000). Estadística para las ciencias sociales y del comportamiento.

Esteller, R., Vachtsevanos, G., Echauz, J., & Litt, B. (2001). A comparison of waveform fractal dimension algorithms. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications, 48*, 177-183.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters, 27*, 861-874.

Gangwar, M., Mishra, R. B., & Yadav, R. S. (2014). Classical and intelligent computing methods in psychiatry and neuropsychitry: an overview. *Int. J. Adv. Res. IT Eng*, *3*, 12.

Garcia, M. (2016). Racist in the machine: The disturbing implications of algorithmic bias. *World Policy Journal, 33*, 111-117.

Gray, R. M., & Davisson, L. D. (2004). *An introduction to statistical signal processing.* Cambridge University Press.

Green, A. (2012). *Computationally Characterizing Schizophrenia.* Ph.D. dissertation, University of Toronto (Canada).

Hasey, G. M. (2013). A review of recent literature employing electroencephalographic techniques to study the pathophysiology, phenomenology, and treatment response of schizophrenia. *Current psychiatry reports*, 1-8.

Hornero, R., Abasolo, D. E., Jimeno, N., & Espino, P. (2003). Applying approximate entropy and central tendency measure to analyze time series generated by schizophrenic patients. *Engineering in Medicine and Biology Society, 2003. Proceedings of the 25th Annual International Conference of the IEEE, 3*, págs. 2447-2450.

Jena, S. K. (2015). Examination stress and its effect on EEG. *Int J Med Sci Pub Health, 11*, 1493-1497.

Jeong, J. a.-H. (1998). Non-linear dynamical analysis of the EEG in Alzheimer's disease with optimal embedding dimension. *Electroencephalography and clinical Neurophysiology, 106*(3), 220-228.

Khodayari-Rostamabad, A., Hasey, G. M., MacCrimmon, D. J., Reilly, J. P., & Bruin, H. (2010). A pilot study to determine whether machine learning methodologies using pre-treatment electroencephalography can predict the symptomatic response to clozapine therapy. *Clinical Neurophysiology, 121*, 1998-2006.

Khodayari-Rostamabad, A., Reilly, J. P., Hasey, G., MacCrimmon, D., & others. (2010). Diagnosis of psychiatric disorders using EEG data and employing a statistical decision model. *Engineering in Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE*, (págs. 4006-4009).

Knight, J. N. (2003). *Signal fraction analysis and artifact removal in EEG*. Ph.D. dissertation, Colorado State University.

Kossoff, E. H., Ritzl, E. K., Politsky, J. M., Murro, A. M., Smith, J. R., Duckrow, R. B., . . . Bergey, G. K. (2004). Effect of an external responsive neurostimulator on seizures and electrographic discharges during subdural electrode monitoring. *Epilepsia, 45*, 1560-1567.

Lac, S. d., Squire, L. R., Bloom, F., & Berg, D. (2008). Fundamental Neuroscience. Elsevier.

Lempel, A., & Ziv, J. (1976). On the complexity of finite sequences. *IEEE Transactions on information theory, 22*, 75-81.

Li, Y.-j., & Fan, F.-y. (2006). Classification of Schizophrenia and depression by EEG with ANNs. *Engineering in Medicine and Biology Society, 2005. IEEE-EMBS 2005. 27th Annual International Conference of the*, (págs. 2679-2682).

Lior, R., & others. (2014). *Data mining with decision trees: theory and applications* (Vol. 81). World scientific.

Majnik, M., & Bosnic, Z. (2013). ROC analysis of classifiers in machine learning: A survey. *Intelligent data analysis, 17*, 531-558.

Manis, G. (2008). Fast computation of approximate entropy. *Computer methods and programs in biomedicine, 91*, 48-54.

Marx, V. (12 de Junio de 2013). *Biology: The big challenges of big data*. Obtenido de Nature: http://www.nature.com/nature/journal/v498/n7453/full/498255a.html

Messias, E. L., Chen, C.-Y., & Eaton, W. W. (2007). Epidemiology of schizophrenia: review of findings and myths. *Psychiatric Clinics, 30*, 323-338.

MINISTERIO DE SALUD. (2016). *Guía Clínica: Para el tratamiento de personas desde el primer episodio de Esquizofrenia.* Santiago: MINSAL, (2016).

Murphy, K. P. (2012). *Machine learning, a probabilistic perspective.* The MIT Press.

Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. *Proceedings of the 22nd international conference on Machine learning*, (págs. 625-632).

NNCI. (s.f.). *Laboratory for Neurophysiology*. Recuperado el 30 de 1 de 2017, de http://brain.bio.msu.ru/eeg_schizophrenia.htm: http://brain.bio.msu.ru/eeg_schizophrenia.htm

Nuwer, M. (1997). Assessment of digital EEG, quantitative EEG, and EEG brain mapping: report of the American Academy of Neurology and the American Clinical Neurophysiology Society. *Neurology, 49*, 277-292.

OMS. (Abril de 2016). *http://www.who.int/mediacentre/factsheets/fs397/en/*. Obtenido de WHO - Schizophrenia: http://www.who.int/mediacentre/factsheets/fs397/en/

OMS. (Abril de 2016). *WHO*. Obtenido de WHO - Schizofrenia: http://www.who.int/mediacentre/factsheets/fs397/en/

Parvinnia, E., Sabeti, M., Jahromi, M. Z., & Boostani, R. (2014). Classification of EEG Signals using adaptive weighted distance nearest neighbor algorithm. *Journal of King Saud University-Computer and Information Sciences, 26*, 1-6.

Pincus, S. M., Gladstone, I. M., & Ehrenkranz, R. A. (1991). A regularity statistic for medical data analysis. *Journal of clinical monitoring, 7*, 335-345.

Purves, D., Augustine, G. J., Fitzpatrick, D., Hall, W. C., LaMantia, A., McNamara, J. O., & Williams, S. M. (2004). Neuroscience. Massachusetts. *Publishers Sunderland*.

Rennie, J. D., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of naive bayes text classifiers. *Proceedings of the 20th international conference on machine learning (ICML-03)*, (págs. 616-623).

Rish, I. (2001). An empirical study of the naive Bayes classifier. *IJCAI 2001 workshop on empirical methods in artificial intelligence, 3*, págs. 41-46.

Rodríguez, J. J.-G. (2009). *Epidemiología de los trastornos mentales en América Latina y el Caribe.* Washington, DC: Pan American Health Org.

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one, 10*, e0118432.

Sammut, C., & Webb, G. I. (2011). *Encyclopedia of machine learning.* Springer Science & Business Media.

Seo, S.-H., & Lee, J.-T. (2010). Stress and EEG. En *Convergence and hybrid information technologies.* InTech.

Theiler, J., & Rapp, P. E. (1996). Re-examination of the evidence for low-dimensional, nonlinear structure in the human electroencephalogram. *Electroencephalography and clinical Neurophysiology, 98*, 213-222.

Timashev, S. F., Panischev, O. Y., Polyakov, Y. S., Demin, S. A., & Kaplan, A. Y. (2012). Analysis of cross-correlations in electroencephalogram signals as an approach to proactive diagnosis of schizophrenia. *Physica A: Statistical Mechanics and its Applications, 391*, 1179-1194.

Treviño, M., & Gutiérrez, R. (2007). Las bases celulares de las oscilaciones neuronales. *Salud Mental, 30*.

van Os, J. e. (Agosto de 2009). Schizophrenia. *The Lancet, Volume 374*(Issue 9690), 635 - 645.

Wikimedia Commons. (2015). File:21 electrodes of International 10-20 system for EEG.svg --- Wikimedia Commons{,} the free media repository. Obtenido de https://commons.wikimedia.org/w/index.php?title=File:21_electrodes_of_International_10-20_system_for_EEG.svg&oldid=169572190

Wikimedia Commons. (2016). File:Kernel Machine.png --- Wikimedia Commons, the free media repository. Obtenido de https://commons.wikimedia.org/w/index.php?title=File:Kernel_Machine.png&oldid=215993536

Wikimedia Commons. (2016). File:Svm max sep hyperplane with margin.png --- Wikimedia Commons, the free media repository. Obtenido de https://commons.wikimedia.org/w/index.php?title=File:Svm_max_sep_hyperplane_with_margin.png&oldid=225162099

Wittchen, H.-U., Jacobi, F., Rehm, J., Gustavsson, A., Svensson, M., Jönsson, B., . . . others. (2011). The size and burden of mental disorders and other disorders of the brain in Europe 2010. *European neuropsychopharmacology, 21*, 655-679.

Yadav, P. S., Arya, R. K., & Yadav, V. (s.f.). An Overview of Various Computing Methods in Psychiatry and Neuropsychiatry.

Capítulo 7: Anexos

7.1 Anexo 1: Ingeniería de características

Implementado en Python 3.

```
import numpy as np
import scipy

# Functions to calculate the geometric mean of the covariance matrix
def M_EXP(A,B=None):
    if B is None:
        R = scipy.linalg.expm(A)
    else:
        q,w,e = scipy.linalg.svd(B)
        P2 = np.dot(q,np.diag(np.sqrt(w)))
        P2_inv = np.dot(np.diag(1/np.sqrt(w)),e)
        R = np.dot(P2,np.dot(scipy.linalg.expm(
            np.dot(P2_inv,np.dot(A,P2_inv.T))),P2.T))
    return R

def M_LOG(A,B=None):
    if B is None:
        R = scipy.linalg.logm(A)
    else:
        q,w,e = scipy.linalg.svd(B)
        P2 = np.dot(q,np.diag(np.sqrt(w)))
        P2_inv = np.dot(np.diag(1/np.sqrt(w)),e)
        R = np.dot(P2,np.dot(scipy.linalg.logm(
            np.dot(P2_inv,np.dot(A,P2_inv.T))),P2.T))
    return R

def Mean_SPD(P_i,epsilon = 1e-3, max_iter = 50, disp=False):
    n,m,_ = P_i.shape
    l = np.copy(P_i)
    P = np.mean(l,0)
    S = np.ones_like(P)
    c=0
    while np.linalg.norm(S)>epsilon and c<max_iter:
        for i in range(n):
            l[i,:,:] = M_LOG(P_i[i,:,:],P)
        S = np.mean(l,0)
        if np.isnan(M_EXP(S,P)[0,0]):
            return [S,P]
        P = M_EXP(S,P)
        if disp:
            print("Norma S \n",np.linalg.norm(S),"\nIteración: ", c+1)
        c=c+1
    return P

def M_NORM(A,B):
    q,w,e = scipy.linalg.svd(B)
    P2 = np.dot(q,np.diag(np.sqrt(w)))
    P2_inv = np.dot(np.diag(1/np.sqrt(w)),e)
    return np.linalg.norm(scipy.linalg.logm(np.dot(P2_inv,np.dot(A,P2_inv.T))))
# Function to calculate te AR coefficients, and the mean among the samples
def AR(signals, p=8, channels=N_channels, samples = 1):
    (a,b) = signals.shape
    b = int(b/channels)
    r = zeros((b+p-1+1,p))
    out2 = zeros((a,(p+1)*channels))
    for i in range(a):
```

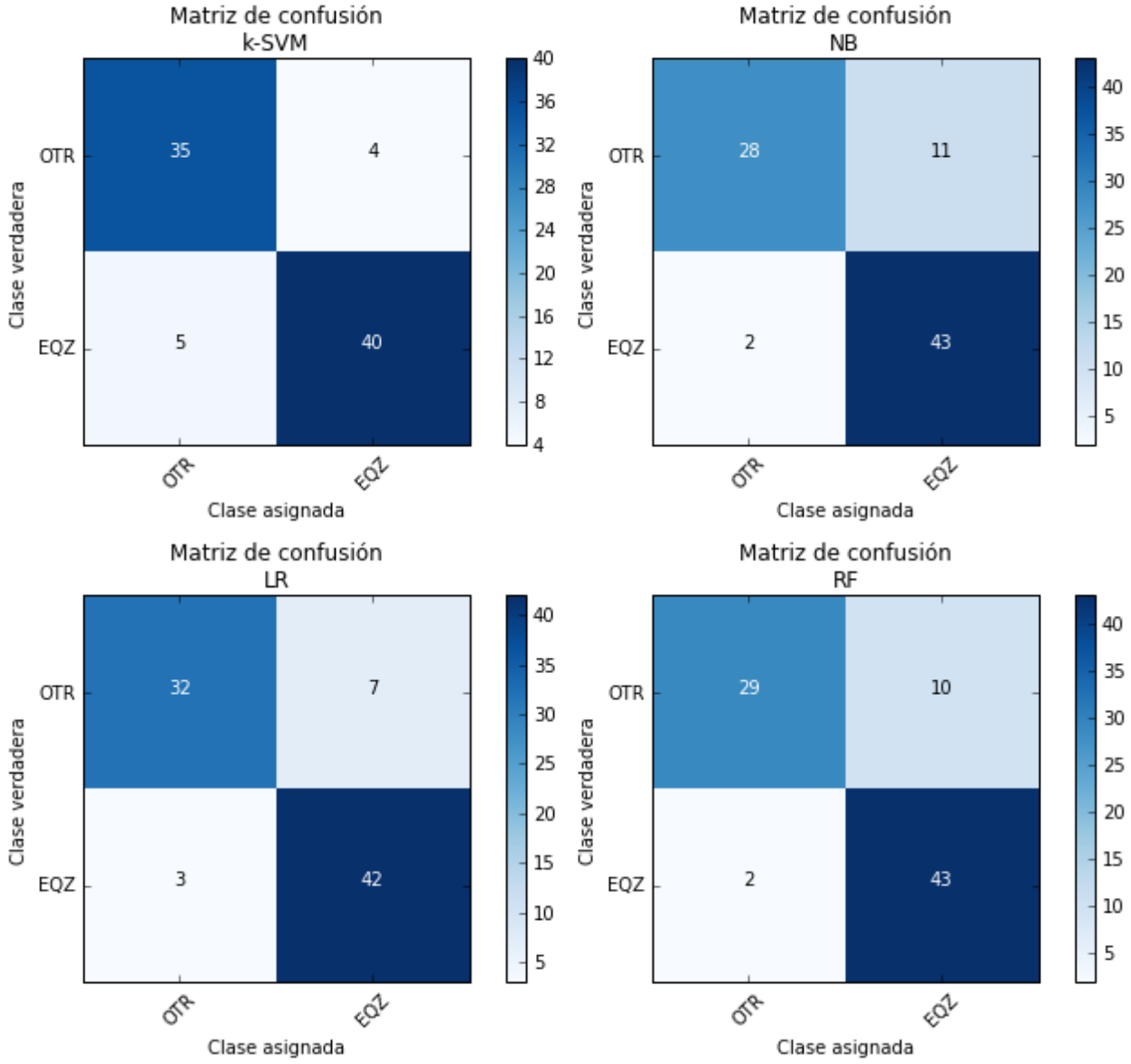
```
        for j in range(channels):
            s_aux = signals[i,j*b:(j+1)*b]
            for k in range(p):
                r[p-k-1:-k-1,k] = s_aux
            L_ = np.concatenate((r[p-1:-p-1,:],np.ones((b-p,1))),axis=1)
            coefs = np.linalg.solve(np.dot(L_.T,L_),np.dot(L_.T,s_aux[p:])).flatten()
            out2[i,j*(p+1):(j+1)*(p+1)] = coefs
    if samples > 1:
        out1 = np.zeros((int(a/samples),(p+1)*channels))
        for i in range(int(a/samples)):
            out1[i,:] = np.mean(out2[i*samples:(i+1)*samples,:],axis=0)
    else:
        out1 = out2
    return (out1, out2)

# RPDE

def embedding(x,m):
    a = x.size
    y = np.zeros((m,a-m+1))
    for i in range(a-m+1):
        y[:,i]=x[i:i+m]
    return y

def closed_returns(x,m,r):
    a = x.size
    closed_ = np.zeros(a)
    r2=r*r
    y = embedding(x,m)
    for i in range(a-m+1):
        j=i+1
        aFlag = False
        while((j < a-m+1) and (not aFlag)):
            x_d = np.sum((y[:,i]-y[:,j])**2)
            if x_d > r:
                aFlag = True
            j+=1
        aFlag = False
        while((j < a-m+1) and (not aFlag)):
            x_d = np.sum((y[:,i]-y[:,j])**2)
            if x_d <= r:
                aFlag = True
                dif = j-i
                closed_[dif] = closed_[dif]+1
            j+=1
    return closed_

def RPDE_(x,m,epsilon,T_max=0):
    if T_max == 0:
        T_max = x.size
    a = closed_returns(x,m,epsilon)
    t = T_max
    a = a[:t]
    s = np.sum(a)
    if s == 0:
        return
    a = a/s
    return max((scipy.stats.entropy(a)/np.log(t),0))

def RPDE(x,m,epsilon,T_max=0):
    if x.shape[0]/x.size==1:
        return RPDE_(x,m,epsilon,T_max=0)
    return np.array(list(map(lambda l: RPDE_(l,m,epsilon,T_max),x)))
```

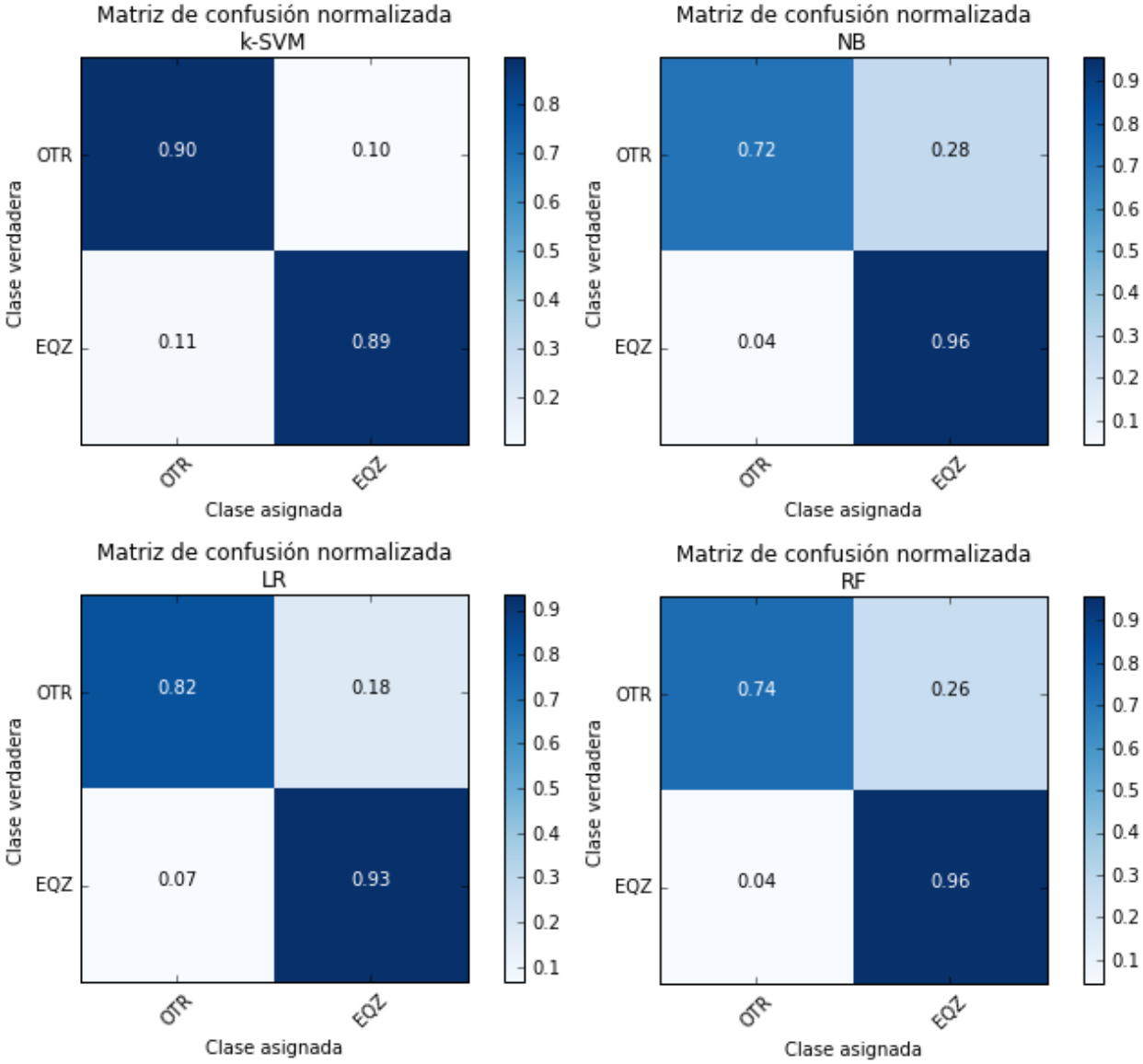
7.2 Anexo 2: Test de hipótesis

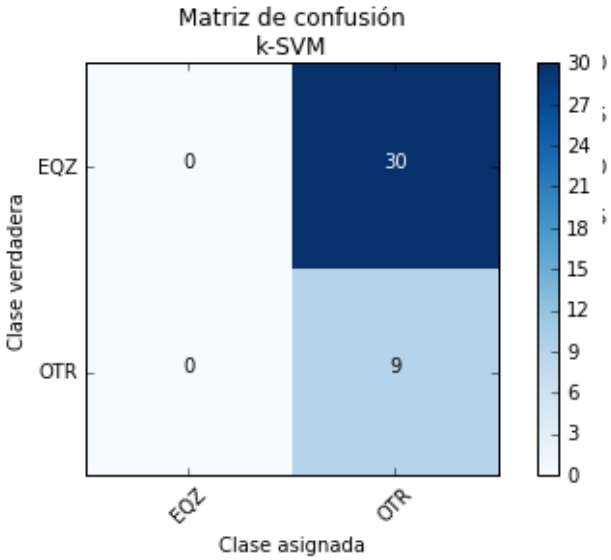
Evaluado usando R v3.

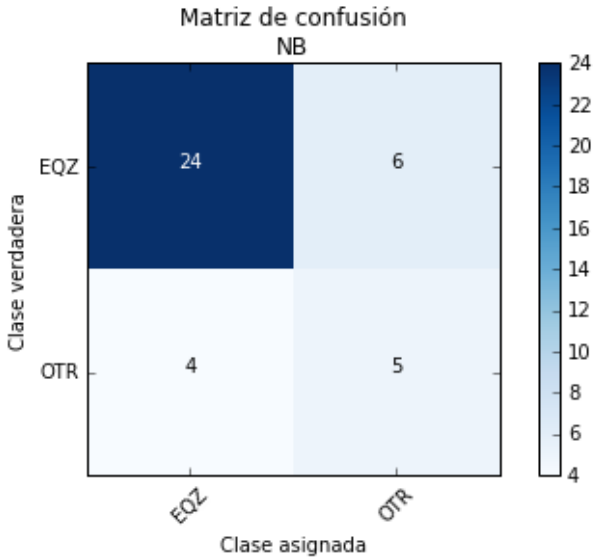
```
library("stats")
print(wilcox.test(pob_n,pob_p, alternative = "less") )# pob_n: Población negative, pob_p: Población positiva
```

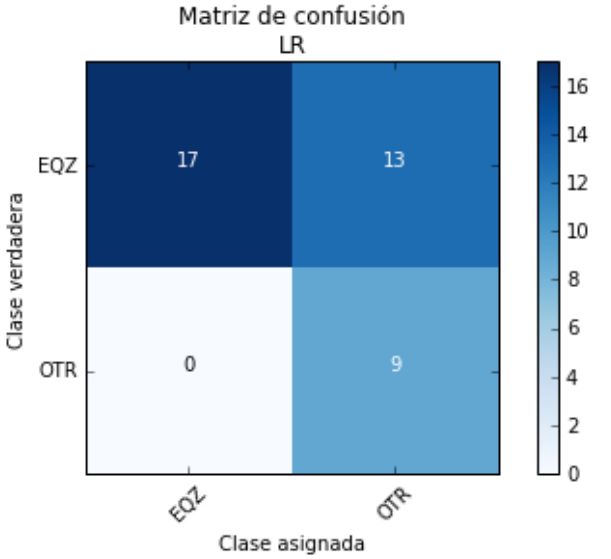
7.3 Anexo 3: Matrices de confusión

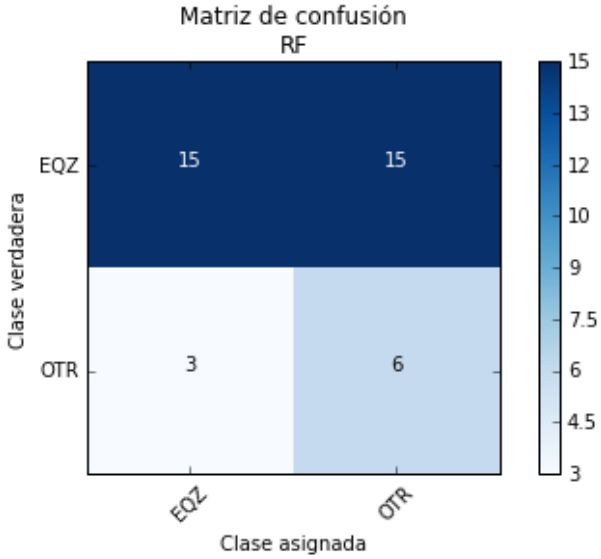


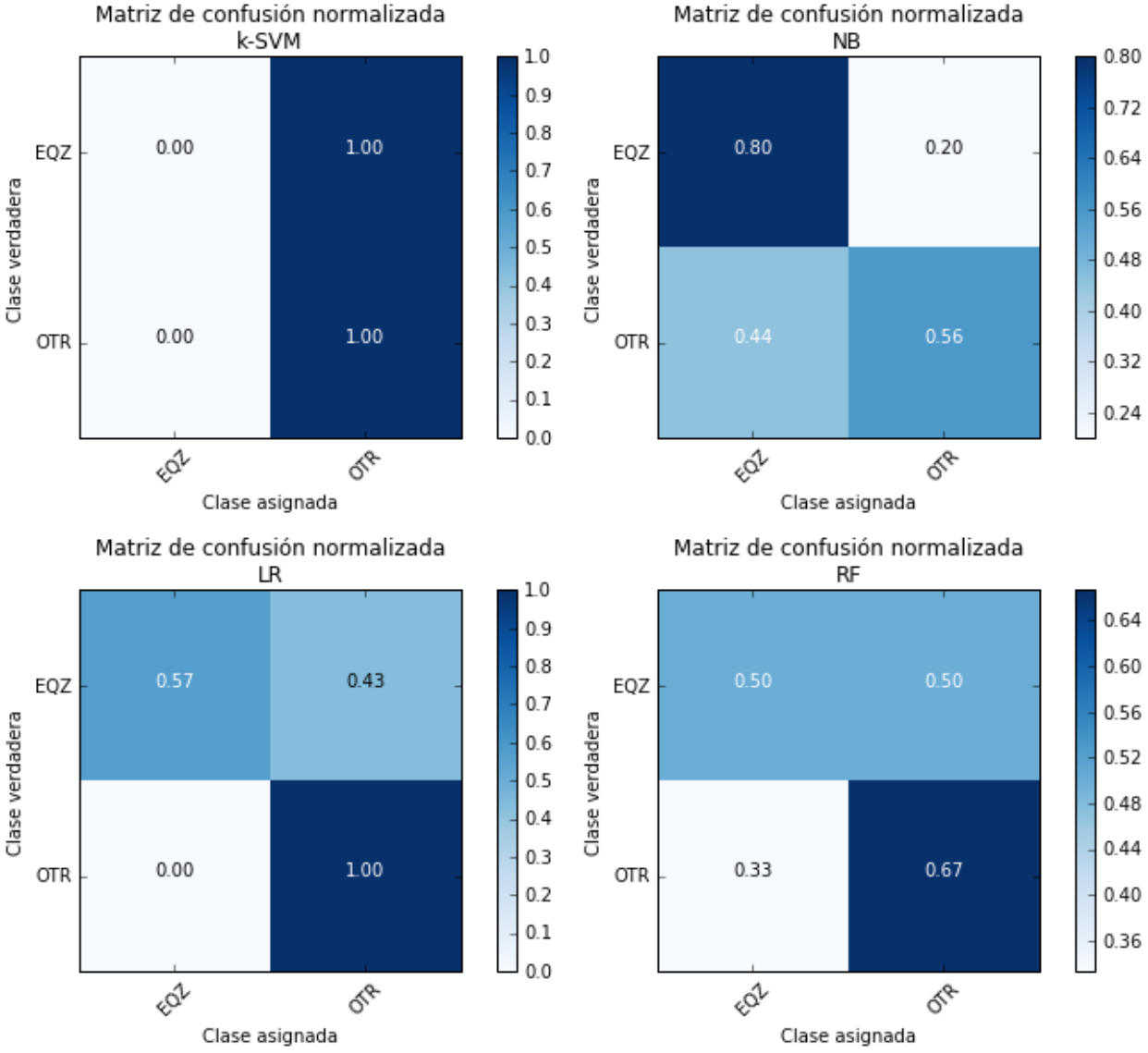

Matriz de confusión normalizada
k-SVM
Clase verdadera
OTR
EQZ
0.90
0.10
0.11
0.89
Clase asignada
Matriz de confusión normalizada
NB
0.72
0.28
0.04
0.96
Matriz de confusión normalizada
LR
0.82
0.18
0.07
0.93
Matriz de confusión normalizada
RF
0.74
0.26
0.04
0.96


Matriz de confusión
k-SVM
EQZ
OTR
0
30
0
9
Clase verdadera
Clase asignada
EQZ
OTR
30
27
24
21
18
15
12
9
6
3
0

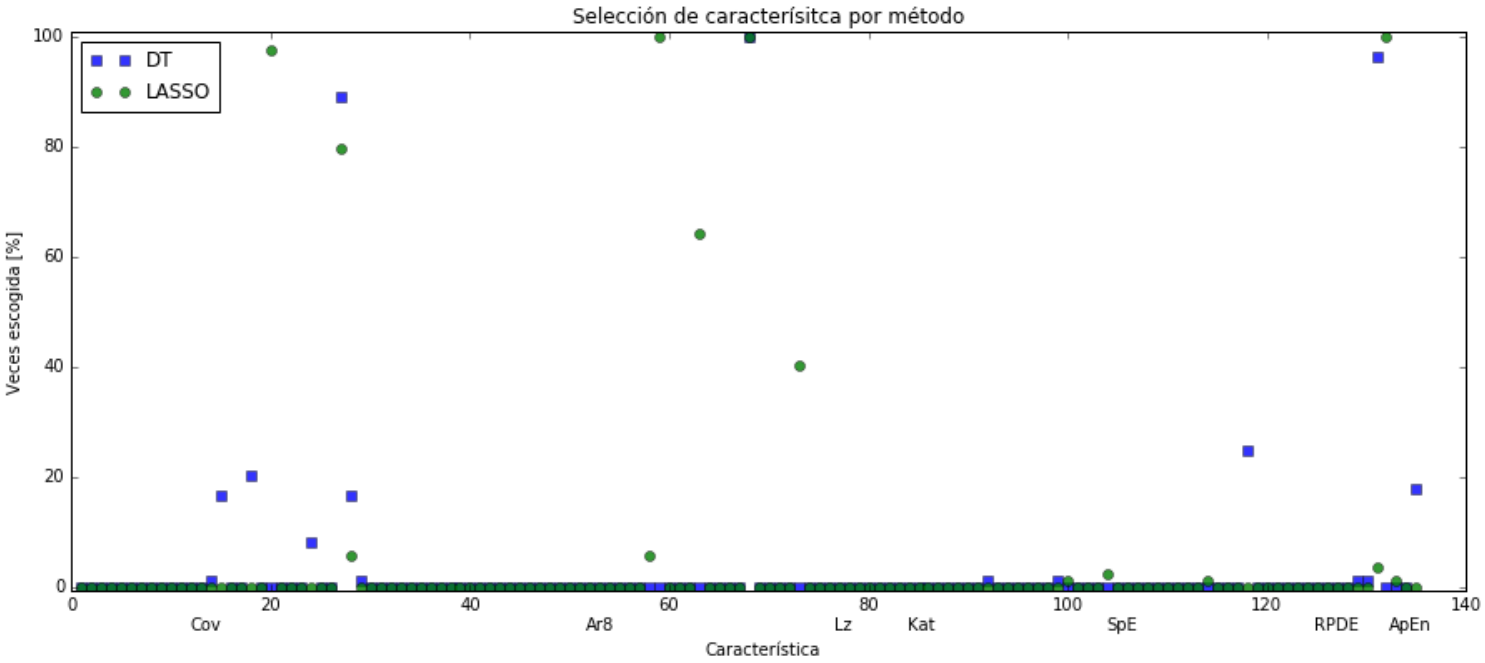

Matriz de confusión
NB
EQZ
OTR
24
6
4
5
Clase verdadera
Clase asignada
EQZ
OTR
24
22
20
18
16
14
12
10
8
6
4

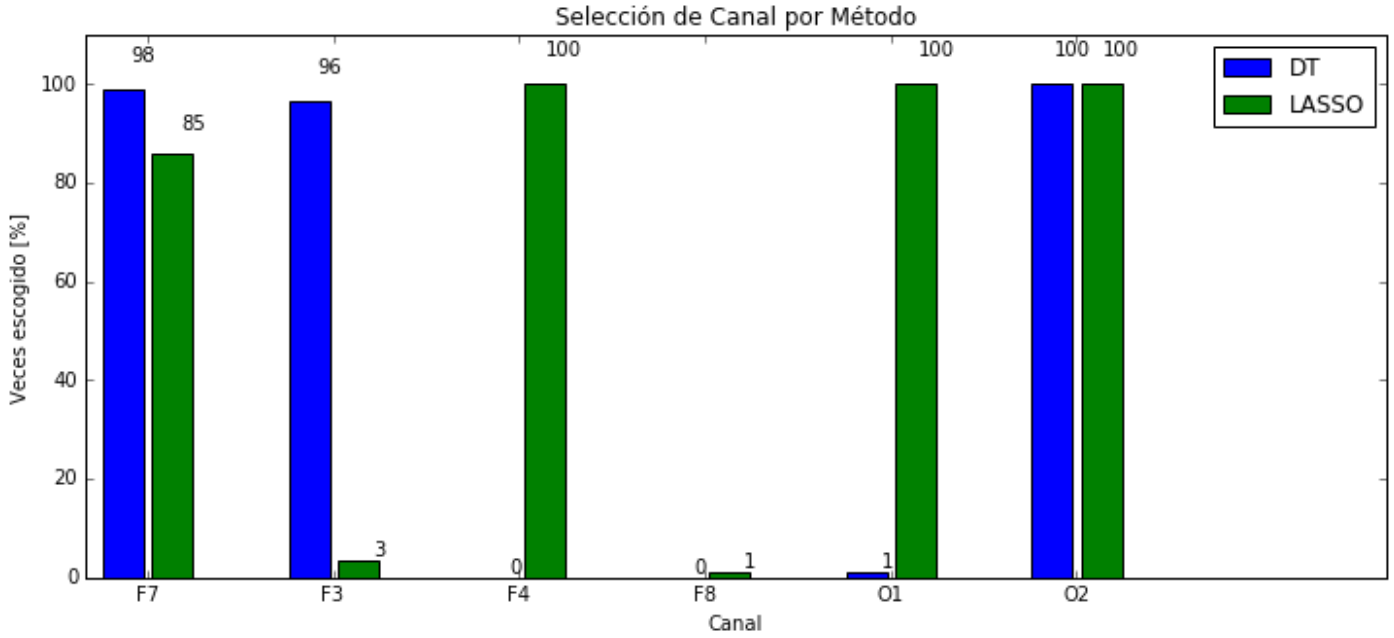

Matriz de confusión
LR
EQZ
OTR
17
13
0
9
Clase verdadera
Clase asignada
EQZ
OTR
16
14
12
10
8
6
4
2
0

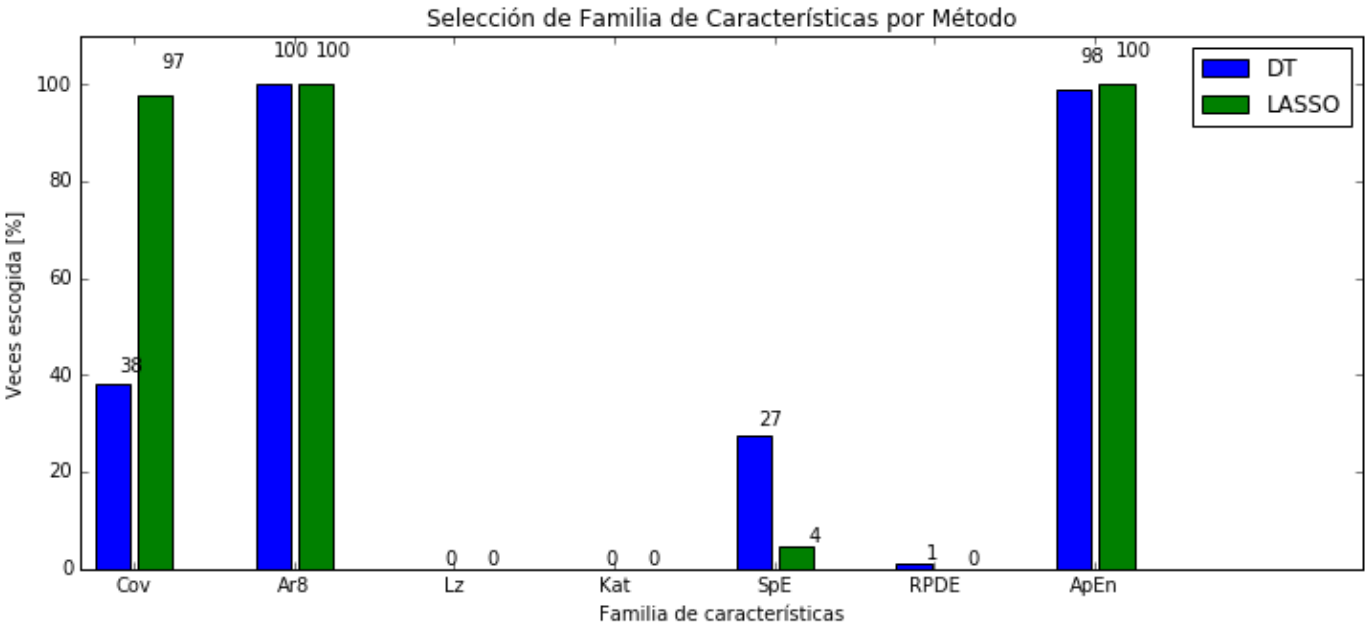

Matriz de confusión
RF
EQZ
OTR
15
15
3
6
Clase verdadera
Clase asignada
EQZ
OTR
15
13
12
10
9
7.5
6
4.5
3


Matriz de confusión normalizada
k-SVM
EQZ
OTR
Clase verdadera
0.00
1.00
0.00
1.00
EQZ
OTR
Clase asignada
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Matriz de confusión normalizada
NB
EQZ
OTR
Clase verdadera
0.80
0.20
0.44
0.56
EQZ
OTR
Clase asignada
0.80
0.72
0.64
0.56
0.48
0.40
0.32
0.24
Matriz de confusión normalizada
LR
EQZ
OTR
Clase verdadera
0.57
0.43
0.00
1.00
EQZ
OTR
Clase asignada
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Matriz de confusión normalizada
RF
EQZ
OTR
Clase verdadera
0.50
0.50
0.33
0.67
EQZ
OTR
Clase asignada
0.64
0.60
0.56
0.52
0.48
0.44
0.40
0.36

7.4 Anexo 4: Selección de características






Selección de Familia de Características por Método
Veces escogida [%]
Familia de características
DT
LASSO
Cov
Ar8
Lz
Kat
SpE
RPDE
ApEn
38
97
100
100
0
0
0
0
27
4
1
0
98
100