INSTITUTO TECNOLÓGICO DE LEÓN

División de Estudios de Posgrado e investigación

"Análisis y tratamiento de señales EEG para la clasificación del proceso cerebral del lenguaje a través de Active Learning"

Tesis

Que presenta:

Eugenio Salvador Martínez Velazquez

Para obtener el grado de:

Maestro en Ciencias
de la Computación

Con la dirección de:

Dra. María del Rosario Baltazar Flores

Con la co-dirección de:

Dr. Carlos Alberto Reyes García

Leon, Guanajuato. Enero 2021

EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Instituto Tecnológico de León

León, Guanajuato, 22/junio/2021

DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN
OFICIO No. DEPI-096-2021

ING. EUGENIO SALVADOR MARTÍNEZ VELÁZQUEZ
PRESENTE

De acuerdo al fallo emitido por la Comisión Revisora, integrada por los: Dra. María del Rosario Baltazar Flores, Dr. Carlos Alberto Reyes García, MC. Miguel Angel Casillas Araiza, MC. Martha Alicia Rocha Sánchez, considerando que llena todos los requisitos establecidos en los Lineamientos Generales para la Operación del Posgrado del Tecnológico Nacional de México, se autoriza la impresión del trabajo de tesis titulado: "Análisis y tratamiento de señales EEG para la clasificación del proceso cerebral del lenguaje a través del proceso Active Learning". Lo que hacemos de su conocimiento para los efectos y fines correspondientes.

ATENTAMENTE

Excelencia en Educación Tecnológica®
Ciencia Tecnología y Libertad

DR. DAVID ISAEL GUTIÉRREZ HERNÁNDEZ
JEFE DE LA DEPI

SEP
TECNOLÓGICO NACIONAL DE MÉXICO
INSTITUTO TECNOLÓGICO DE LEÓN
DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

C.c.p. Expediente

Av. Tecnológico s/n Fracc. Industrial
Julián de Obregón C.P 37290
León, Gto. México Tel. 01 (477) 7105200,
e-mail: tecleon@leon.tecnm.mx
tecnm.mx | leon.tecnm.mx

México 2021 Año de la Independencia

EDUCACIÓN

TECNOLÓGICO NACIONAL DE MÉXICO

León, Gto., a 31 de mayo del 2021

C. ING. LUIS ROBERTO GALLEGOS MUÑOZ
JEFE DE SERVICIOS ESCOLARES
P R E S E N T E

Por este medio hacemos de su conocimiento que la tesis titulada **"Análisis y tratamiento de señales EEG para la clasificación del proceso cerebral del** lenguaje a **través de Active Learning"**, ha sido leída y aprobada por los miembros del Comité Tutorial para su evaluación por el jurado del acto de examen de grado al alumno (a) **C. Eugenio Salvador Martínez Velazquez,** con número de control **M19240009** como parte de los requisitos para obtener el grado de Maestro(a) en Ciencias de la Computación (MCCOM-2011-05).

Sin otro particular por el momento, quedamos de Usted.

A T E N T A M E N T E
COMITÉ TUTORIAL

Dr. María del Rosario Baltazar Flores
DIRECTOR

Dr. Carlos Reyes García
CODIRECTOR ó REVISOR

MC. Martha Alicia Rocha Sánchez
REVISOR (a)

MC. Miguel Ángel Casillas Araiza
REVISOR (a)

SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

DECLARACION DE AUTENTICIDAD Y DE NO PLAGIO

Yo, **Eugenio Salvador Martínez Velazquez** identificado con No. Control **M19240009**, alumno (a) del programa de la **Maestría en Ciencias de la Computación**, autor (a) de la Tesis titulada: **"Análisis y tratamiento de señales EEG para la clasificación del proceso cerebral del lenguaje a través de Active Learning"** DECLARO QUE:

1.- El presente trabajo de investigación, tema de la tesis presentada para la obtención del título de **MAESTRO (A) EN CIENCIAS DE LA COMPUTACIÓN** es original, siendo resultado de mi trabajo personal, el cual no he copiado de otro trabajo de investigación, ni utilizado ideas, fórmulas, ni citas completas "stricto sensu", así como ilustraciones, fotografías u otros materiales audiovisuales, obtenidas de cualquier tesis, obra, artículo, memoria, etc. en su versión digital o impresa.

2.- Declaro que el trabajo de investigación que pongo a consideración para evaluación no ha sido presentado anteriormente para obtener algún grado académico o título, ni ha sido publicado en sitio alguno.

3.- Declaro que las pruebas o experimentos derivados de esta investigación fueron realizados bajo el consentimiento de los involucrados y con fines estrictamente académicos conforme a criterios éticos de confidencialidad.

Soy consciente de que el hecho de no respetar los derechos de autor y hacer plagio, es objeto de sanciones universitarias y/o legales por lo que asumo cualquier responsabilidad que pudiera derivarse de irregularidades den la tesis, así como de los derechos sobre la obra presentada.

Asimismo, me hago responsable ante el Tecnológico Nacional de México/Instituto Tecnológico de León o terceros, de cualquier irregularidad o daño que pudiera ocasional por el incumplimiento de lo declarado.

De identificarse falsificación, plagio, fraude, o que el trabajo de investigación haya sido publicado anteriormente; asumo las consecuencias y sanciones que de mi acción se deriven, responsabilizándome por todas las cargas pecuniarias o legales que se deriven de ello sometiéndome a las normas establecidas en los Lineamientos y Disposiciones de la Operación de Estudios de Posgrado en el Tecnológico Nacional de México.

León, Guanajuato a 16 de agosto de 2021.

Eugenio Salvador Martínez Velazquez

SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

ACUERDO PARA USO DE OBRA (TESIS DE GRADO)

A QUIEN CORRESPONDA

PRESENTE

Por medio del presente escrito, **Eugenio Salvador Martínez Velazquez** (en lo sucesivo el AUTOR) hace constar que es titular intelectual de la obra denominada: **"Análisis y tratamiento de señales EEG para la clasificación del proceso cerebral del lenguaje a través de Active Learning"**, (en lo sucesivo la OBRA) en virtud de lo cual autoriza al Tecnológico Nacional de México/Instituto Tecnológico de León (en lo sucesivo TECNM/IT León) para que efectúe resguardo físico y/o electrónico mediante copia digital o impresa para asegurar su disponibilidad, divulgación, comunicación pública, distribución, transmisión, reproducción, así como digitalización de la misma con fines académicos y sin fines de lucro como parte del Repositorio Institucional del TECNM/ITLeón.

De igual manera, es deseo del AUTOR establecer que esta autorización es voluntaria y gratuita, y que de acuerdo a lo señalado en la Ley Federal del Derecho de Autor y la Ley de Propiedad Industrial el TECNM/IT León cuenta con mi autorización para la utilización de la información antes señalada, estableciendo que se utilizará única y exclusivamente para los fines antes señalados. El AUTOR autoriza al TECNM /IT León a utilizar la obra en los términos y condiciones aquí expresados, sin que ello implique se le conceda licencia o autorización alguna o algún tipo de derecho distinto al mencionada respecto a la "propiedad intelectual" de la misma OBRA; incluyendo todo tipo de derechos patrimoniales sobre obras y creaciones protegidas por derechos de autor y demás formas de propiedad intelectual reconocida o que lleguen a reconocer las leyes correspondientes. Al reutilizar, reproducir, transmitir y/o distribuir la OBRA se deberá reconocer y dar créditos de autoría de la obra intelectual en los términos especificados por el propio autor, y el no hacerlo implica el término de uso de esta licencia para los fines estipulados. Nada de esta licencia menoscaba o restringe los derechos patrimoniales y morales del AUTOR.

De la misma manera, se hace manifiesto que el contenido académico, literario, la edición y en general de cualquier parte de la OBRA son responsabilidad de AUTOR, por lo que se deslinda al (TECNM/ITLeón) por cualquier violación a los derechos de autor y/o propiedad intelectual, así como cualquier responsabilidad relacionada con la misma frente a terceros. Finalmente, el AUTOR manifiesta que estará depositando la versión final de su documento de Tesis, OBRA, y cuenta con los derechos morales y patrimoniales correspondientes para otorgar la presente autorización de uso.

En la ciudad de León, del estado de Guanajuato a los 16 días del mes de agosto de 2021.

Atentamente,

Eugenio Salvador Martínez Velazquez

Agradecimientos

”La discapacidad no te define; te define como haces frente a los desafíos que la discapacidad te presenta”. Jim Abbott

Agradecido estoy con Dios y con la vida, por permitirme cumplir una meta más; por darme la fuerza para levantarme y sostenerme a pesar de los embates de la vida misma; por darme la sabiduría para aprender a superar los desafíos que hoy me han llevado a culminar este gran logro.

Agradezco enormemente al CONACYT y al Instituto tecnológico de Leon, por permitirme ser parte de este mundo de la investigación científica y por ser los entes que coadyuvaron a mi formación. De manera muy especial agradezco a la Dra. Rosario Baltazar, por creer en mi potencial para emprender juntos este trabajo de investigacion, así como por su apoyo y guía como asesora de esta tesis. Al Dr. Carlos Reyes del INAOE, al MC. Ángel Casillas, y a la MC. Martha Rocha; por sus invaluables conocimientos que abonaron a mi formación.

De corazon agradezco a mis padres Salvador Martínez y Lorena Velazquez, por su apoyo incondicional, por ser mis mas grandes consejeros, por todas sus palabras de aliento y por no permitirme dejarme vencer ante las dificultades; a ustedes dedico este trabajo de investigacion que he logrado culminar con todo mi esfuerzo y su incondicional apoyo. Por ultimo agradezco a dos personitas muy especiales Paulina y Alexa, que me motivaron e inspiraron para realizar este trabajo de investigacion.

Publicaciones

- Multiagent system as support for the diagnosis of language impairments using BCI-Neurofeedback: Preliminary Study

  14th International KES Conference on Agents and Multi-agent Systems - Technologies and Applications (AMSTA-20).

  Eugenio Martínez, Rosario Baltazar, Carlos A. Reyes-García, Miguel Casillas, Martha-Alicia Rocha, Socorro Gutierrez, Ma. Del Consuelo Martínez Wbaldo

  Virtual Conference.

  17-19 june 2020. Split, Croatia

  Smart Innovation, Systems and Technologies

  ISSN: 2190-3018

Tabla de Contenido

Índice de Figuras

Índice de Tablas

Introducción

Según el Banco Mundial, el 15 % de la poblacion a nivel mundial padecen discapacidad, de los cuales, la quinta parte de la poblacion del mundo experimentan alguna discapacidad considerable. Dentro del territorio mexicano, el Estado de Guanajuato es el quinto con mayor porcentaje de habitantes con discapacidad; donde 57 de cada mil habitantes es diagnosticado. Las discapacidades con mayor porcentaje dentro de la entidad, son aquellas relacionadas con problemas auditivos, así como problemas de lenguaje.

Debido a los descubrimientos de Hans Berger, quien es considerado el padre de la electroencefalografía; se abrieron diversas líneas de investigacion cuyo fin es el explorar los confines de la mente humana. Una de esas líneas de investigacion se centra en identificar como se procesa el lenguaje a nivel cerebral, así como identificar las patologías neuronales consecuentes de alguna alteracion, que repercuten en el correcto desarrollo del lenguaje. Con el pasar del tiempo, la tecnología se ha sumado a todas estas disciplinas dedicadas a desentrañar los misterios del proceso cerebral, aportando un gran valor que propicia adquirir un mejor entendimiento del procesamiento del lenguaje a nivel cerebral.

El objeto de la presente investigacion, es mostrar la aplicación de técnicas de Aprendizaje Activo (Active Learnig) en el campo de la lingüística, que con el sustento teorico que se tiene tanto médico como computacional; se plantea el objetivo de determinar el proceso cerebral del lenguaje en niños de edad escolar, con el fin de brindar una herramienta de apoyo al diagnostico, así como el aportar conocimiento para aprender más sobre el comportamiento del lenguaje a nivel cerebral.

En los inicios de este trabajo de investigación, se había planteado que la señal EEG fuera obtenida de pacientes como se menciona en el quinto capítulo de esta tesis. Pero debido a la pandemia COVID-19, suceso historico que comenz desde finales del años 2019, que ha afectado a nivel mundial dejando un sin número de muertes a su paso. Dicho suceso provoc replantear la manera de trabajo en el desarrollo de esta investigacion, ya que debido al largo periodo de cuarentena y a las medidas de prevencion para evitar contagios, se descartó totalmente el trabajo de adquisicion de la señal Electroencefalografica con pacientes en edad escolar; por lo cual se inicio la búsqueda de un conjunto de datos elaborado por algún estudio de investigación previo, el cual encajar con los objetivos propuestos dentro de esta investigacion, los cuales buscan identificar el proceso cerebral del lenguaje en pacientes y/o voluntarios en edad escolar.

A continuacion, se relatar todo el trabajo de investigación que se llevo a cabo para cumplir con los objetivos propuestos.

Planteamiento del problema

La comunicación oral y escrita es sin duda alguna una herramienta elemental para el desarrollo y que a nivel personal, se vuelve imprescindible para poder expresar y compartir conocimiento, manifestar emociones, etc. Desgraciadamente, parte de la poblacion mundial padecen dificultades en el desarrollo del lenguaje, que les impiden establecer una correcta comunicacion con su entorno. Estos trastornos de lenguaje, que pueden ser provocados por un mal funcionamiento de la química cerebral, un proceso infeccioso o algún daño severo en el cerebro, dichos trastornos, son diagnosticados por técnicas especializadas, así como basadas en el criterio del médico tratante al examinar los analisis clínicos como el electroencefalograma (EEG) y las diversa pruebas que se realizan.

A pesar del avance tecnológico, no se cuenta con algún equipo que sirva de apoyo para el diagnostico, que permita emplear las técnicas de la inteligencia artificial como el Aprendizaje Activo (Avtive Learning), con el cual se pueda extraer informacion necesaria de la bioseñal provista por el EEG y posteriormente determinar si existe alguna discrepancia entre los trastornos de lenguaje como Disartria, Afasia y Disfasia, para obtener un diagnóstico mas certero.

Dada esta situacion, se ha planteado abordarla desde el área de la ciencia de datos, ya que la naturaleza del problema nos permite explorar y manejar los datos para extraer información relevante; que nos permita diferenciar características dentro de la bioseñal del EEG, como lo son las ondas cerebrales que tiene un papel tanto cognitivo como en los diferentes procesos del lenguaje. Aunado al reciente acontecimiento a nivel mundial, por la pandemia de COVID-19 y como medida de seguridad, se har uso de un conjunto de datos de dominio público para la realización de la etapa experimental del proyecto.

Preguntas de Investigación

Del registro de la actividad cerebral durante las pruebas aplicadas en la etapa de adquisicion de la señal de Electroencefalograma, provistas por el set de datos Child Mind Institute - Multimodal Resource for Studying Information Processing in the Developing Brain (MIPDB) [Simon P. Kelly, 2016].

- ¿Como se puede mejorar el diagnostico de los trastornos de lenguaje, a partir de EEG, empleando técnicas de Aprendizaje Activo (Active Learning)?
- ¿Cual es el mejor método de Aprendizaje Activo (Active Learnig) que permita etiquetar las ondas cerebrales que se involucran en el proceso cerebral del lenguaje?
- ¿Qué aportes nos brinda la técnica de Aprendizaje Activo o/y Desarrollo Experimental Óptimo dentro de los fines de la investigación?
- ¿Qué características representan la actividad cerebral de los pacientes del conjunto de datos usado los cuales fueron obtenidos del estudio elaborado por Child Mind Institute?

Justificacion

En México un 7% de los niños sufre algún nivel de deterioro en el desarrollo del lenguaje, que puede ser provocado por una falta de estimulación neurosensorial adecuada, mas sin embargo, no se descarta la presencia de un daño cerebral como lesiones, mal funcionamiento químico, proceso infeccioso, entre otras afectaciones. De igual manera, con mayor frecuencia se presentan mas casos de este tipo a nivel mundial. La presente investigacion, se centrará en determinar en que medida se puede someter la señal EEG a las técnicas de Aprendizaje Activo (Active Learnig), para determinar si es posible establecer un método de etiquetado eficiente para el proceso cerebral del lenguaje que involucra a las ondas cerebrales Alfa, Beta, Delta y Theta. Esto con el fin, de identificar el proceso cerebral del lenguaje a través de estas etiquetas, así como aprender más del comportamiento de estas alteraciones del lenguaje a nivel cerebral y brindar una herramienta apoyo para el diagnostico médico, así como una herramienta que permita cuantificar el avance de los pacientes en su proceso de terapia.

Hipótesis

Es posible que, mediante una interfaz BCI así como a través de técnicas de tratamiento de la señal EEG, en conjunto con la aplicacion de técnicas de Aprendizaje Activo (Active Learning), entonces se pueda identificar informacion relevante que permita etiquetar los procesos cerebrales del lenguaje; esto a través de la cuantificacion de las frecuencias de las ondas cerebrales Alfa, Beta, Delta y Theta que se desencadenan con la aplicación de los paradigmas Sequence Learning y Naturalistic Viewing (ver figura 3.1).

H1 = El proceso de etiquetado llevado por la técnica de Aprendizaje Activo (Active Learning) nos permite identificar el proceso Cerebral del Lenguaje.

H0 = El proceso de etiquetado llevado por la técnica de Aprendizaje Activo (Active Learning) NO nos permite identificar el proceso Cerebral del Lenguaje.

Objetivo general

Modelar un sistema BCI para que a través de técnicas de Aprendizaje Activo y a partir de señales EEG, permita obtener un buen desempeño en el proceso de etiquetado para identificar el Proceso Cerebral del Lenguaje de la señal EEG entre individuos de edad escolar .

Objetivos específicos

- Evaluar las técnica de Analisis de Componentes Independientes (ICA) y aplicar las más adecuada para la eliminacion de artefactos presentes en el EEG.
- Extraer las características relevantes de la señal EEG en el proceso de concentracion de los pacientes, mediante la Transformada Wavelet Discreta (DWT).
- Reducir la dimensionalidad de los datos mediante la técnica de Análisis de Componentes Principales (PCA).
- Evaluar técnicas de aprendizaje activo, que permitan tener un mejor desempeño durante el proceso de etiquetado de las clases que permitan agrupar de manera eficiente las señales EEG, con el fin de identificar el proceso cerebral del lenguaje en pacientes y/o voluntarios.
- Analizar los resultados obtenidos.

Capítulo 1

Estado del Arte

Debido a los avances y descubrimientos de Hans Berger padre de la electroencefalografía, se planteó el uso de la actividad cerebral del humano como medio de comunicacion y control de su ambiente; esto sin la intervención de los nervios periféricos y músculos. De esta idea inicial, surge el concepto de Interfaz Cerebro Computadora (BCI, por sus siglas en inglés), la cual Hoffmann (2007) define como un sistema de comunicacion que interpreta las señales de la actividad cerebral, transformandola en comandos que pueden ser ejecutados por una computadora u otro dispositivo. Así mismo, el autor establece que para obtener una buena comunicacion entre la interfaz BCI y el usuario, depende la interacción efectiva entre ambos; esta interaccion conlleva un ciclo donde se codifica la actividad cerebral del usuario, se extrae la información, se clasifica y se retorna al usuario una respuesta que puede ser un estímulo o la ejecucion de alguna interaccion con su ambiente.

Construccion de una interfaz Cerebro-Maquina por medio de un traductor de señales neuronales a palabras Dicotomicas textuales

Dentro de los trabajos realizados en la Division de Estudios y Posgrados, del Instituto Tecnologico de Leon, podemos destacar la tesis ”Construccion de una interfaz Cerebro-Maquina por medio de un traductor de señales neuronales a palabras Dicotómicas textuales” [Serna, 2018], la cual busca incorporar un estudio alternativo de comunicación, que por medio del entrenamiento cerebral implementado en una interfaz BMI (Brain Machine Interface); que al incorporar disciplinas tales como las matematicas, electrónica, inteligencia artificial y lingüística, se plantea reducir la cantidad de tiempo del entrenamiento.

El objetivo de esta tesis fue el diseño e implementacion de un modelo BMI capaz de clasificar palabras dicotomicas de manera textual, esto a través del uso de potenciales evocados provocados por el estímulo auditivo a manera de preguntas. La mayor aportacion de este trabajo de investigación, fue la introduccion del concepto del sesgo cognitivo para la clasificacion de palabras obtenidas del pensamiento, que demuestra en base a los experimentos realizados de clasificacion ser una alternativa viable; ya que cada palabra pensada genera una pulsacion eléctrica diferente entre sí, y demuestra que la concentración durante las pruebas facilita la deteccion de las pulsaciones eléctricas.

Análisis y clasificación de electroencefalogramas (EEG) registrados durante el habla imaginada

Debido a la complejidad de la señal del electroencefalograma, y al ser esta la parte esencial de presente trabajo de investigacion; se consult la tesis doctoral ”Analisis y clasificación de electroencefalogramas (EEG) registrados durante el habla imaginada”
[Torres García et al., 2016], dicho trabajo de investigación nos brinda un modelo de procesamiento y clasificación aplicado a la señal de electroencefalograma, el cual consta de la seleccion automatica de características destacadas de la señal EEG, así mismo plantea un filtrado especial que combina varios canales en uno simple para la extracción de la información y finalmente mediante las técnicas de seleccion de canales nos muestra como se puede obtener informacion mas útil del EEG para su interpretacion.

A continuacion, se detallará el estado del arte de la lineas diferentes investigaciones relacionadas con este proyecto de investigación, que se han realizado a través del tiempo.

Hallazgos electroencefalográficos en los pacientes con trastorno específico del desarrollo del lenguaje

Dentro del área de la Neurofisiología, se destaca el articulo arriba mencionado; ya que para los fines de esta investigación, el saber que existe una identificacion de la frecuencia y características de las anormalidades de tipo epileptiformes presentes en un grupo de niños con Trastorno Especifico del Desarrollo del Lenguaje. [Aguilar et al., 2015], nos detalla el proceso de su investigacion, cuyo estudio estuvo compuesto por niños con un retardo en la adquisición del lenguaje, que acudieron al Hospital Pediátrico ”Juan Manuel Marquez”para su valoracion; donde fueron eva-

luados mediante un examen neurologico, pruebas metabolicas de orina, potenciales evocados auditivos, valoración psiquiátrica y psicométrica; para determinar la parecencia del TEDL.

Todos los pacientes fueron sometidos a un EEG durante el sueño espontáneo postpandrial, durante aproximadamente 30 minutos, obteniendo trazos en etapas I y II de sueño sin movimientos oculares. Esto mediante un electroencefalografo digital de 32 canales, de los cuales solo se usaron 19 canales ubicados según el estandar 10-20; con una ganancia de los amplificadores de 1000, una frecuencia de muestreo de 200Hz y filtros con un ancho de banda de 0.5 a 30Hz. Los registros fueron evaluados fuera de linea en montajes monopolares y bipolares por tres especialistas en Neurofisiología clínica. En su articulo [Aguilar et al., 2015], nos relata que no es infrecuente encontrar alteraciones epileptiformes en personas que no hayan experimentado crisis epilépticas; pero resulta de gran interés conocer el verdadero efecto de estas descargas en el EEG de sujetos que no han sufrido ningún tipo de crisis epiléptica, pero que presentan alteraciones del neurodesarrollo en la infancia, como los trastornos de lenguaje, trastornos de aprendizaje, etc.

Los resultados de este estudio, encontraron que de la poblacion evaluada, el 89.3 % de los EEG se diagnosticaron como anormales, cabe destacar que la actividad de las descargas epileptiformes no se dio de manera generalizada; es decir, la localización de la actividad lenta fue en las regiones temporales en un 100 % y tuvo lateralizacion sobre el hemisferio izquierdo en el 80 % de los casos. [Aguilar et al., 2015] concluye, que los hallazgos de su investigacion puede ser posible que las alteraciones en e EEG tengan un impacto en el desarrollo del lenguaje, ya que existe un predominio de las alteraciones en el hemisferio izquierdo; así como la presencia de descargas epileptiformes interictales sobre los lobulos frontales y temporales, dichas estructuras también se involucran ampliamente en la función del lenguaje.

Como seres socialmente activos, el comunicar nuestras ideas, compartir informacion y transmitir nuestra cultura, se da gracias al proceso de comunicacion, ya que es

fundamental para interactuar con nuestro entorno. La falta de estos medios de comunicacion (habla, escritura) limitan en gran medida como hacemos frente a las actividades diarias. En seguida, describiremos un dispositivo tecnologico que rompe con esa barrera de comunicacion, al proporcionar a los pacientes un medio de expresion que facilita la interacción con su entorno.

The thought-translation device (TTD): Neurobehavioral mechanisms and clinical outcome

[Birbaumer et al., 2003] propone en su artículo un traductor del pensamiento (TTD, por sus siglas en inglés) el cual consiste en un dispositivo al que se somete a un periodo de entrenamiento y es auxiliado por un software de deletreo (spelling program). El dispositivo TTD basa su funcionamiento en la señal SCP (slow cortical potential) que a comparacion de otras señales EEG posee dos ventajas significativas:
[Birbaumer et al., 1990]

- Estan presentes en la actividad cerebral de todas las personas, a pesar del diagnostico o padecimiento que se tenga.
- Esta señal fisiológica tiene una duración de entre 500 ms hasta 10s, lo que le permite ser mejor interpretada por los mecanismos neurofisiologicos.

En su artículo [Birbaumer et al., 2000] describe detalladamente el proceso de adquisicion de la señal SCP; donde hace uso de de un amplificador EEG de 8 canales tomando la señal del electrodo Cz (vertex) a una frecuencia de muestreo de 256 Hz. Una vez extraída la señal, es filtrada para aislarla del movimiento ocular; para posteriormente la señal ser convertida en un estímulo visual para el paciente, que es mostrado como una esfera de luz, la cual se desplaza a través de la pantalla.

A 20-questions-based binary spelling interface for communication systems

Una de las aportaciones mas significativas dentro del área de la comunicacion a través del EEG, fue propuesta por [Tonin et al., 2018], los cuales desarrollaron una interfaz BCI basada en la espectroscopia de infrarrojo cercano, la cual almacena un set de preguntas basada en el juego de las 20 preguntas, que permitirá al paciente comunicarse el responder utilizado solo respuestas "SI o NO". Esta interfaz, implementa una Red Neuronal Artificial para el proceso de clasificacion, la cual estima una declaracion pensada por el paciente en base a la respuesta de al menos 20 preguntas, estas declaraciones son frases que también estan en un set de datos. Los resultados de la experimentacion, demuestran que este sistema basado en 20 preguntas, puede ser una interfaz valida para cualquier BCI que utilice una señal lenta como la espectroscopia de infrarrojo cercano o con una tasa de precisión baja, ademas puede ser aplicado en una interfaz basada en EEG, ya que este sistema mejora el rendimiento al predecir frases enteras utilizando al menos 20 entradas binarias. El único inconveniente es que el sistemas solo puede predecir las frases que están almacenadas en la base de datos, por lo que el paciente no será libre de formular sus propias frases.

Análisis de Señales Electroencefalográficas para la clasificacion del Habla Imaginada

Dentro del área del análisis de la señal electroencefalografica, podemos resaltar la brillante aportación de [Torres García et al., 2013], donde a través la interpretacion de la información extraída del EEG, provenientes de la región del modelo lingüístico de Geschwind-Wernicke, donde se logr la clasificación de cinco palabras mediante el registro del habla imaginada. Dicho trabajo de investigacion, emplea un método de tratamiento y clasificacion de la señal EEG, que consta de la eliminacion del

ruido de la señal al aplicar la técnica de referencia promedio común, enseguida se da la extraccion de las características usando el modelo de la transformada Wavelet discreta (DWT, por sus siglas en inglés), esta técnica permite efectuar un modelado de las variaciones de la señal EEG, en dominio tiempo-escala, que como resultado nos brinda una representacion eficiente al restringir la variacion en la traslación de la escala. Para el proceso de clasificacion se probaron los clasificadores Naive Bayes, Random Forest y Máquina de Vector Soporte, de los cuales se obtuvieron los siguientes resultados:

- Del Cross Validation empleando los electrodos de la región lingüística de Geschwind-Wernicke.

  El mejor porcentaje de exactitud obtenida de esta validación cruzada, fue alcanzado por el algoritmo de clasificación Random Forest, teniendo una puntacion de 44.43 %.

- Del Cross Validation empleando los electrodos de la región lingüística de Geschwind-Wernicke más los 10 canales restantes.

  De este proceso de clasificación, el mayor porcentaje de exactitud fue del 60.11 %, logrado por el algoritmo de clasificacion Random Forest, empleando 50 árboles, y una validacion cruzada de 10 pliegues.

Specificity of spontaneous EEG associated with different levels of cognitive and communicative dysfunctions in children

Con respecto al estudio de los trastornos del lenguaje, el articulo de [Kozhushko et al., 2018], se centra en el estudio de las habilidades cognitivas y déficit comunicativos en niños con Trastorno del Espectro Autista (TEA); donde mediante el análisis de potencia espectral del EEG en un estado de reposo, buscan encontrar la correlacion de la actividad cerebral anormal con la severidad de la disfunción cognitiva y de comuni-

cacion de los niños con TEA.

Para realizar este analisis, se usaron 19 electrodos colocados en base a el estándar 10-20, a la señal obtenida se le removi parpadeo de los pacientes empleando la técnica de Analisis de Componentes Independientes (ICA, por sus siglas en inglés) con un filtrado de 0 -2 Hz para ondas lentas y de 20 - 35 Hz para ondas rapidas. Para realizar el análisis de la potencia espectral, se tomaron las frecuencias theta, alpha y beta, se normalizaron los valores de estas frecuencias usando el logaritmo decimal; posteriormente se realiza una prueba ANOVA para comparar por pares cada electro, teniendo como valor de p=0.0002; se empleo la prueba no perimétrica Spearman's para comprobar la correlacion entre el déficit cognitivo y el déficit fisiologico de los pacientes. Para poder tener una mejor visualizacion de los datos, el autor explica el uso de un software conocido como sLORETA (Low Resolution Electromagnetic Tomography)[Pascual-Marqui, 2002]. Como resultado de este trabajo de investigacion, se comprob en los pacientes con TEA la relacion entre las frecuencias theta/alpha y las frecuencias theta/beta, siendo la frecuencia theta la que genera un mayo afectacion en combinación con las otras.

A resource for assessing information processing in the developing brainusing EEG and eye tracking

Para la fase experimental de esta investigacion, y debido a la pandemia de COVID-19 y a las medidas tomadas para la prevención de contagios se acord con el comité evaluador, el uso de un set de datos de dominio público; el cual fue desarrollado por el Child Mind Institute de los Estados Unidos. En su artículo [Langer et al., 2017], nos describen de manera detallada el proceso de creacion de este set de datos; cuyo objetivo es proveer un recurso de informacion, que permita desarrollar herramientas pera el mejor desarrollo cerebral o el diagnóstico de patologías.

El National Institute of Mental Health ha tomado un papel de liderazgo, en el esfuerzo por establecer el Research Domain Criteria (RDoC); como marco de referencia para la caracterización de enfermedades mentales. El principal aspecto de este marco de referencia, es la integración de información de múltiples niveles; así como el reconocimiento de imagenes, unidades de medida y potenciales de la actividad neurofisiologica. [Langer et al., 2017] Nos presenta una serie de pruebas para estimular la actividad cerebral de manera activa y pasiva; que incluye tres paradigmas que evalúan distintas tareas. El primero de ellos permite rastrear el procesamiento cerebral durante la toma de sesiones simples; el segundo paradigma involucra la secuencia de aprendizaje de tareas simples, finalmente, el tercer paradigma evalúa la velocidad del procesamiento cerebral durante la ejecución de tareas.

Este estudio estuvo integrado por 126 voluntarios, cuyas edades van de los 6 años hasta los 44, los cuales fueron reclutados del Child Mind Institute. De esta poblacion que participó en el estudio, el 80.2 % son individuos con un desarrollo normal, mientras que, el 19.8 % fueron diagnosticados con uno o mas desordenes clínicos. Del total de los participantes el 54.8 % son del sexo masculino y el 45.2 % son del sexo femenino. Previo al estudio en el laboratorio, los pacientes fueron evaluados mediante una entrevista telefonica, esto con el fin de confirmar su elegibilidad para el estudio, así como cuestiones de seguridad.

El estudio completo tiene una duracion de cinco horas, en las cuales el participante podría elegir si dividir el estudio en sesiones o realizar el estudio completo. Para la adquisicion de la señal electroencefalografica, [Langer et al., 2017] se uso un equipo EEG Geodesic Hydrocel de 128 canales; y se aseguró la buena impedancia de los electrodos, al mantenerlos por debajo de 40 kOhm; esto con el fin de captar una señal EEG de mayor calidad. Así mismo, un electrodo podría ser identificado como un mal electrodo, al obtener una varianza de más de tres desviaciones estandar, de la media en comparacion con los otros electrodos.

Para asegurar una mejor integridad de la señal EEG [Langer et al., 2017], realizó la correccion o eliminación de artefactos (movimientos involuntarios, latidos del corazon, latidos, etc.). En dicho proceso, realizó una fase de filtrado haciendo uso de un filtro de de respuesta finita a impulsos (Hamming windowed-sinc finite impulse response). Los artefactos oculares fueron eliminados a través la regresión lineal de los canales EOG de los canales EEG del cuero cabelludo. Para eliminar el ruido que pudiera existir en la señal EEG, se aplicó un robusto algoritmo de Analisis de Componentes Principales (PCA); así como el algoritmo Aumentado de Multiplicadores de Lagrange (ALM100).

Efficient Labeling of EGG signal Artifics using Active Learning

Es durante este suceso de la pandemia de COVID-19, donde se reestructura el proceso de experimentacion de la investigación, el cual fue analizado y aceptado por el comité que integra esta investigacion. Por tal motivo, se realiza una búsqueda literaria, donde nos hemos encontrado con el artículo Efficient Labeling of EGG signal Artifics using Active Learning [Lawhern et al., 2015], en el cuál el autor no presenta un método para mejorar la extraccion de artefactos empleando el Aprendizaje Activo (Active Learning), dicho método consiste en modelo Query-by-Committe (QBC, por sus siglas en inglés); este algoritmo toma una decision a través de una votación para identificar los datos, mientras que aquellos datos que no logran una votación definida pasan a ser identificados por el programador (Oracle). Es importante mencionar que si bien se ha demostrado que el Aprendizaje Activo ha obtenido buenos resultados en trabajos de investigacion como Spam Filter o en la clasificación de imagenes; su aplicacion en trabajos de investigación relacionados con la señal EEG no ha sido ampliamente explorado.

Dentro de este trabajo de investigacion, el autor [Lawhern et al., 2015] adquiri la señal de 7 voluntarios del Army Research Laboratory; a los cuáles se les pidió que

ejecutaran dos tareas la primera de ellas, completaron una prueba donde tenían que realiza ocho tipos de movimientos definidos como pestañeos, movimientos oculares, movimiento de los músculos de la mandíbula, músculos de las cejas y giros del cuello, así como la condicion nula al no generar ningún movimiento; la segunda tarea consistía en realizar una tarea de discriminacion de imagenes en la cual se presentaban imagenes de enemigos, así como imágenes de soldados amigos.

Mientras se realizaban dichas tareas, se recolectaba la señal EEG, usando un electroencefalograma de 64 canales a una frecuencia de muestreo de 512Hz, haciendo la referencia de los canales en los mastoides; posteriormente la señal fue ajustada (down-samplig) a 256Hz y pasada por un filtro pasa altas a 1 Hz; es importante destacar que este procesamiento de la señal se llevo acabo con el EEGLAB Toolbox.

Para la ejecucion del Aprendizaje activo, el autor [Lawhern et al., 2015] consideró dos formas, la primera de ellas uso un modelo Query-by-Committe estándar y un modelo K committee a través de K-fold cross-validation, como se muestra en la figura 1.1.

Los resultados de este trabajo de investigación arrojaron que, se pude obtener un buen porcentaje de clasificacion al clasificar menos del 25 % de los datos; lo que sugiere un ahorro significativo del tiempo, en comparacion al tiempo obtenido cuando se realiza el etiquetado de los artefactos de manera manual un una gran grupo de datos.

1. Use K-fold cross-validation to train K models $C_1, \ldots, C_K$ on L.
2. For each $C_i, i = 1, \ldots, K$, predict the labels for all epochs in U.
   a. If using Decision-Confidence QBC, calculate the confidence for each C_i for all epochs in U.
3. Sort all epochs in U by the amount of disagreement among the committee. For example, if $K = 5$, then a single epoch with 2 committee members in agreement is more severe than a single epoch with 3 committee members in agreement.
   a. If using Decision-Confidence QBC, take the sum of all the confidence values for each of the K classifiers. The sum of all confidence values will range from $[0, K]$.
4. Select up to M epochs with the greatest amount of committee disagreement for oracle labeling. Remove the M epochs from U and add the M epochs into L.
   a. If using Decision Confidence QBC, within each level of disagreement the lowest summed confidence epochs are selected first.
5. Re-train $C_1, \ldots, C_K$ on L using the new data labeled in Step 4.
6. Using the learned model, calculate the classification accuracy on the validation set V.
7. Repeat steps 2-6 until desired level of convergence is obtained.

Figura 1.1: Pseudocodigo para el Aprendizaje Activo [Lawhern et al., 2015]

Oscilo-patología en trastornos del espectro Autista: Las ondas Cerebrales en los procesos del lenguaje

Con el pasar de las décadas y el avance tecnologico que se genera en diferentes ramas de la ciencia, la medicina, entre otras; han contribuido a que dichas disciplinas se combinen y colaboren entre sí para buscar nuevas formas de estudio y avances tecnologicos para abordar diferentes topicos. Prueba de ello tenemos el articulo Oscilo-patología en trastornos del espectro Autista: Las ondas Cerebrales en los procesos del lenguaje' [Morales, 2020].

En este artículo, el autor nos brinda una recopilación de diversas fuentes de informacion sobre la importancia de las ondas cerebrales en los procesos del lenguaje normal y patológico. Uno de los principales aportes es la información que nos brinda con respecto al papel que juegan las ondas cerebrales en el proceso del lenguaje, en donde nos presenta un modelo conocido como Dynamic Cognomics, que describe este proceso, como se muestra en la figura 1.2. [Morales, 2020]

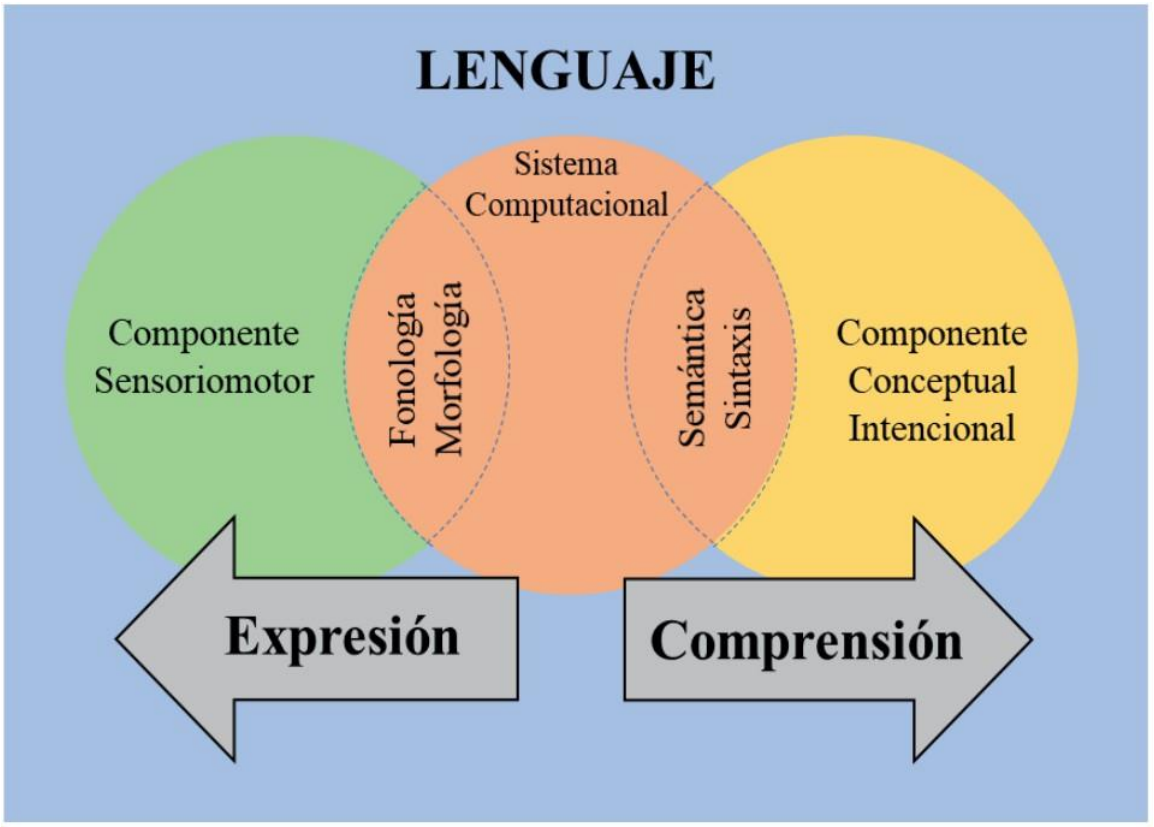

Figura 1.2: MOdelo Dynamic Cognomics [Morales, 2020]

El autor describe este modelo como un sistema que organiza las ondas cerebrales en parejas que emplean frecuencias cruzadas, este intercambio se interpreta como un potencial de accion la cual inicia la actividad cognitiva de comprension lingüística; el autor detalla que este modelo utiliza la estructura minimalista del lenguaje planteada por Chomsky, el cual cuenta con dos componentes, el primero de ellos es el componente Cognome o componente de la conceptualización que se encarga de la interpretacion; el segundo componente se le conoce como Dynome o componente

sensorio-motor encargado de la expresion [Morales, 2020].

Es dentro de este modelo que ocurre el cruce de las onda cerebrales, el autor nos explica que el cruce de las ondas Alfa y Gamma activan los sistemas viso-espaciales, que son importantes para los procesos lingüísticos escritos. También se argumenta dentro de este articulo que diversos estudios de electroencefalografía, han indicado que las ondas Theta al entrar en la porcion del lobulo temporal, activan procesos de comprensión ligústica que requieren de la memoria. Otro ejemplo es el cruce de las ondas Gamma, Beta y Alfa las cuales activan los proceso de etiquetado y categorizacion de palabras. El autor concluye que este modelo explica como las ondas cerebrales mutan y pasan de una banda de frecuencia a otra, lo que concuerda con la definición física de onda: la energía no se crea ni se destruye, solo se transforma [Morales, 2020].

Dentro de las conclusiones de este articulo, se destaca el rol protagónico de las ondas cerebrales para los diversos procesos cognitivos del cerebro humano, entre ellos el proceso cerebral del lenguaje; ya que el comportamiento de estas ondas cerebrales tanto de manera individual como el cruce y transformación de estas da lugar a los diferentes procesos lingüísticos.

A continuación, se desglosar los temas que auxiliaron el desarrollo de esta investigacion dentro del marco teórico.

Capítulo 2

Marco teórico y Conceptual

Anatomía Cerebral

Cerebro

En su libro, [Von Neumann et al., 1999] nos detalla el funcionamiento cerebral, partiendo del componente basico del sistema nervioso, la neurona, la cual genera y propaga el impulso nervioso por toda la masa encefalica. Dicho impulso, combina una variedad de procesos eléctricos, químicos y mecanicos, lo que denota su complejidad [Lopez et al., 2008]. Describe a la neurona como un cuerpo celular del cual se originan directa o indirectamente ramas, denominadas dendritas, que se especializa junto con el soma en recibir el impulso nervioso, mientras que los axones son los encargados de propagar el impulso nervioso; el aspecto principal con el que se identifica a este impulso nervioso, es con una perturbación eléctrica. Que consiste en un potencial eléctrico de alrededor de 50 milivoltios y con una duracion de un milisegundo.[Von Neumann et al., 1999]

Este organismo ha atravesado por un largo proceso de evolucion, como lo explica [Florian, 2010] en su libro. El proceso evolutivo del cerebro no fue contante, ya que la evidencia fósil sugiere que hubo periodos cortos de rapida evolucion, seguidos de largos periodos de estancamiento. Sin embargo, estos periodos de estancamiento,

permitieron asegurar la estructura neuronal y cimentar una mejor imagen mental del mundo exterior, así como un mejor procesamiento de datos. En seguida, describiremos los estímulos que desencadenan la actividad neuronal, cuya funcion es la transmision de la información entre neuronas, el cual se denomina sinapsis.

Sinapsis Química

En este tipo de sinapsis, no existe una union estructural entre la célula presinaptica y la postsinaptica, las cuales están separadas por un espacio intracelular de aproximadamente 20 nm, denominada hendidura sinaptica. En este tipo de sinapsis, le célula presinaptica libera a la hendidura sinaptica una biomolécula, llamada transmisor, el cual se unirá al receptor proteico de la célula postsinaptica, lo que desencadena cambios en la permeabilidad de célula postsináptica. [Vicario, 1999]

En un principio, se pens que la neurona solo podía liberar un único tipo de transmisor, e implícitamente todos los terminales sinápticos de la misma neurona liberarían el mismo tipo de transmisor, a este concepto se le conoce como el principio de dale, dicho principio creía que la neurona tendería a la excitacion o a la inhibicion dependiendo de la sustancia que liberase, sin embargo, una misma neurona puede cumplir ambas funciones, aunque esta libere la misma sustancia transmisora. [Vicario, 1999]

Figura 2.1: Sinapsis Química

Sinapsis Eléctrica

En este tipo de sinapsis, no existe una diferencia entre las células presináptica y la postsináptica, ya que la hendidura sinaptica es muy estrecha, de alrededor de 20 nm de separacion, además, existe la presencia de uniones llamadas canales de gap-junctions, los cuales funcionan como vías de alta conductividad eléctrica, lo que permite la despolarizacion o la hiper-polarización de la neurona. [Cardinali, 2007]

La sinapsis eléctrica, no tiene retardo sináptico y es bidireccional, aunque esta bi-direccionalidad se puede ver limitada por resistencia entre ambas células sinapticas, las sinapsis eléctricas son menos frecuentes, aunque se encuentran diseminadas por todo el sistema nervioso central. [Cardinali, 1991]. En la figura 2.2

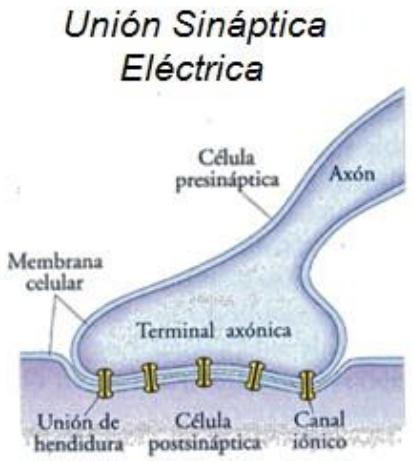

Figura 2.2: Sinapsis Eléctrica

Trastornos de lenguaje

La perdida del lenguaje ocasionado por algún factor como una lesión cerebral, la presencia de algún tumor o cierto proceso infeccioso, es estudiado por una rama de la ciencia llamada Neurolingüística; la cual se compone del estudio del procesamiento del lenguaje en el cerebro; donde los neurólogos son los encargados del estudio del cerebro y el sistema nervioso en especial en la presencia de alguna lesión.

Mientras que los lingüistas se encargan del estudio de la estructura del lenguaje. [Obler et al., 2001]

A continuación, detallaremos los trastornos de lenguaje que se abordaran en el presente trabajo de investigacion.

Disartria

La Disartria, es el resultado de la activación neuromuscular anormal de los músculos articuladores del habla, ocasionando una afectación en la velocidad, fuerza y sincronizacion de dichos músculos. Una manifestacion evidente de este trastorno es la distorsion de los sonidos, esto es relacionado con una disfunción del sistema nervioso central, particularmente en la union neuronal y muscular, lo que también puede provocar déficit sensoriales. Este trastorno de lenguaje puede ser diagnosticado, mediante el analisis de las regiones especificas del deterioro del habla, o la identificacion de un trastorno neurologico. Así mismo, se le realiza al paciente un examen del habla, donde se hace la repeticion de sílabas, palabras y frase, para comprobar el cambio de condonacion de los sonidos consonánticos. [Bradley et al., 2009]

Afasia

Cuando se presenta un daño generalizado en el hemisferio izquierdo, los pacientes pueden presentar una deficiencia lingüística severa; que puede ir desde la articulación de unas pocas palabras o silabas, hasta la inhabilidad del habla completa, a esto se le conoce como ”Afasia Global” [Obler et al., 2001]. En su libro [Bradley et al., 2005] describe los síntomas de la afasia que pueden ayudar a su diagnostico, de los cuales se encuentran el mutismo o perdida total del habla, lo que refleja una afasia grave; cabe mencionar que este síntoma puede presentarse en otro tipo de trastorno de lenguaje llamado Disartria que se describi anteriormente. Otro síntoma importante

datado por el autor, al someter al paciente a prueba de habla y escritura, este realiza

un notable esfuerzo para vocalizar y existen indicios de déficit de comprensión, son indicios alarmantes de un trastorno afasico. El habla titubeante al empezar a hablar expresa una dificultad en la búsqueda de palabras; la anonimia o la incapacidad de reproducir un nombre en especifico manifestando pausas al hablar, es un indicador fiable de la presencia de este trastorno de lenguaje. Bradley (2005) remarca que el síntoma fundamental del trastorno afasico es la falta de comprension de las conversaciones, así como la comprension y reproduccion del lenguaje escrito.

Dentro de este trastorno de lenguaje, podemos encontrar síndromes que se manifiestan de acuerdo al nivel de daño en el hemisferio izquierdo, a continuación, describiremos algunos de los síndromes mas comunes derribados de este trastorno del lenguaje, los cuales son mencionados en el libro "El lenguaje y el cerebro" [Obler et al., 2001].

- Afasia de Broca: Descubierta por el neurologo Francés Broca, la cual se caracteriza por una dicción lenta, vacilante y a frecuentemente se presenta la omision de marcadores gramaticales, pero la compresión del lenguaje se mantiene intacto.

- Afasia de Wernicke: Descubierta por el neurólogo Alemán Carl Wernicke, el paciente presenta un habla fluida, aparentemente normal, pero con la inclusion de estructuras semanticas inusuales llamadas "circunloquios", en lugar de usar palabras simples, y a diferencia de la afasia anterior, los pacientes presentan una severa falta de comprension del lenguaje.

- Afasia de Conducción: Cuyo síntoma más relevante es la incapacidad de reproducir el lenguaje hablado.

Disfasia o Trastorno Específico del Desarrollo del Lenguaje (TEDL)

Este trastorno de lenguaje, se presenta con la dificultad para la adquisición y manejo de la comprensión y expresión del lenguaje; con la singularidad de no presentar ningún déficit neuromotor, cognitivo o socio emocional, por lo que se asocia a un problema intrínseco del procesamiento del lenguaje. Un factor dominante de esta trastorno del lenguaje, es la parecencia de dificultad del aprendizaje secuencial verbal, como los sonidos de las palabras, la formacion de palabras, aprendizaje de vocabulario, etc. [Albesa and Ayala, 2017]

Dicho trastorno de lenguaje tiene un origen poligénico, es decir, causado por la accion simultanea de varios genes, siendo los varones los mas afectados por este trastorno. Se manifiesta a inicios del desarrollo lingüístico del niño, siendo una característica predominante el retraso y distorsion del lenguaje; otro marcador útil para diferenciar este trastorno es la repetición de pseudo palabras sin significado alguno, lo que evidencia las dificultades del procesamiento lingüístico. Albesa (2017), señala seis criterios clínicos que son críticos para la pronta intervención de especialistas, que se detalla a continuacion. [Albesa and Ayala, 2017]

1. Ninguna palabra inteligible a los 18 meses de edad (ademas de 'papa/mama').

2. Falta de desarrollo de protodeclarativos como la señalizacion.

3. Limitadas respuestas de intencionalidad compartida.

4. No asocia dos palabras en un enunciado a los 2.5 años.

5. Vocabulario limitado a una cuantas palabras a los 3 años.

6. Enunciados de solo dos palabras a los 4 años.

Neurociencias del Lenguaje

Esta disciplina estudia la organización del lenguaje en el cerebro, cuyas raíces se remontan décadas atras con los diversos estudios que se enfocaban en descubrir las bases neurologicas del lenguaje, a través de métodos de observacion, mediante autopsias de cerebros dañados para encontrar la relacion entre las áreas dañadas y los trastornos lingüísticos que haya tenido en vida el paciente. La Neurociencias del Lenguaje, se considera una disciplina joven, debido a como la Neurociencia aborda los temas de investigación, tanto en el enfoque como en la metodología de los proyectos de investigacion. [Vega, 2012]

Uno de los factores claves que marcan la diferencia entre la Neuropsicología clásica y la Neurolingüistica clasica, es el avance tecnologico que se ha visto en las dos ultimas décadas; las cuales han surgido grandes invenciones, que permiten visualizar el funcionamiento del lenguaje mientras el paciente hace uso de este en tiempo real. Otro de los factores, que marcan de diferencia en esta disciplina, es le desarrollo de modelos cada vez mas detallados de la estructura y organizacion de todos los componentes del sistema de procesamiento del lenguaje; los cuales han sido creados por parte de las áreas de la Psicolingüística. Dichos modelos son fundamentales en el proceso de exploración de la organización neuronal del lenguaje, ya que estos nos brindan la informacion necesaria para poder interpretar los datos proporcionados por las técnicas de neuroimagen y electroencefalografía.[Vega, 2012]

Estas aportaciones que trajo consigo la Neurociencia del Lenguaje, lograron demostrar que en el proceso del lenguaje intervienen muchas mas áreas cerebrales, de las que en un principio se creían. Lo que llev a la comprensión de que el proceso cognitivo cerebral se lleva acabo a través de redes neuronales, que se expanden por amplias zonas del cerebro; lo que descarta la creencia clasica de que el cerebro estaba organizado por centros o módulos, los cuales eran responsables de determinados procesos lingüísticos. [Vega, 2012]

Son todos estos descubrimientos y aprendizajes, los que han permitido a la Neurociencia del Lenguaje, ser una ciencia interdisciplinaria como la lingüística, psicolingüística, neuropsicológa o la inteligencia artificial; cuyo ideal es contar con modelos de procesamiento lingüístico que interpreten todas las actividades del lenguaje, que permitiran encontrar todos los componentes de la funcion cerebral, y finalmente predecir y explicar los trastornos afásicos en funcion de esos modelos lingüísticos y neurologicos. [Vega, 2012]

Comprension Oral del Lenguaje

Como seres socialmente activos, la expresión oral es el medio de comunicacion fundamental; que permite la expresión como la comprensión de pensamientos, ideas, sentimientos. La comprension oral, es un proceso sofisticado que requiere la participacion de múltiples procesos cognitivos, los cuales pueden verse afectados por diversos factores que complican en gran medida el proceso de comprensión oral. Uno de estos factores es el ruido ambiental que acompaña el mensaje lingüístico, ya que una de las primeras tareas de este proceso cerebral es la separacion de la informacion lingüística de otros estímulos auditivos que llegan al oído al mismo tiempo. Otro de los factores que afectan el proceso de comprensión oral, es el hecho de que el lenguaje oral es un proceso continuo, que carece de fragmentacion como la lengua escrita; dicha fragmentacion del estimulo lingüístico supone en si toda una dificultad en el proceso de comprensión. [Vega, 2012]

Para poder decodificar y entender el mensaje oral, el oyente debe realizar varios procesos cognitivos, como lo son:

- Acústico, en el que se analizan las propiedades físicas de la onda de sonido.
- Fonético, en el cual se identifican los rasgos fonéticos de esos sonidos.

- Fonológico, en el que se clasifican los segmentos fonéticos identificados en el proceso anterior como fonemas de la lengua oyente. [Vega, 2012]

Una vez completados estos tres niveles de procesamiento, las siguientes operaciones se dirigen al reconocimiento de palabras que componen el mensaje; lo que da pie a la segmentación del habla y la identificacion de palabras que forman las diferentes secuencias de fonemas. Culminado así, en las operaciones destinadas en acceder al significado de esas palabras. A pesar de la basta acumulación de conocimiento sobre el sistema de procesamiento de lenguaje, fue hasta hace poco que se logró integrar datos neuropsicologicos y psicolingüísticos; que gracias a los avances de la neuroimagen, se logr crear un modelo que permite comprender la organización cortical de la comprension oral. Este modelo es conocido como Hickok y Poeppel, el cual describe que los codigos sensoriales del habla deben de interactuar con dos sistemas: el sistema conceptual y el sistema motor-articulatorio.

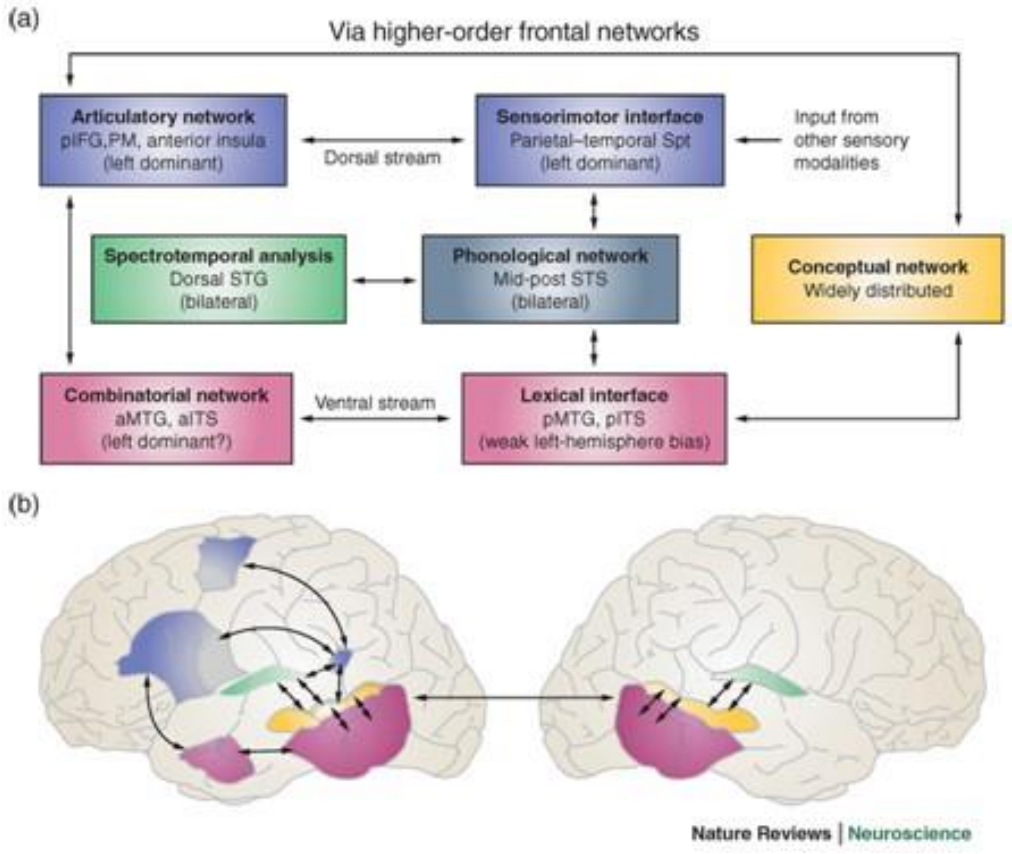

Figura 2.3: Modelo Hickok y Poeppel.[Poeppel and Hickok, 2004]

Bioseñales

Los organismos vivientes, estan compuestos de un conjunto de sistemas, los cuales llevan a cabo diferentes procesos fisiológicos, que en conjunto, realizan tareas importantes para mantener con vida al organismo, como lo es el transportar nutrientes, oxigenar y desechar toxinas. Estos procesos fisiológicos son fenómenos complejos, que estimulan y controlan las entradas y salidas de la materia física, neurotransmisores o procesos químicos, los cuales provocan una accion mecánica, eléctrica o bioquímica.Una vez comprendido el funcionamiento de estos procesos fisiologicos, se vuelve posible observar y medir el efecto que produce, esto se facilita, ya que los efectos de este funcionamiento son detectados en la superficie del cuerpo, lo que permite medir de manera cuantitativa con instrumentacion adecuada. [Rangayyan, 2015]

Ondas Cerebrales

Las señales neuronales dependen de las propiedades eléctricas de la membrana celular, dependiendo de la region examinada, las neuronas presentan un potencial de reposo y cuatro tipos de señales eléctricas: [Cardinali, 1991]

1. Señal de entrada

2. Señal de integración

3. Señal de conducción

4. Señal de salida

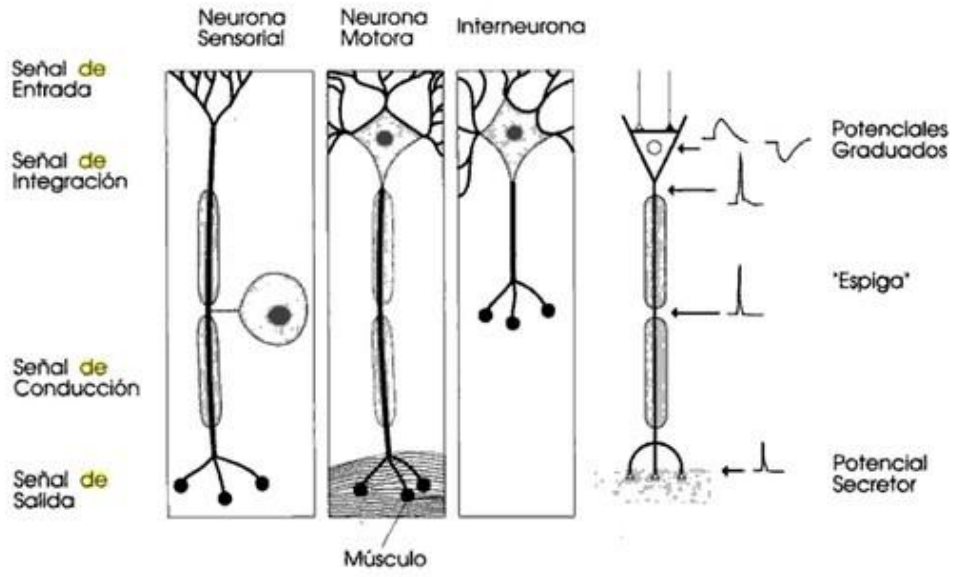

Figura 2.4: Tipos de señales [Cardinali, 1991]

En 1929, Hans Berger registra por primera vez, a través de la invencion de electroencefalograma la actividad eléctrica cerebral; años más tarde, Loomis, Harvey y Hobart sistematizaron esta actividad eléctrica y la dividieron en cinco grandes grupos que determinan diferentes planos de conciencia. [Martínez Perez, 2009]

Con estos grandes descubrimientos, y desde el uso de la electroencefalografía, se logr describir mas profanamente a estos cinco grandes grupos. Los cuales se categorizaron según su banda de frecuencias y estas se relacionan con diferentes estados o funciones cerebrales. [Malik and Amin, 2017] y [Vega, 2012], describe estos grupos de la siguiente manera.

- Ondas Gama: Tiene más de 30 oscilaciones por segundo, La alta actividad gamma en las ubicaciones temporales est asociada a los procesos de memoria.

  Debido a la pequeña amplitud y a la alta contaminacion por artefactos musculares, las ondas gamma estan subestimadas y no se han estudiado ampliamente en comparación con otras ondas cerebrales lentas.

- Ondas Beta: Oscilan entre 14 y 30 Hz de frecuencia por segundo. Se encuentran mayormente en las regiones frontales y centrales del encéfalo, dichas

ondas corresponden al estado de vigilia e implica la accion espontanea tanto física como mental. La potencia de esta onda aumenta en los sitios occipitales durante las tareas de discriminación espacial y la atencion visual en los participantes de alto rendimiento, tanto e jovenes como en adultos.

- Ondas Alfa: Oscilan entre siete y 14 Hz de frecuencia por segundo, se original en la parte posterior del encéfalo, se caracterizan por la sensación de paz y quietud, por lo que se usan mucho en la meditacion y relajación. Esta onda se subdivide en alfa inetriro y en alfa superios; y esta onda cambia con la carga durante la retención de la memoria de trabajo.

- Ondas Theta: Oscilan entre 4 y 7 Hz de frecuencia por segundo, se originan en las regiones temporales, y provocan un estado profundo de relajacion, y son mas comunes en los niños que en los adultos. En los adultos, la alta actividad theta frontal est relacionada con la falta de respuesta al tratamiento antidepresivo en los pacientes con depresion.

- Ondas Delta: Oscila por debajo de los 4 Hz de frecuencia, presentes en todas las regiones cerebrales, se caracterizan por encontrarse en las etapas mas profundas de sueño no REM. Los componentes de baja frecuencia del EEG, especialmente las bandas delta, son los principales contribuyentes al pico P300 de potenciales relacionados con eventos (ERP). El P300 es un indicador ampliamente estudiado y conocido del procesamiento cognitivo.

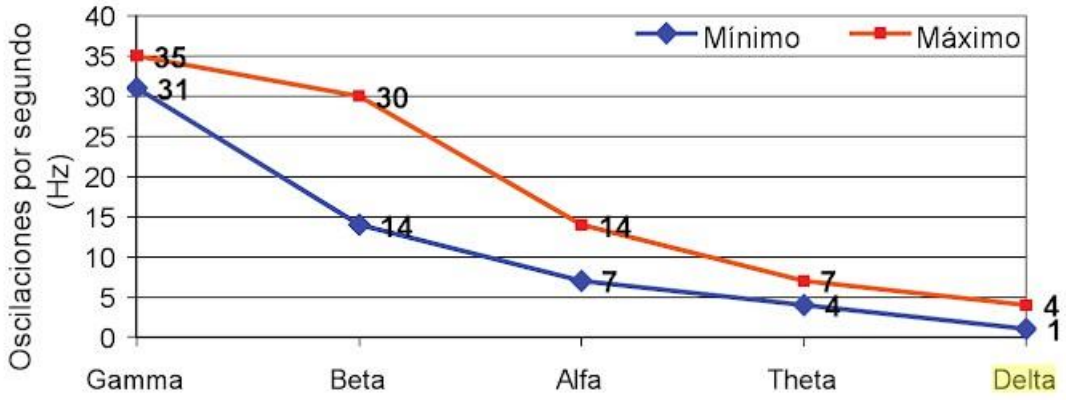

Figura 2.5: Ondas Cerebrales

Técnicas de Electroencefalografía y Neuroimagen

La aparición de técnicas electroencefalográficas y de neuroimangen es de suma importancia en el estudio de la base neurologica del lenguaje, ya que permite observar al momento la activación cerebral de las personas sanas, mientras realizan una actividad lingüística. Gracias a ello, se han confirmado muchos de los hallazgos obtenidos de los estudios previos con pacientes; así como se han desestimado muchos otros hallazgos. [Vega, 2012]

La resonancia magnética (RM), es un procedimiento que presenta mediante imágenes la disposición de protones de una muestra dada. Se basa en la propiedad física de los protones de absorber y devolver ondas de radiofrecuencias cuando se encuentran orientados en un campo magnético. En cambio, la resonancia magnética funcional (RMF) agrega mas características de contenido, modo, temporalidad y control para identificar de mejor manera las tareas particulares que realiza el individuo. De los cuales se tienen dos paradigmas para RMF: pasivos y activos, el primero solo requiere que el paciente se mantenga quieto y ver, sentir u oír; mientras que en el segundo paradigma, se requiere que el paciente realice una actividad voluntaria. [Zuleta, 2007]

Dentro de la disciplina de la Neurociencia del lenguaje, existen dos técnicas de neuro-

imagen mas usadas; una de ellas es la Tomografía por Emision de Positrones (TEP), que se basa en la deteccion de marcadores radioactivos integrados en agua, que se inyectan en la sangre; esta se diluye en la sangre y llega a todo el cuerpo incluido el cerebro, la tomografía detecta ese marcador resaltando las zonas del cerebro con mayor concentracion de sangre, que es consecuencia de una mayor actividad cerebral. Otra de las técnicas usadas en la Neurociencia del lenguaje es la Resonancia Magnética Funcional (RMF), que visualiza la actividad cerebral directamente a través de los cambios de concentracion de oxigeno que tiene la sangre, es decir, detecta el aumento de oxihemoglobina de una determinada área cerebral, esto a través de sus propiedades magnéticas, con lo que se logra contrastar las zonas ricas en oxihemoglobina con las regiones de flujo sanguíneo normal.[Vega, 2012]

Una de las técnicas más usada en las investigaciones y detecciones de patologías neurologicas es la Electroencefalografía (EEG), ya que esta técnica es la menos invasiva, ya que registran as corrientes eléctricas generadas por la actividad cerebral. Esto a través de electrodos sobre el cuero cabelludo, que recolectan y amplifican las corrientes eléctricas de amplios grupos de neuronas y se corrobora en que áreas del cerebro se tiene una mayor actividad cerebral. [Vega, 2012]

La Electroencefalografía es el registro de la actividad producida por las células cerebrales, como resultado de la suma de los potenciales sinápticos excitatorios e inhibitorios de las neuronas [Mayor et al., 2013]. En la práctica clínica, la adquisición simultanea de la señal EEG de varios electrodos (canales), ubicados en diferentes regiones del cuero cabelludo, son empleados para el analisis de la actividad cerebral y el diagnostico clínico. [Rangayyan, 2015] Con respecto a la electroencefalografía, se estableci el Sistema Internacional de Medicion 10-20, que consta de un mínimo de veintiún electrodos, de los cuales 19 son electrodos craneanos y dos electrodos diferenciales. Los electrodos puestos de lado izquierdo se identifican con números impares, y los electrodos de lado derecho con números pares; la línea central se denomina con la letra Z, que se deriva del símbolo anglosajón de cero. [Mayor et al., 2013]

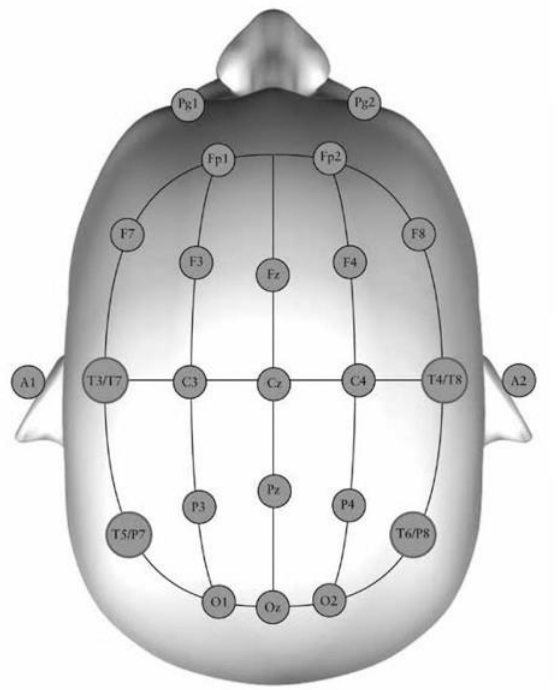

Figura 2.6: Sistema Internacional de Medicion 10-20

Una de las técnicas electroencefalograficas pero de mejor resolucion espacial es la Magnetoencefalografía (MEG), que recoge los campos magnéticos generados por las corrientes eléctricas cerebrales, gracias a la poca distorsion de estos campos magnéticos, son capaces de obtener una resolución espacial buena. Uno de los principales inconvenientes de esta técnica, son los elevados costes de adquisicion y mantenimiento. [Vega, 2012]

Adquisición de señales mediante Electroencefalografía

La maquina de electroencefalografía se dividen en equipos analogos y equipos digitales, los primeros utilizan agujas para el registro sobre el papel, mientras que los equipos digitales se valen de un software que convierte la señal eléctrica en digital para ser observada en pantalla. El número mínimo para la adquisición de la señal es de ocho electrodos, sin embargo, es posible encontrar equipos con diez, doce o dieciséis canales. [Mayor et al., 2013]

Para la adquisición de la señal, [Mayor et al., 2013] explica que el equipo de electro-

encefalografía se puede dividir en:

- Calibracion: Es el proceso mediante el cual se determina el voltaje de un potencial electroencefalografico, el cual es comparado con una actividad eléctrica de voltaje conocido.
- Amplificación: Cada canal del electroencefalograma tiene su propio amplificador, que se encarga de hacer un diferencial entre dos potenciales eléctricos conocidos como input, aumentan el voltaje de la señal obtenida, la transforman en micro-voltios a voltios y permite que se registre en el electroencefalógrafo.
- Impedancia: Es la resistencia al paso de la corriente alterna en una zona del circuito.
- Filtrado: Dentro de la actividad eléctrica cerebral, existe un rango de frecuencias, que es útil en la interpretacion del estudio; este rango se encuentra de 1 a 30 Hz.

Interfaces BCI

En su libro [Fazel-rezai, 2011], describe a la Interfaz Cerebro-Computadora (BCI, por sus siglas en inglés) como un sistema que se compone de una señal entrada, el procesamiento de la señal mediante un algoritmo, el cual mapea la señal de entrada para dar una señal de salida. Esta interfaz puede hacer uso de diferentes señales provenientes del cuerpo, de manera que puedan ser procesadas para el control de dispositivos externos, para ello, el paciente es conectado a la BCI por medio de electrodos o sensores a una unidad de amplificación de Bioseñales y a una unidad de adquisicion de señales que convierten la señal analoga a una señal digital, como se muestra en la figura 2.6.

Una vez que los datos son adquiridos, son enviados a un sistema de procesamiento y clasificacion en tiempo real, que muestra al paciente una retroalimentacion o

estímulo, que permite que el paciente aprenda a controlar la BCI. El concepto de Interfaz Cerebro-Computadora (BCI) fue introducido por el Dr. Jacques J. Vidal, en su articulo ”Toward Direct Brain-Computer Communication”, en el cual el Dr. Vidal se plantea la siguiente pregunta, ”¿Se podría trabaja la comunicación humano-computadora para trasportar informacion con el proposito de controlar dispositivos externos, pro tesis o inclusive naves espaciales”. [Vidal, 1973]

En virtud del gran aporte científico del Dr. Vidal, surgieron diversas lineas de investigacion dedicadas a crear aplicaciones, especialmente para los pacientes con paralisis, ya que el hecho de brindarle una oportunidad de comunicación a estos pacientes se ha convertido en una prioridad. [Schreuder, 2014]

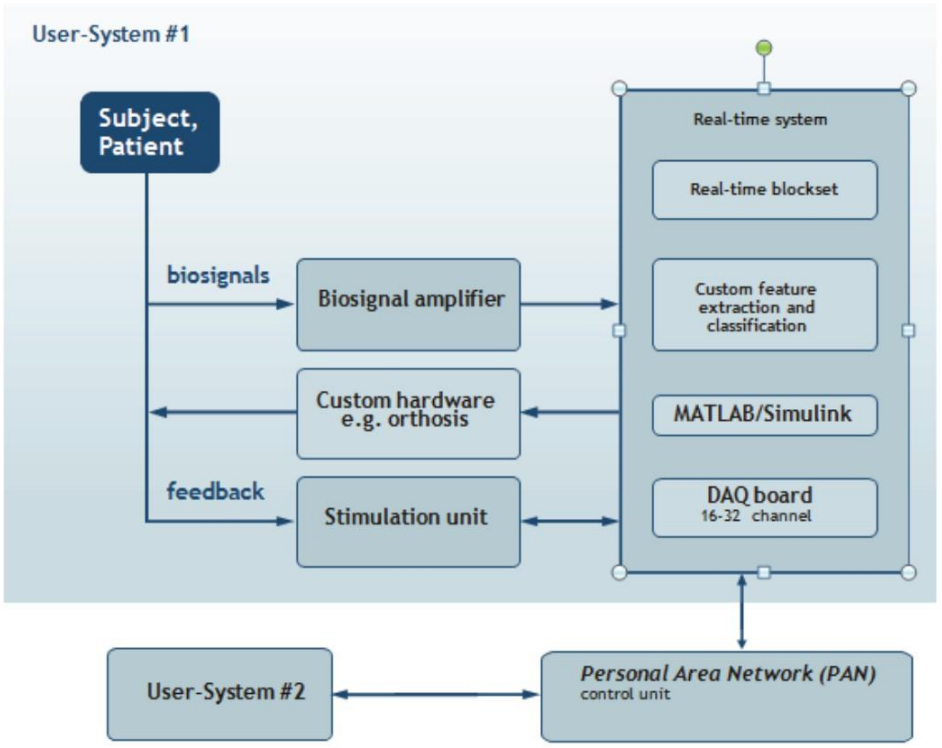

Figura 2.7: Componentes de un sistema BCI [Fazel-rezai, 2011]

Diseño de experimentos para el estudio del cerebro humano

En el campo de la investigación Biomédica, Neurociencia, Médica, entre otras; la técnica mas común usada para el estudio de las funciones cerebrales es la Electroencefalografía (EEG); ya que es un medio no invasivo para obtener datos que le permiten al investigador conocer las distintas fusiones cerebrales como el percepcion, la imaginacion motora, etc; así como detectar anomalías como los desordenes del sueño, epilepsia, depresión entre otras. Esta actividad cerebral se representa por medio de una señal de micro-voltaje que es capturada por el equipo de electroencefalografía; que con los avances de la tecnología dichos datos pueden ser almacenados de manera digital, lo que permite realizar el procesamiento de la señal EEG por medio de modelos computacionales.

Partiendo de esta técnica de recoleccion y medida de la actividad cerebral, los investigadores podrán plantear el rumbo de su investigación; al formularse una pregunta o al observar un área de oportunidad, que los conducirá al planteamiento de un hipotesis con la cual buscan dar una solución al problema planteado. Para llevar acabo un buen estudio de investigacion, es necesario que el estudio tenga un buen diseño experimental, el cual deber de tener las siguientes características: [Malik and Amin, 2017]

- Ser lo mas simple posible, para que puede ser fácil de operar por el equipo de investigación.
- Probar una hipótesis específica y proveer estimaciones justas de los resultados y los riesgos que conlleva.
- Generar los costos mínimos al realizar los experimentos, sin dejar de lado la búsqueda de informacion significativa.
- Planificar el analisis de los datos y la interpretacion de resultados.

- Realizar conclusiones que tengan validez.

Implicaciones Éticas

Como se ha narrado a lo largo de la historia, el verdadero avance médico y científico se ha envuelto en serios dilemas; ya que se ha requerido de la experimentacion en humanos y animales, que muchas veces puede ser catalogada como cruel he inhumana. Por lo que esto puede llegar a generar inquietudes éticas, políticas, humanistas y legales . Aunque el electroencefalograma, al ser una técnica no invasiva; no queda exenta de estas preocupaciones como lo describe [Malik and Amin, 2017].

El autor hace hincapié en que el elemento más importante para la investigación, es el consentimiento informado del sujeto, es decir, si los sujetos son competentes para decidir sobre su participacion. Ya que la falta de conocimiento de los sujetos sobre el experimento puede inducirlos a ser incompetentes para el experimento. La finalidad del consentimiento informado es, proporcionar a los participantes informacion completa sobre los experimentos de investigación en la medida de lo posible para que quede claro para los participantes o sus padres/tutores, si se trata de niños, para así decidir su participación. Dicho consentimiento deber ser dado por escrito a los participantes, y ser firmado antes de comenzar el experimento; en caso en que el participante no posea la capacidad de dar el consentimiento o sea menor de edad, se debe obtener la firma del consentimiento por parte del padre o tutor legal.

Los participantes tendrán pleno derecho de decidir no participar en el experimento, a lo cual el investigador deber respetar dicha decision. Por lo que, el personal involucrado en la investigación, deber de asegurar que la decision de rechazo por parte del paciente, no afectara el desarrollo del experimento; así mismo deberá brindar la seguridad al participante de que no tendrá ninguna afectacion al rechazar la participacion.

Tamaño de la muestra

El autor [Malik and Amin, 2017], nos menciona otro aspecto importante en la experimentacion para el estudio de la actividad cerebral, se trata del calculo del tamaño de la muestra. Aunque no resulta una tarea sencilla, la mayoría de las investigaciones científicas recurren a una pregunta en común "¿cuantos individuos deben ser incluidos en el estudio de investigacion?"; por lo que se selecciona un pequeño conjunto de individuos de la poblacion que es de tamaño reducido pero estadísticamente suficiente para representar a la población objetivo. En la práctica, el tamaño de la muestra se determina en función de los gastos de la reunión de datos y debe ser lo suficientemente grande como para tener suficiente poder estadístico.

Se debe considerara tener en claro el objetivo y la hipotesis de la investigacion, antes de realizar el cálculo para determinar del tamaño de la poblacion. El autor puntualiza que, tener los objetivos bien definidos llevaran a los investigadores a extraer información relevante de estudios anteriores para utilizarla en el calculo del tamaño de la muestra, por ejemplo, las diferencias medias, la varianza, la desviación estandar y el tamaño del efecto.

Señales digitales

Una señal, se defina como un fenomeno físico que varia en tiempo, espacio o cualquier otra variable independiente, matematicamente podemos describir una señal como una función de una o mas variables independientes, como se muestra a continuacion. [Proakis and Manolakis, 1996]

$$S_2(t) = 20t^2 \tag{2.1}$$

Donde la ecuación 2.1, es una señal que varia con respecto a la variable del tiempo. La señal digital, codifica sus valores para que estos sean analizados en términos de valores discretos, en lugar de valores continuos. Este tipo de señales, poseen las siguientes características. [Proakis and Manolakis, 1996]

- Pueden ser amplificadas y reconstruidas al mismo tiempo.
- Se puede contar con un sistema de deteccion y correccion de errores.
- De fácil procesamiento.
- Mínima pérdida de calidad.

Tratamiento de señales digitales

El tratamiento de señales digitales es un área de la ciencia y la ingeniería, que se ha desarrollado por mas de 30 años. Como resultado de este rapido desarrollo, se han logrado los avances tan significativos en el tecnología computacional y la fabricación de circuitos integrados, que han permitido la construccion de sistemas digitales altamente sofisticados, capaces de ejecutar tareas complicadas. [Proakis and Manolakis, 1996]

El procesamiento de señales digitales, es un método alternativo del procesamiento analógico. Para ello, se necesita una interfaz entre la señal analoga y el procesador digital, dicha interfaz se denomina convertidor análogo-digital (A/D), que da como resultado de salida una señal digital, que pude ser introducida a un procesador digital. Este procesador digital, puede ser una computadora digital programable o un pequeño microprocesador o cualquier otro hardware programable. [Proakis and Manolakis, 1996]

El proceso de conversión de la señal análogo a digital, es descrita por [Usategui et al., 2007], donde comienza por un circuito operacional en como comparación con la señal analogica de entrada, el cual amplifica su ganancia para ofrecer una alta impedancia, con lo que se garantiza que el circuito que provee la señal no la deforma. La señal resultante es de muy baja impedancia, por lo que es enviada a un condensador hasta que disponga de una tension igual a la aplicada al momento de la entrada operacional, cuando el condensador haya tomado una muestra de la tension de entrada, la señal es enviada a otro amplificador operacional.

Filtro de Respuesta finita al impulso (FIR)

Uno de los tipos de filtros digitales con una amplia aplicación en distintas problemáticas son los Filtros de Respuesta Infinita (FIR, por sus siglas en ingles). Dichos filtros pueden usarse en la reconstruccion de la señal sin distorsión; ya que son inherentemente estables. Para entender mas sobre su funcionamiento, observemos la siguiente ecuacion. [Elliott, 1988]

$$H(z) = h(0) + h(1)z^{-1} + \cdots + h(N-1)z^{-(N-1)}$$

Figura 2.8: Ecuacion del Filtro FIR.

Donde $N-1$ es el orden del filtro y h(n) es el coeficiente de la respuesta al impulso. Como el filtro tiene un respuesta linear de la fase, este puede ser simétrico o antisimétrico; esto dependiendo si el orden del filtro dado es par o impar. La funcion de transferencia de este filtro puede ser implementado con N multiplicadores y $N-1$ de orden: a esto se le conoce como forma directa; como se observa en la figura 2.9.

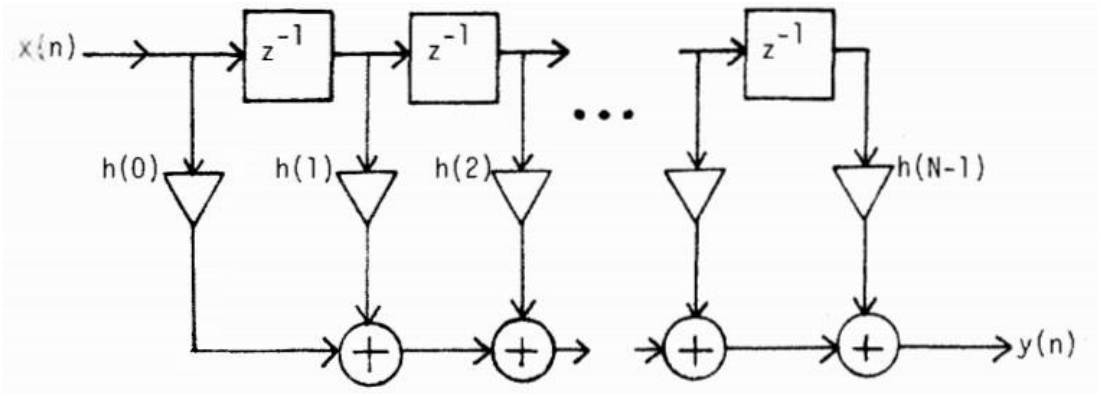

Figura 2.9: Implementación de la forma directa del filtro FIR.

Teorema de Nyquist

El teorema de Nyquist se plantea de la siguiente manera. Sea una señal x(t) ∈ L_2(R) con CTFTX(ω). Si X(ω) = 0 para todo |ω| > Π/T , entonces x(t) podrá ser reconstruida a partir de sus muestras x(nT) usando la siguiente formula de reconstruccion. [Eldar, 2015]

$$x(t) = \sum_{n \in \mathbb{Z}} x(nT)\,\mathrm{sinc}((t - nT)/T)$$

where

$$\mathrm{sinc}(t) = \frac{\sin(\pi t)}{\pi t}.$$

Figura 2.10: Formula de reconstrucción del Teorema de Nyquist.[Eldar, 2015]

El teorema de Nyquist afirma que una señal Π/T puede recuperarse a partir de sus muestras uniformemente espaciadas con el período T, o en la frecuencia de muestreo f = 1/T , la cual hace referencia a la frecuencia de muestreo de Nyquist. La recuperacion es posible a partir de muestras uniformes en cualquier caso igual o superior a la tasa de Nyquist, es decir, la señal puede obtenerse utilizando la ecuación de reconstrucción (figura 2.10) con T sustituyendo a T para cualquier T ≤ T.

Transformada Wavelet Discreta

La transformada Wavellet Discreta (DWT,por sus siglas en inglés), se puede entender bajo el enfoque de la descomposición multiresolución de las ondas de las señales o imagenes; el cual consiste en crear un componente de aproximación utilizando una funcion de escalado (un filtro pasa-bajas) y componentes de detalle utilizando funciones ondulatorias (filtros pasa-altos). La principal ventaja que nos ofrece esta descomposicion multiresolucion es, que podemos brindar una alta resolucion a los objetos pequeños, mientras que los objetos grandes pueden tomar una baja resolucion [Sundararajan, 2016].

El autor [Sundararajan, 2016] nos indica que, la DWT es similar a la DFT del análisis de Fourier y se utiliza ampliamente en aplicaciones practicas. El principio sigue siendo el mismo que el de las otras transformaciones. La señal se transforma en una forma diferente utilizando funciones básicas mas adecuadas para el procesamiento requerido, y la transformacion puede llevarse a cabo de manera eficiente utilizando algoritmos rápidos. Una característica distintiva de las funciones de la base de transformacion wavelet es que todas ellas se derivan de dos funciones transitorias, una funcion de escalado y una funcion de la wavelet, por desplazamiento temporal y escalado. La funcion de onda en sí misma se define como una combinacion lineal de funciones de escalado y de escalado desplazado. El análisis de la DWT se basa en el hecho de que la combinación de un grupo continuo de componentes de frecuencia del espectro produce una señal transitoria en el dominio del tiempo. Las señales de base de la DWT estan relacionadas con este tipo de señales y se derivan de esa relacion. Por lo tanto, la DWT es local en la medida de lo posible tanto en tiempo como en frecuencia. Es decir, se puede determinar el intervalo de tiempo en que se produce un componente compuesto por un grupo continuo de frecuencias. Sin embargo, no se puede encontrar con precision la instancia temporal de la ubicación de un solo componente de frecuencia.

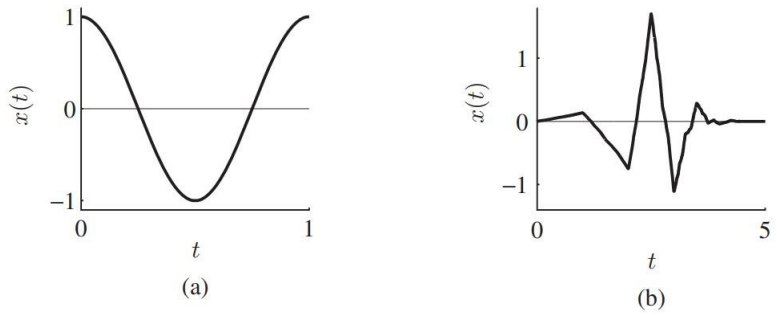

Figura 2.11: a) Transformada de Fourier, b) Transformada Wavelet [Sundararajan, 2016]

El analisis DWT es una herramienta para convertir una señal del dominio del tiempo en una señal del dominio de la frecuencia y viceversa. Como las señales de base son de naturaleza transitoria y corresponden a una parte del espectro, podemos localizar la ocurrencia de un evento en el dominio del tiempo. Es decir, la correlacion entre una señal dada y una señal de base puede determinarse en varios instantes de tiempo desplazándola. La segunda observacion es que el espectro de la señal se divide en partes desiguales. Esto se adapta al análisis de la señal, ya que un mayor ancho de banda e intervalos de tiempo mas cortos son apropiados para detectar un componente de alta frecuencia y viceversa [Sundararajan, 2016].

Ciencia de datos

Ciencia de datos o mejor conocida como Data Science, es descrita por el autor [Saltz and Stanton, 2017] como un termino que nos evoca el pensar en personas de bata blanca dentro de un laboratorio; pero dicho pensamiento no puede estar mas alejado de la realidad. La ciencia de los datos es mucho mas que el simple analisis de los datos; ya que nos ofrece una gama de funciones y requiere una serie de habilidades. Por lo cual el autor nos ilustra con el siguiente ejemplo; consideremos esta idea pensando en algunos de los datos involucrados en la compra de una caja de cereales.

Cualquiera que sea su preferencia en cuanto a los cereales -frutal, chocolatada, fibrosa o con nueces- usted se prepara para la compra escribiendo cereal en su lista de compras. Ya su compra planeada es un dato, también llamado dato, aunque un garabato de lapiz en la parte de atrás en un sobre que solo usted puede leer. Cuando llegas a la tienda de comestibles, usas tu dato como recordatorio para coger esa caja gigante de FruityChocoBoms del estante y ponerla en tu carrito. En la caja, el cajero escanea el codigo de barras de su caja, y la caja registrará el precio. De vuelta en el almacén, una computadora le dice al gerente de la tienda que es hora de pedir otro pedido al distribuidor, porque su compra fue una de las últimas cajas de la tienda. También tienes un cupón para tu caja grande, y el cajero lo escanea, dandote un descuento predeterminado.

Al final de la semana, un informe de todos los cupones escaneados del fabricante se sube a la compañía de cereales para que puedan emitir un reembolso a la tienda de comestibles por todos los descuentos de los cupones que han repartido a los clientes. Finalmente, a finales de mes, el gerente de la tienda mira una colorida coleccion de graficos de pastel que muestran todos los diferentes tipos de cereal que se vendieron y basandose en las fuertes ventas de cereales frutales, decide ofrecer mas variedades de estos en el limitado espacio de la tienda el próximo mes.

Así que la pequeña informacion que comenzó como un garabato en su lista de la compra terminó en muchos lugares diferentes, sobre todo en el escritorio de un gerente como una ayuda para la toma de decisiones. En el viaje de tu lapiz al escritorio del gerente, el dato pasó por muchas transformaciones. Además de las computadoras en las que el dato pudo haber pasado o haber permanecido encendido a largo plazo, muchas otras piezas de hardware, como el escaner de codigo de barras, participaron en la recoleccion, manipulacion, transmision y almacenamiento del dato.

Ademas, se utilizaron muchos programas informaticos diferentes para organizar, agregar, visualizar y presentar los datos. Finalmente, muchos sistemas humanos diferentes estuvieron involucrados en el trabajo con el dato. La gente decidía qué sistemas comprar e instalar, quién debía tener acceso a qué tipo de datos y qué pasaría con los datos una vez cumplido su propósito inmediato. El personal de la cadena de supermercados y sus socios tomaron otras mil decisiones y negociaciones detalladas antes de que el escenario descrito anteriormente pudiera hacerse realidad [Saltz and Stanton, 2017].

Aprendizaje Activo (Active Learning)

Aprendizaje activo (AL por sus siglas en inglés) o también conocido como ”diseño experimental óptimo”, es un subcampo del Machine Learning (Aprendizaje Maquina); cuya hipotesis indica que si a un algoritmo se le permite escoger los datos de los cuales aprende, esto mejorar el proceso de aprendizaje. Es decir, el algoritmo puede

consultar activamente a un anotador humano (en inglés se conoce como oracle cuyo rol es cubierto por el programador) cuales son las consultas que pueden ser etiquetadas y cuales no pueden ser etiquetadas. Es importante mencionar que según la teoría del AL, el rol del anotador también puede ser cubierto por un conjunto de datos previamente etiquetados por algún procesos de agrupacion o clusterizacion como por ejemplo el algoritmo K-nn; por consiguiente y como se menciona con anterioridad, el algoritmo de Aprendizaje activo (AL) puede realizar la consulta a un ente humano si un dato pertenece a una determinada clase o bien, valerse un conjunto pequeño de datos previamente etiquetados con el cual el algoritmo determina por si solo la pertenecía del dato a una determinada clase. Actualmente el Aprendizaje Activo, es usado con frecuencia en diversos problemas de Machine Learning, donde el común denominador son dos escenarios, el primer escenario se caracteriza por la parecencia de los datos proporcionados son abundantes y el proceso de etiquetado es nulo o muy costoso de obtener. El segundo escenario se caracteriza por la parecencia de datos poco abundantes, a los cuales se les plantea sacar el mejor provecho. [Settles, 2010]

Esta técnica de aprendizaje semi-supervisado tiene un ciclo de funcionamiento, que comienza con un pequeño grupo de instancias del conjunto de entrenamiento ya etiquetado; posteriormente solicita la etiqueta de una o mas instancias seleccionadas cuidadosamente, aprende del resultado de esta consulta, entonces aprovecha este nuevo conocimiento para elegir a que instancia consultar después; como se describe en la siguiente figura 2.12.

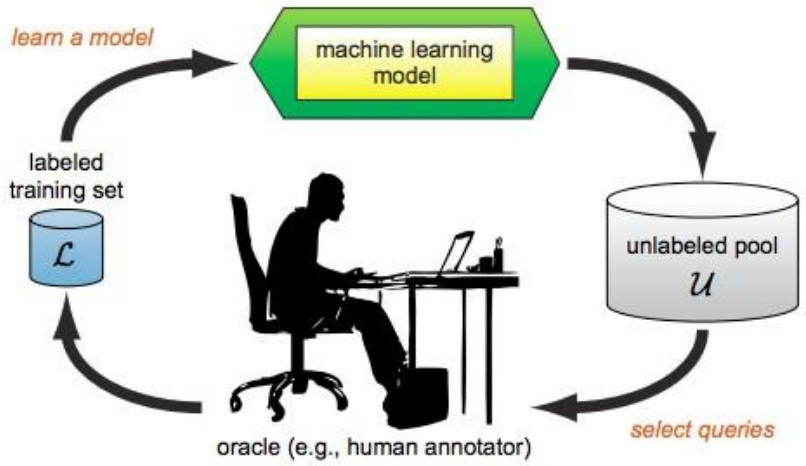

Figura 2.12: Ciclo del Aprendizaje[Settles, 2010]

En el aprendizaje activo, existen tres escenarios de uso común, los cuales asumen que las consultas se toman de instancias sin etiquetar, para ser etiquetadas por un anotador. Estos escenarios son: 1) síntesis de consultas de pertenencia, 2) muestra selectiva basada en flujo, 3) muestreo en grupo. Como se muestra en la siguiente figura 2.13.

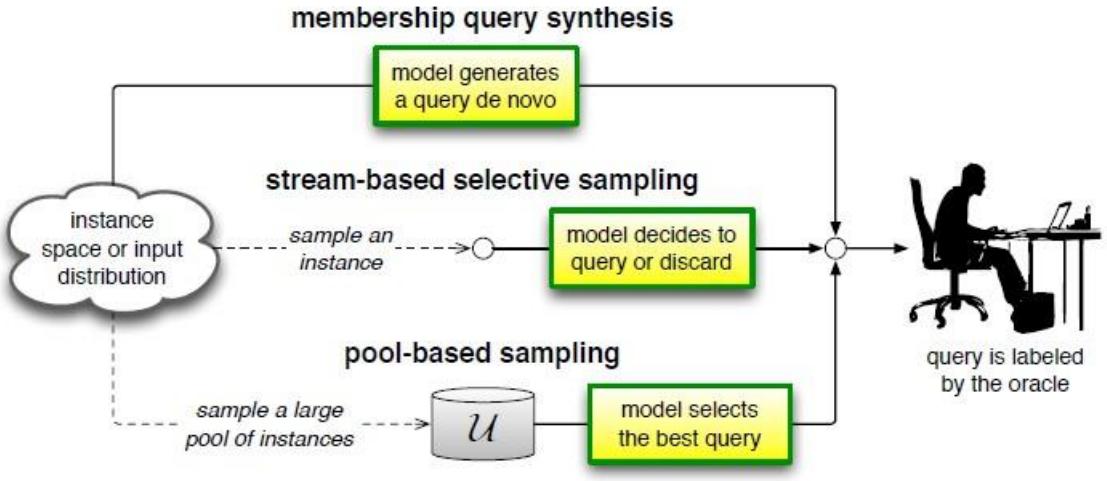

Figura 2.13: Escenarios del Aprendizaje Activo [Settles, 2010]

Síntesis de consultas de pertenencia: El algoritmo, deberá solicitar las etiquetas de cada una de las instancias sin etiquetar dentro del espacio de entrada, incluyendo aquellas consultas que el algoritmo haya generado desde cero, en lugar de solicitar alguno de los datos ya muestreados de alguna otra distribucion. Este escenario resulta útil en set de datos pequeños, ya que el etiquetar gran cantidad de datos por medio de un anotador podría resultar bastante complicado. [Settles, 2010]

Muestra selectiva basada en flujo: Este escenario asume que obtener una una instancia sin etiquetar no genera ningún costo computacional, por lo que puede ser tomada de la distribucion actual, y el algoritmo podr decidir si se solicita o no

que la instancia sea etiquetada. Por lo que cada punto de los datos sin etiquetar es examinado uno a uno, evaluando la informacion de cada elemento con respecto a los parametros de consulta. [Settles, 2010]

Muestreo en grupo: Se asume que existe un pequeño set de datos ya etiquetados y un gran grupo de datos sin etiquetar. Comúnmente, la consulta de la instancia a etiquetar se lleva acabo de una manera selectiva, ya que se le asigna una puntacion informativa que es usada para evaluar todas las instancias en el set de datos. Enseguida, el algoritmo selecciona aquella instancia que proporcione una puntuacion

mayor y consulta al anotador para proceder a etiquetar la instancia. [Settles, 2010]

Teoría del limite central

Como nos describe el autor [Alvarado and Batanero, 2008], el Teorema de Límite Central, es uno de los mas usados al momento de analizar el comportamiento de variables aleatorias. Dicho teorema, se ha desarrollado a lo largo de la historia, generando un gran impacto en la inferencia estadística ya que diferentes parametros probabilísticos, pueden ser expresados en una suma de variables.

Dicho teorema puede expresarse como "Para muestras grandes, se puede obtener una aproximacion cercana de la distribucion muestral de la media con una distribución normal"[Freund and Simon, 1994]. Es decir, este teorema se puede aplicar tanto a poblaciones infinitas como a poblaciones finitas; cuando n, a pesar de ser grande no es mas que una pequeña parte de la población; aunque resulte difícil señalar cuan grande es el tamaño de n, por lo general se considera que $n = 30$ es un tamaño lo suficientemente alto. Para comprender un poco mas este teorema, es necesario conocer los siguientes principios que nos describe el autor [Triola et al., 2012]:

1. Si $n > 30$, entonces las medias muestrales tienen una distribucion que se puede aproximar por medio de una distribucion normal, con una media m y una desviacion estandar (éste es el lineamiento que suele utilizarse, independientemente de la distribucion de la población original).

2. Si $n \leq 30$ y la poblacion original tiene una distribucion normal, entonces las medias muestrales tienen una distribucion normal con una media m y una desviacion estandar $\delta/\sqrt{n}$.

3. Si $n \leq 30$, pero la poblacion original no tiene una distribución normal, entonces no se aplican los métodos de esta seccion.

Capítulo 3

Desarrollo

”La lengua no es la envoltura del pensamiento sino el pensamiento mismo.”Miguel de Unamuno.

Recursos Humanos

Este trabajo de investigacion multidisciplinaria, se llevo acabo con la colaboración de diversas instituciones, dentro del ámbito médico y clínico; se conto con la participacion de la Dra. María del Consuelo Martínez Wbaldo, coordinadora de la Unidad de Investigacion Sociomédica del Instituto Nacional de Rehabilitación y la Dra. Socorro Gutiérrez jefa del Departamento de Audiología del Centro Estatal de Rehabilitación del Estado de Guajanuato; quienes aportaron todo su experiencia en el área del lenguaje.

En el área tecnologica, se cont con el apoyo del Dr. Carlos Reyes Garcia del Instituto Nacional de Astrofísica, Óptica y Electronica, el cual aportó su basta experiencia en el tratamiento de señales biologicas, en específico de la señal de Electroencefalograma; así como el su apoyo como coasesor de este trabajo de investigacion. Dentro de esta área tecnologica, se cont con la direccion para este trabajo de investigación de la Dra. Rosario Baltazar Flores, así como la revision de MC. Miguel Ángel Casillas y la MC. Martha Alicia Rocha.

Un espacial agradecimiento a la Lic. Lorena Erendira Velazquez Morales quien participp en este trabajo de investigacion con su experiencia en el área de la docencia,

así como en facilitar a varios de sus alumnos, los cuales formaron parte como sujetos de prueba en el grupo de control; este agradecimiento se hace extensivo a la supervisora de la Zona y a la directora del plantel, por proporcionar el permiso para realizar la recolección de datos dentro del plantel.

Recursos Materiales

Para el desarrollo de este trabajo de investigacion, se tuvo a disposición un equipo portátil de EEG inalambrico via Bluetooth, cuyo nombre es Cyton Biosensing Board de ocho canales de la empresa OpenBCI. Para la recoleccion de la señal EEG se usaron electrodos de copa de oro, así como la pasta conductora de uso comercial Ten-20.

Para el desarrollo del software de análisis y clasificación de la señal EGG se uso el entorno de desarrollo integrado MatLab, en conjunto con la herramienta de tratamiento de señales EEGlab. La totalidad del sistema es soportado por una Laptop con un procesador Intel Core i7 de octava generacion.

Recursos Financieros

El presenta trabajo de investigación, fue financiado por el Consejo Nacional de Ciencia y Tecnología (CONACYT), todos los gastos que implicaron este trabajo de investigacion fueron cubiertos por la beca Programa Nacional de Posgrados de Calidad.

Consideraciones éticas

Este trabajo de investigación esta clasificado como de riesgo mínimo, ya que no presenta ningún tipo de riesgo para el sujeto de prueba; el cual aceptó su participacion al firmar una carta de consentimiento informado (Ver anexos).

Consideraciones de Bioseguridad

El trabajo de investigacion no genera ningún tipo de daño mayor a la salud y bienestar del sujeto de prueba, solo se debe considerar la presencia de ansiedad por el uso del equipo de EEG ya que se trata de cables dispuesto por la cabeza, así como una posible reaccion alérgica a la pasta conductora Ten-20.

Modelo del problema

El correcto desarrollo del lenguaje en el individuo es esencial para el desenvolvimiento pleno de este. Podemos encontrar que los trastornos del desarrollo del lenguaje, se manifiestan con mayor frecuencia en pacientes en edad pediatrica; lo que resulta alarmante ya que de no someter a estos pacientes a tratamientos de correccion o mejora, estos podr perder la capacidad de desarrollar esta habilidad fundamental que

le permitiría expresarse y desenvolverse con su entrono. Es por ello que los expertos médicos y terapistas de lenguaje, han identificado momentos clave, los cuales otorgan la capacidad de adquirir dichas habilidades de manera perdurable. Esto gracias al diagnostico temprano del trastorno en específico que posee cada paciente, lo cual se logra a través de evaluaciones para identificar los aspectos del lenguaje de acuerdo a su edad cronologica. Estas evaluaciones involucran diferentes aspectos emocionales, sociales, conductas, rutinas, ademas de la aplicacion de pruebas psicologicas o estudios médicos; los cuales podría agotar el tiempo de estos momentos clave, reduciendo así la oportunidad de que el paciente adquiera las habilidades necesarias del lenguaje.

Por consiguiente, este trabajo de investigación propone el diseño de una solución computacional que podría brindar una herramienta de apoyo para el pre-diagnostico de estas afectaciones al desarrollo del lenguaje. La solucion que se propone en este trabajo de investigacion parte de aplicar la técnica de Aprendizaje Activo (Active Learning), en conjunto con técnicas de tratamiento de señales digitales aplicadas en señales Electroencefalograficas de niños en edad escolar, que nos permita desarrollar un proceso de etiquetado de datos a traves del Aprendizaje Activo (Active Learning) capas de identificar el ”Proceso Cerebral del Lenguaje”.

Adquisicion de la señal EEG

En los inicios de este trabajo de investigación, se había planteado que la señal EEG fuera obtenida de pacientes como se menciona con anterioridad en este trabajo de investigación, pero debido a la pandemia COVID-19, suceso historico que comenzó desde finales del años 2019, que ha afectado a nivel mundial dejando un sin numero de muertes a su paso. Dicho suceso provoco replantear la manera de trabajo de esta investigación, ya que debido al periodo de cuarentena y a las medidas de prevencion para evitar contagios, se descartado totalmente el trabajo de adquisicion con pacientes; por lo cual se inici la búsqueda de un set de datos ya elaborado de algún estudio de investigacion previo, el cual encajara con los objetivos propuestos dentro de esta investigación, el cual permitiera identificar características de la señal EEG durante el proceso cerebral del lenguaje, haciendo énfasis en la comprension auditiva.

Es por ello que, dentro de esta búsqueda se ha tomada lo decision conjunta de trabajar con el set de datos Child Mind Institute - Multimodal Resource for Studying Information Processing in the Developing Brain (MIPDB) [Simon P. Kelly, 2016]. El estudio de Simmon P. Kelly, pretende cubrir un amplio espectro de enfoques que se examinan en la neurociencia cognitiva moderna; con la finalidad de proveer datos para el incremento de nuevas investigaciones en el campo del desarrollo cerebral así como en la deteccion de procesos patologicos, tal como se investigo en esta tesis.

A continuacion, detallaremos como fue le proceso llevado a cabo por el equipo de investigación a cargo del PhD Simmon P. Kelly para la adquisicion de de las señales EEG usadas para este trabajo de investigacion.

Set de datos Child Mind Institute - Multimodal Resource for Studying Information Processing in the Developing Brain (MIPDB)

Dicho set de datos incluye dos paradigmas experimentales; el primero de ellos realiza un serie de tareas activas que analizan los principales componentes del desarrollo cerebral. El segundo paradigma experimental, contiene una serie de tareas pasivas, que permiten examinar el funcionamiento de la red neuronal durante diversas estimulaciones externas. Este conjunto de datos fue elegido por decision conjunta del comité, ya que se analizó con detalle, los paradigmas correspondientes a las tareas activas y pasivas que son incluidas dentro de este estudio podrían ser útiles para identificar señales del procesamiento cerebral del lenguaje, en específico, aquellas relacionadas a la comprension auditiva. Y que para los fines de esta investigacion nos enfocaremos en los paradigmas Sequence Learning y Naturalistic Viewing como se muestra en la siguiente figura 3.1.

Task	Depth of processing/degree of stimulation	Description
Active (Task-Dependent) Paradigms		
Contrast change	Minimal	Probes basic elements of sensorimotor translations, e.g., sensory evidence encoding, decision formation and motor preparation, providing dynamic measurements of each processing stage in isolation.
Sequence learning	Moderate	Assesses successive visuo-spatial sequence learning by using semantically unloaded stimuli, tracks the progress of gradual memory formation
Symbol search	Complex	A computerized version of a clinical pediatric assessment measuring processing speed capacity in a visual search task, which involves multiple perceptual decisions, short-term memory and motor response.
Passive (Task-Independent) Paradigms		
Resting-state	None	Measures endogenous brain activity during rest.
Surround suppression	Minimal	Measures excitatory (using the steady-state visually evoked potential; SSVEP) and inhibitory (using the surround-suppression effect) neurophysiological activity during sensory processing with semantically unloaded stimuli.
Naturalistic viewing	Complex	Measures neurophysiological activity during higher-level audio-visual stimulation (movies).

Figura 3.1: Paradigmas experimentales.[Simon P. Kelly, 2016]

Para evaluar cada uno de estos paradigmas, el equipo de investigacion a cargo del PhD Simmon P. Kelly uso un modelo de tres etapa que consiste en:

1. Percepción sensorial
2. Integración de informacion (es decir, acumulación, integracion y elaboracion de informacion)
3. Generación de respuesta (es decir, preparación motora, ejecucion).

El protocolo de adquisición que se sigui dentro del estudio llevado a cabo por el PhD Simmon P. Kelly para la recoleccion de la señal EEG, comenzó con el proceso de seleccion; los participantes potenciales (o sus padres, si eran menores de 18 años) fueron examinados por teléfono, o en persona en el Child Mind Institute, por un asistente de investigación capacitado. La entrevista de 10 minutos evaluó la elegibilidad y seguridad de los participantes para participar, obteniendo informacion sobre:

- Antecedentes de enfermedades psiquiatricas, incluyendo tratamientos pasados y presentes, medicamentos y diagnosticos.
- Antecedentes de trastornos neurológicos y/o epilepsia.

Descripción del set de datos

Dentro de este estudio, se incluyeron datos de un total de 126 pacientes de entre 6 a 44 años de edad; los cuales son pacientes recurrentes del Child Mind Institute de acuerdo al [Simon P. Kelly, 2016].

De estos 126 pacientes incluidos en el trabajo de investigación del PhD Simmon P. Kelly el 54.8 % son masculinos y el 45.2 % son femeninos; como se muestra en la figura 3.2.

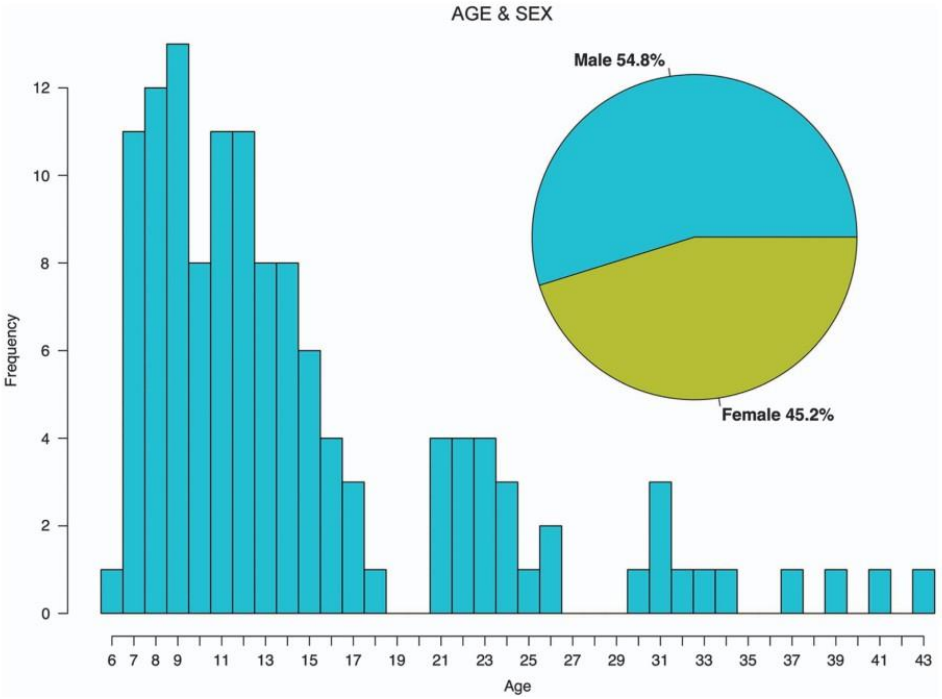

Figura 3.2: Distribución por edad de la poblacion.[Simon P. Kelly, 2016]

De estos participantes seleccionados en el estudio, el 80.2 % han tenido un desarrollo normal sin ningún diagnóstico; mientras que el 19.8 % ha sido diagnosticado con uno o más desordenes clínicos; en la figura 3.3 podemos observar con mas detalle la distribución de los diagnosticos, que se han dividido en once categorías; donde se muestra la frecuencia de cada uno de los diagnósticos; así como los porcentajes correspondientes a la muestra.

Diagnostic Category	Frequency	% of clinical sample	% of total sample
No Diagnosis	101	NA	0.80
Attention	12	0.48	0.10
Anxiety	10	0.40	0.08
Learning	7	0.28	0.06
OCD	4	0.16	0.03
ASD	2	0.08	0.02
Depressive	2	0.08	0.02
Trauma	2	0.08	0.02
Disruptive	2	0.08	0.02
Motor	2	0.08	0.02
Language	1	0.04	0.01
Mood	1	0.04	0.01

Figura 3.3: Distribución de los diagnosticos de los participantes.[Simon P. Kelly, 2016]

Para llevar a cabo la recolección de la señal EEG en el trabajo de investigacion del PhD Simmon P. Kelly, se uso un electroencefalógrafo EEG Geodesic Hydrocel system de 128 canales, al cual se le aplico una frecuencia de muestreo de 500 Hz con un filtro pasa banda de 0.1 a 100 Hz. Se probo la impedancia de cada electrodo para asegurar una correcta adquisición de la señal, la cual se mantuvo por debajo de los 40 kOhm; el electrodo usado como referencia para la adquisicion de la señal fue el electrodo Cz (vertex), de acuerdo al estandar 10-20 de posicionamiento de electrodos.

El estudio se realizo en una habitacion obscura y aislada del sonido, se coloco al participante a una distancia de 70 cm de un monitor de 17 pulgadas. Para asegurar que la cabeza tuviera una posición estable, se uso un soporte para descansar la barbilla. La presentación de los estímulos fue programada en MATLAB usando el PsychToolbox; donde se sigui el mismo orden de presentación para todos los participantes (ver figura 3.4), las instrucciones de cada una de las tareas se presentaban en el monitor y uno de los investigadores realizaba una serie de preguntas a través de un intercomunicador. Si los participantes eran menores de 12 años, un segundo investigador los acompañaba dentro de la habitacion para asegurar el procedimiento

y acompañar al participante. El estudio tuvo una duración de 5 horas, donde los participantes podían elegir hacerlo en varias sesiones o bien, completarlo en una sola sesion con algunos descansos; durante este tiempo se les ofrecía a los participantes bocadillos y jugo, así como alentar a que descansaran entre cada una de las pruebas.

Time	Child	Parent	Adult
0:00	Assent	Consent	Consent
0:30	Cap measurement, Digit span	Demographics, Hollingshead	Demographics, Hollingshead, Cap measurement, Digit span
0:45	EEG Paradigms	Parent Questionnaires (SWAN, CBCL, KINDL, EEG History and Demographics Questionnaire, BRIEF)	EEG Paradigms
2:45	Break		Break
3:00	Questionnaires (IAT, CTAS, KINDL)		Questionnaires (ASR, CAARS, History and Demographics Questionnaire, KINDL, CTAS, IAT)
3:30	Cognitive Testing (WASI, WIAT)		Cognitive Testing (WASI, WIAT)
5:00	Payment	Payment	Payment

Figura 3.4: Agenda del estudio.[Simon P. Kelly, 2016]

Como podemos observar en la figura 3.4, que es tomada del trabajo del autor [Simon P. Kelly, 2016], la cual nos detalla la agenda de actividades efectuadas durante el estudio. Dichas actividades se dividen durante el periodo del estudio, esta agenda se divide en los dos grupos de pacientes principales, por un lado tenemos al grupo de niños (Child) en donde los padres (Parent) apoyaron al grupo de investigadores durante los primeros 45 minutos del estudio para realizar actividades como la firma del consentimiento y el llenado de cuestionarios y al final del estudio para recibir la compensacion economica. Se les pregunt a los padres si querían estar presentes durante las pruebas, o bien si los niños eran menores de 12 años.

El segundo grupo del estudio, corresponde al grupo de adultos (Adult), realizaban cada fase del estudio por si solos a menos que requirieran de algún apoyo especial.

Preparacion de los datos

El acceso al set de datos obtenido del estudio del PhD Simmon P. Kelly es de manera gratuita y abierta a todo el público, por lo que se puede acceder a ellos desde la pagina oficial del Child Mind Institute [Simon P. Kelly, 2016]; los datos de ese estudio están divididos en seis grupos correspondientes a los diferentes rangos de edad; estos grupos están compuestos por pacientes masculinos y femeninos, que son identificados con los colores azul y rojo respectivamente, como se muestra en la figura 3.5.

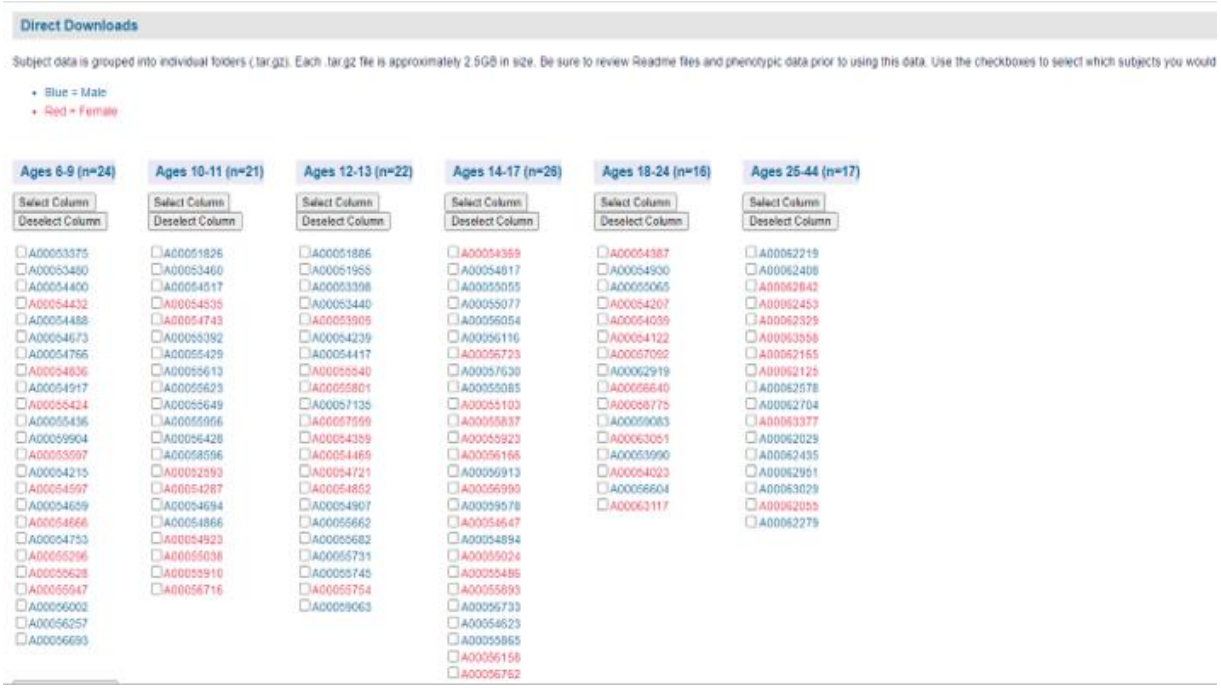

Figura 3.5: Distribución de los datos.[Simon P. Kelly, 2016]

Es preciso destacar, que cada paciente posee su carpeta individual que contiene los archivos recopilados durante el estudio, entre los que se encuentras tres subcarpetas con los archivos de la señal EEG, el rastreo ocular y las pruebas conductuales; cada una de las carpetas por paciente tiene un peso promedio de 2 Gb. Los archivos de interés para esta tesis investigación se encuentran dentro de la carpeta EEG, donde podemos observar dos subcarpetas una con los archivos en crudo del set de datos (Raw) y otra con los archivos con un leve preprocesamiento (preprocessed)

ya que como expluca el estudio del PhD Simmon P. Kelly se han omitido 17 de los 128 electrodos (canales) que no representan mayor relevancia en los datos, como se muestra en la figura 3.6.

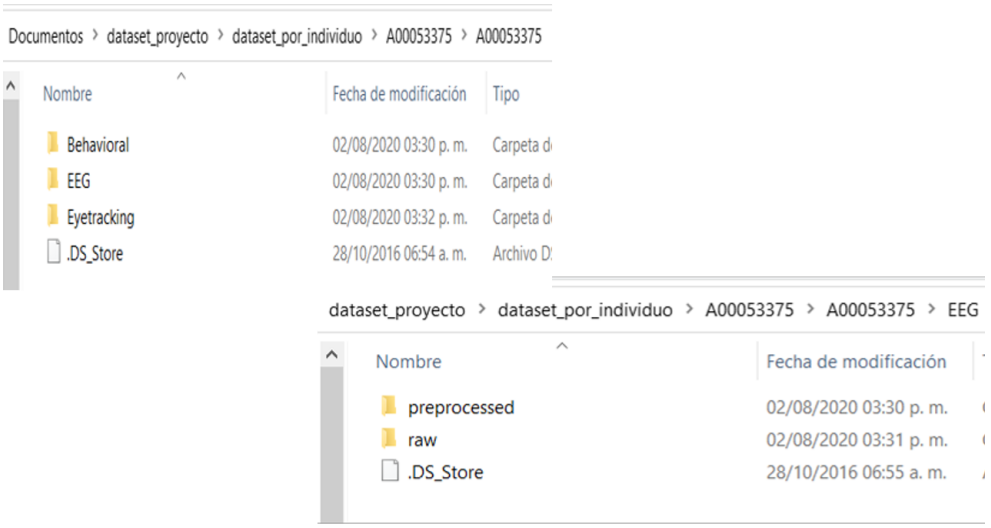

Figura 3.6: Contenido de la carpeta por paciente

Como primera etapa, se trabajar con los archivos .RAW de un total de 6 pacientes de los 126 incluidos el estudio del PhD Simmon P. Kelly; los cuales manipularemos desde el entorno de desarrollo integrado (IDE) de MATLAB; aplicando la librería para el procesamiento de señales electroencefalograficas llamada EEGLAB. Dentro de este IDE, tomaremos los datos de la señal EEG, los cuales estan contenidos dentro del archivo RAW, que contienen la informacion de ese segmento de la señal EEG, que para los propositos de esta investigacion, nos es de utilidad la frecuencia de muestro, la localizacion de los electrodos y los datos de la señal EEG. Otro punto que es preciso destacar, es que cada una de las carpetas que existen por cada uno de los pacientes, no poseen la misma cantidad de informacion, ya que la cantidad de archivos .RAW varia entre cada uno de ellos; pero se tiene un promedio 1.8 Gb de información por cada una de las carpetas individuales por paciente.

Preprocesamiento de los datos

Para esta etapa del proceso de investigación, se trabaj con el ToolBox del IDE de MATLAB llamado EEGLAB; dicha herramienta nos permitir procesar la señal EEG en una primera etapa, con la finalidad de limpiar los datos para solo conservar toda aquella información relevante que sea de utilidad, esta etapa a la que llamaremos Preprocesamiento y que consistir en los siguientes puntos:

- Importar los datos del archivo .RAW.
- Importar la informacion de la localizacion de los canales (electrodos.)
- Reducir la Frecuencia de Muestro (Downsamplig) de 500 Hz a 250 Hz.
- Remover el ruido aleatorio (Baseline).
- Filtrar los datos.
- Eliminar artefactos (ICA).

Habitualmente, esta serie de pasos es la mas usada en el tratamiento de la señal EEG; por lo que se usara en cada uno de los archivos .RAW de cada uno de los 6 pacientes, la finalidad de hacer esto de manera individual, es para asegurar la calidad en esta etapa de preprocesamiento para así trabajar con datos lo más contundentes posibles. A continuacion, se detallar cada uno de estos pasos del preprocesamiento de la señal EEG.

Importar los datos del archivo .RAW.

Como se mencionó anteriormente, cada paciente posee su carpeta individual con los archivos del estudio (Ver figura 3.6). Para este trabajo de investigacion, se trabajo con los archivos .RAW (Ver figura 3.7); dicha extinción de archivos, almacenan los datos sin procesar de la señal EEG, es decir, se almacena toda la señal EEG con la mayor calidad e informacion posible de cada uno de los pacientes; lo que permite cambiar o manipular la informacion de acuerdo a los fines de la investigación.

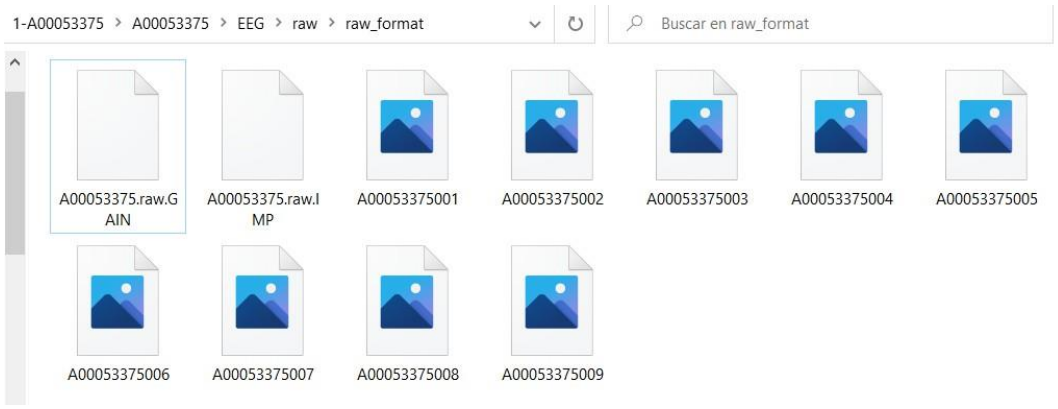

Figura 3.7: Archivos RAW del paciente Uno.

Dentro de la interfaz grafica del ToolBox EEGLAB, se pueden importan una gran variedad de extinciones de archivos de uso mas común relacionados con el análisis de las señales EEG, lo que permite trabajar con la extensión de archivo en crudo (RAW). Para ello se usara el complemento (plugin) FILE-IO interface, que al ser una interfaz gráfica nos otorga la facilidad de buscar el archivo; para posteriormente ser leído por el ToolBox EEGLAB, como se muestra en la figura 3.8.

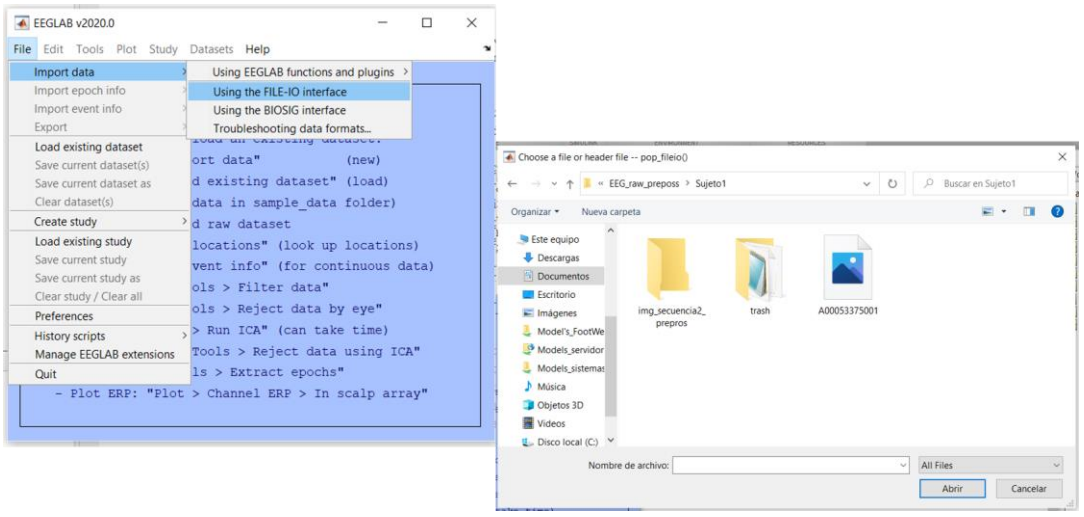

Figura 3.8: Importando el archivo en crudo (Raw)

Importar la información de la localización de los canales o electrodos

Una vez que se han importado los datos del archivo en crudo, importamos los datos referentes a la ubicacion de cada uno de los electrodos o canales; dichos datos son coordenadas en un plano tridimensional. Esta información es importante, ya que identifica a cada uno de los electrodos y de esta manera resulta mas sencillo aplicar el preprocesamiento de la señal e identificar que electrodos fueron afectados o removidos. El grupo de investigacion, encargado de elaborar este set de datos como se mencion en la seccion 'Adquisicion de señales', también nos proveé del archivo con los datos de la ubicacion de los electrodos, por lo que solo queda importar el archivo usando la interface del ToolBox EEGLAB, como se muestra en la figura 3.9.

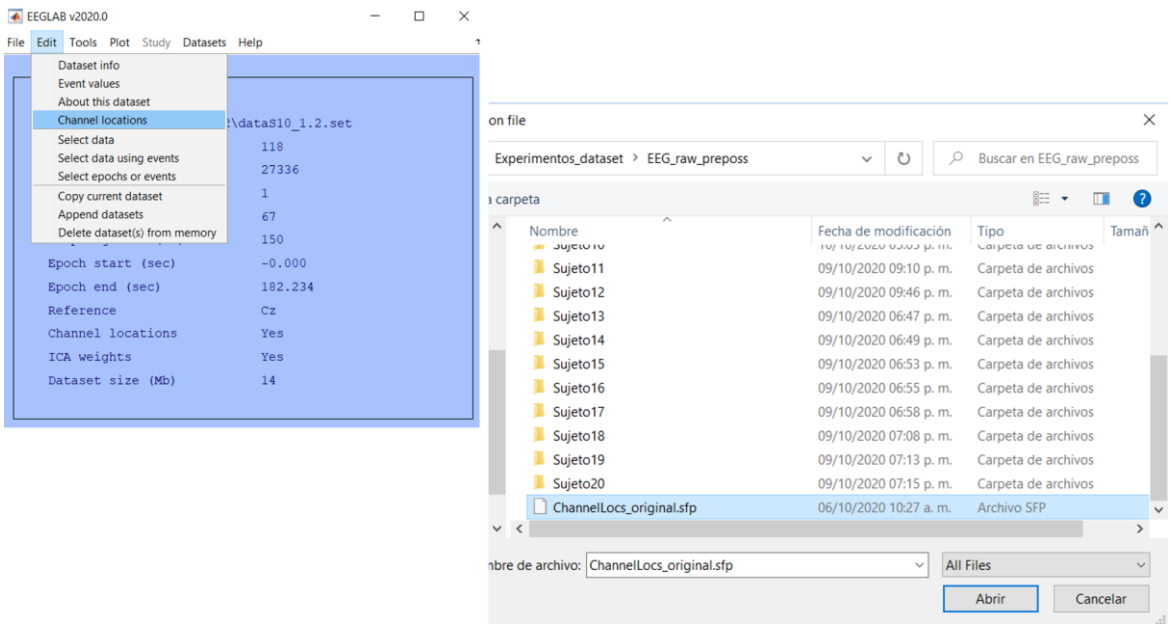

Figura 3.9: Importando el archivo de ubicación de canales.

Reducción de la Frecuencia de Muestro (Downsamplig) de

500 Hz a 250 Hz.

La seleccion de la frecuencia de muestro debe de estar orientada a los fines de la investigación, ya que esto beneficia directamente a los tiempos de computo del proceso posterior en la limpieza de los datos. Para los fines de esta investigacion se ha decidido establecer la frecuencia de muestreo en 250 Hz; ya que solo se analizar el espectro de la frecuencia de 0.5 a 50 Hz, que corresponde a las ondas oscilatorias Alpha, Beta, Teta, Delta de la señal EEG; las cuales están estrechamente relacionadas con el procesamiento cerebral del lenguaje.

Remover el ruido aleatorio (Baseline).

En esta etapa del preprocesamiento se analiz el espectro de la señal EEG con el fin de detectar y eliminar el ruido que pudiera afectar a la señal EEG, que puede ser proveniente de los aparatos médicos, luces fluorescentes, etc. Con la herramienta CleanLine del Toolbox de EEGLAB, procederemos a remover la frecuencia de 60 Hz mas su frecuencia armónica de 120 Hz, con el fin de eliminar el mayor ruido posible de la señal EEG, como se puede observar en la figura 3.10.

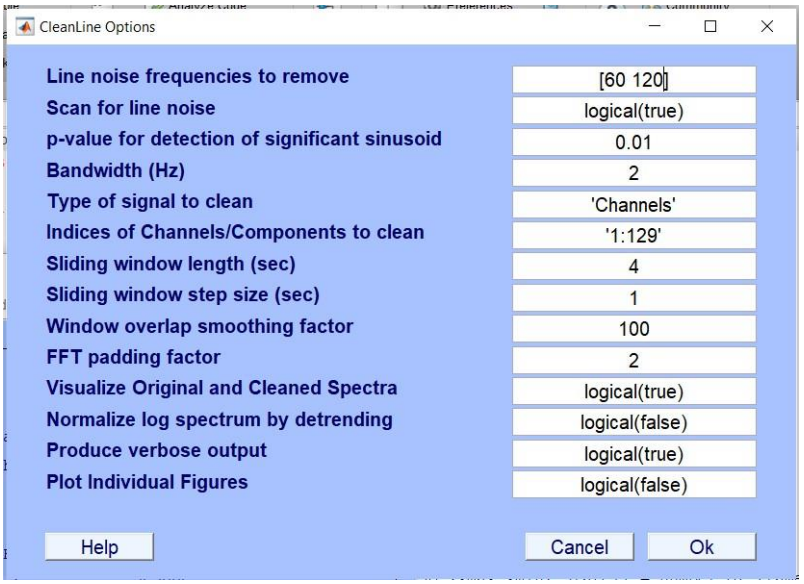

Figura 3.10: Limpieza de ruido de la señal EEG.

Con estos parametros, buscaremos eliminar la mayor cantidad de ruido de la señal EEG. Este algoritmo recorrer la señal con un ancho de ventana (time-bandwidth product) de 4 pulsos con respecto a su duración temporal y el espectro de la frecuencia, donde escaneará seccion por seccion la presencia de ruido lineal de la señal EEG, como se muestra en la figura 3.11.

```
Multi-taper parameters follow:
    Time-bandwidth product:  4
    Number of tapers:    7
    Number of FFT points:    2048
I'm going try to remove lines at these frequencies: [60  120] Hz
I'm going to scan the range +/-1 Hz around each of the above frequencies for the exact line frequency.
I'll do this by selecting the frequency that maximizes Thompson's F-statistic above a threshold of p=0.01.

OK, now stand back and let The Maid show you how it's done!

Cleaning Chan 1...
Computing spectral power...
Average noise reduction: 60.06 Hz: 16.41 dB | 75 Hz: -0.9593 dB
Cleaning Chan 2...
Computing spectral power...
Average noise reduction: 60.06 Hz: 16.07 dB | 75 Hz: -0.9221 dB
Cleaning Chan 3...
Computing spectral power...
Average noise reduction: 60.06 Hz: 16.45 dB | 75 Hz: -1.115 dB
Cleaning Chan 4...
Computing spectral power...
Average noise reduction: 60.06 Hz: 16.21 dB | 75 Hz: -2.728 dB
```

Figura 3.11: Resultados de la limpieza del ruido de linea.

Filtrado de los datos.

Para los fines de este trabajo de investigacion, se trabajarán con cuatro de las cinco frecuencias que se generan durante la actividad cerebral, dichas frecuencias son Alfa, Delta, Theta y Beta; las cuales involucran gran parte de la actividad relacionada con el procesamiento cerebral del lenguaje, como se describe en el capítulo dos en la sección ”Neurociencia del lenguaje”. Es por ello, que en el ToolBox EEGLAB, se filtra la señal a 50 Hz; con ello abarcamos las cuatro frecuencias mencionadas con anterioridad; este filtrado fue aplicado como se observa en la figura 3.12.

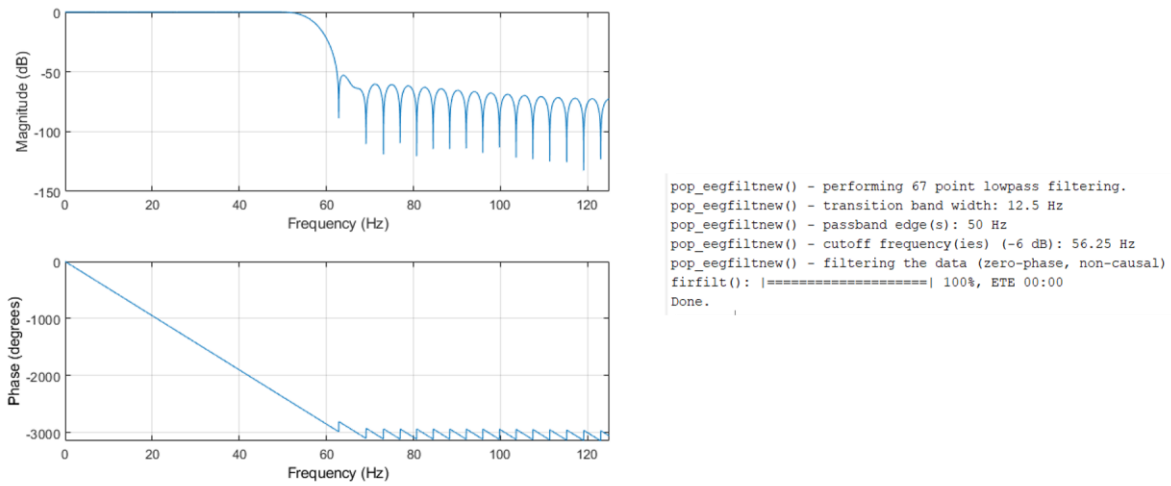

Figura 3.12: Filtrado de la señal EEG a 50Hz.

Para este proceso de filtrado de la señal EEG, se aplico un filtro del tipo FIR (Finite Impulse Response); el cual se caracteriza por ser un sistema no recursivo (ver figura 2.9), donde la respuesta al impulso es de duracion finita puesto que si la entrada se mantiene en ceros por N cantidad de periodos sucesivos el resultado de la salida también ser de cero. A pesar de su alto costo computacional, se tomo la decisión de aplicar este filtro con la finalidad de minimizar la retroalimentación de la señal EEG y brindarle estabilidad a la misma. En la figura 3.12, podemos observar el resultado de la aplicacion de este filtro, donde tenemos una frecuencia de corte 56.25 Hz en la fase y una frecuencia de corte de -6dB en la magnitud, ademas de un ancho de banda de transicion de 12. Hz.

Limpieza de artefactos en las señal EEG con la técnica ICA

Dado a la naturaleza sensible al ruido de la señal EEG, es importante someter a dicha señal a una etapa de limpieza de artefactos, como el pestañeo, latidos del corazón o movimientos involuntarios de la cabeza o señal mioelectrica proveniente de los musculos cercanos a los electrodos. Como primera instancia se realizara el Análisis de Componentes Independientes (ICA), que consiste en encontrar una fuente de

separacion lineal de las componentes, suponiendo que el origen de esta señal tiene una independencia estadística y es del tipo No Gausiano. Para llevar a cabo este proceso, se usara el ToolBox EEGLAB; el cual se eligen los componentes independientes para producir señales maximamente independientes; dicho proceso es lineal y que no añade información extra a la señal, al contrario identifica aquellas señales distintas he independientes, como se muestra en la figura 3.13; donde se observa como este proceso identifica claramente la deteccion del pestañeó en la señal EEG.

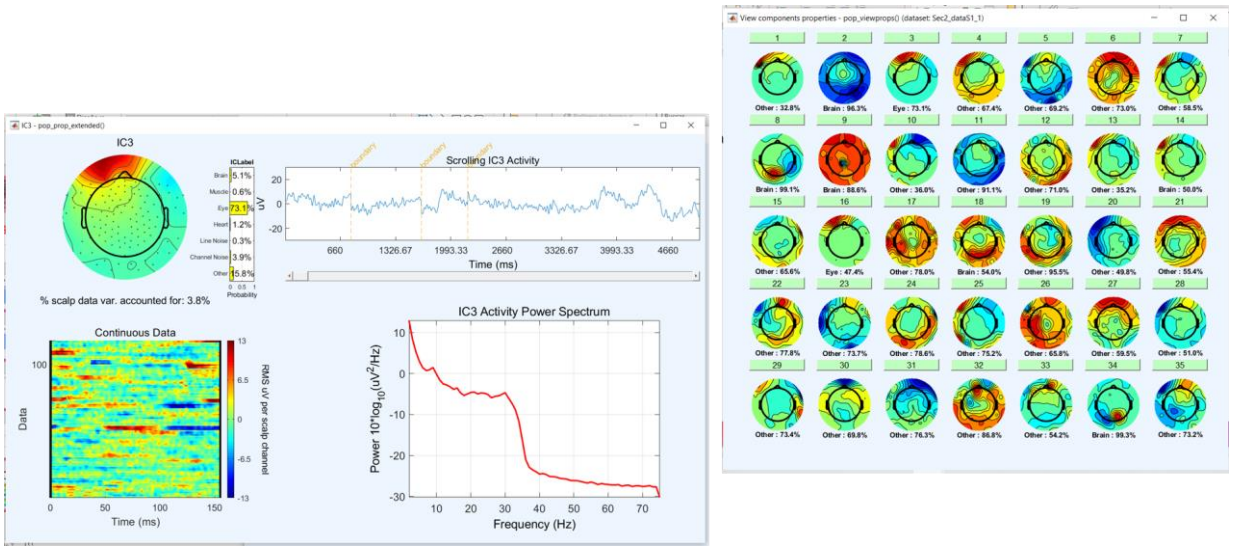

Figura 3.13: Componentes independientes de la señal EEG.

Para los fines de este trabajo de investigacion, se decidi formular un criterio de seleccion de que componentes resultado del Analisis de Componentes Independientes conservar y que componentes descartar; con el fin de trabajar un una señal con el menor porcentaje de ruido. Si bien, se tiene conciencia que la señal de EEG no podrá ser ”limpiada”del todo; por lo que se removeran todos aquellos componentes que contengan ruido de origen fisiologico o no fisiologico identificado, como se describe en la tabla 3.1:

Ruido Fisiológico	Ruido NO Fisiologico
Pestañeos Latidos del corazon Señal de músculos	Ruido en el canal/electrodo Ruido de la línea eléctrica

Tabla 3.1: Componentes a eliminar.

Es importante mencionar, que el resultado del analisis ICA arroj como producto un tipo de componente identific como Other; el cual se puede catalogar como una mezcla de señales de los componentes anteriores (tabla 3.1) y la señal del Electro-encefalograma. En este caso, se tomo la decision de conservar todo aquel canal que contenga un porcentaje de la señal Other; ya que este componente resulta de las mezclas de la señal EEG y el ruido identificado y por tal motivo, se trabajará con ese ruido añadido para no eliminar toda la señal EEG del análisis como se muestra en la figura ??.

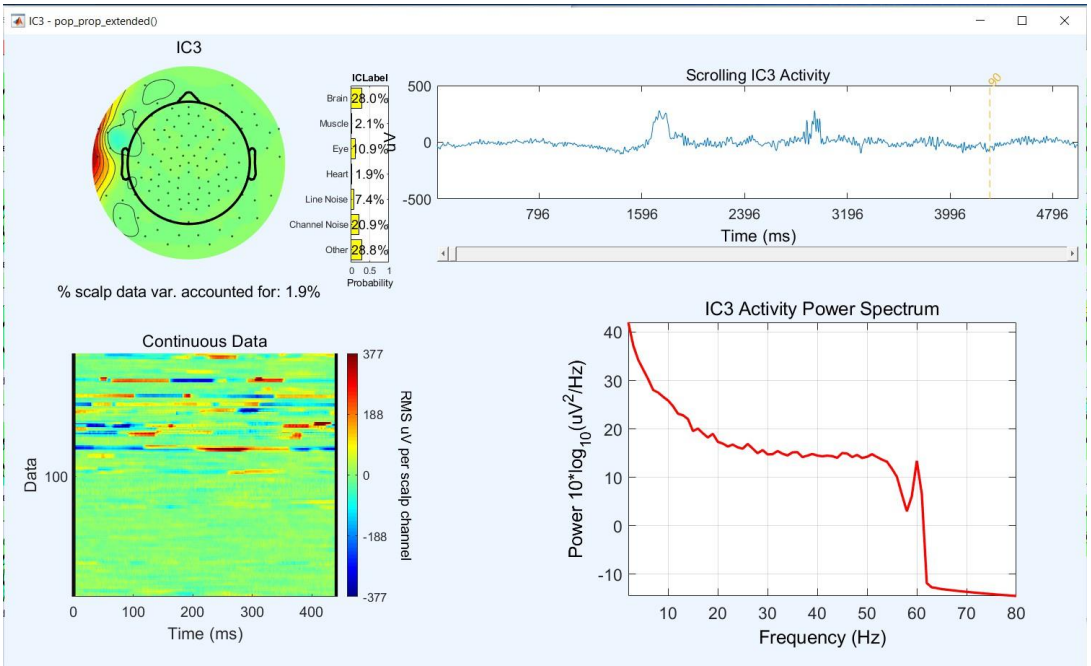

Figura 3.14: Componente identificado como Other.

Como podemos apreciar en la figura ??, este componente tiene un porcentaje del

28.8 % de esta mezcla de señales, así como un 28 % de señal cerebral; en combinacion con porcentajes provenientes de los artefactos a eliminar. Es conveniente recalcar que, solo se eliminaron todos aquellos componentes con un porcentaje mayor de 95 %, por debajo de dicho porcentaje se conserva el componente, ya que por la naturaleza de la señal es conveniente trabajar con cierto porcentaje de ruido.

Extraccion de características de la señal EEG a través de la Transformada Wavelet Discreta (DWT)

Una vez encontrados y procesados los Componentes Independientes de la señal EEG, se someterá la señal a una etapa de transformación a través la aplicacion de la Transformada Wavelet Discreta (DWT), donde se exploraran la familias Wavelet llamada Daubechies, las cuales han sido ampliamente usadas en diversos trabajos de investigacion [Daubechies, 1992, Torres García et al., 2013, Torres García et al., 2016] y cuyos resultados ha demostrados ser satisfactorios en su aplicacion dentro del área de las señales electroencefalograficas. Durante este proceso, se analizarán los niveles de descomposicion y la Wavelet madre de estas dos familias para seleccionar la mas adecuada para la identificación de los trastornos de lenguaje propuestos.

En esta etapa del procesamiento de la señal, se realizara la codificacion de los algoritmos en MATLAB, en el cual se empleara el Wavelet Toolbox [The MathWorks, 2019]; que mediante su función `mdwtdec()` podemos realizar la extraccion de las características de la señal; para ello se usaran los siguientes parametros:

- Indicador de Direccion: el cual indica al algoritmo la direccion de la descomposición por la cual recorrerá la matriz de datos; este Toolbox puede recorrer dos direcciones ya sea por filas ('r') o por columnas ('c').
- Datos de entrada: la señal de entrada puede ser matriz u otro objeto iterable. Que en nuestro caso es un archivo `.mat`, el cual contiene la señal EEG preprocesada.

- Wavelet Madre: Wavelet para usar en la transformación, correspondiente a una de las familias disponibles dentro del Toolbox, la cual puede ser ortogonal o biorthogonal.
- Nivel de descomposición: El cual deber ser identificado como un entero positivo. Cada nivel filtra la señal a través de un banco de filtros pasa bajas y un filtro de bancos pasa altas.

Teniendo en cuenta los parametros anteriores, la funcion mdwtdec() nos devuelve una lista ordenada de matrices de coeficientes, donde se denota cada nivel de descomposicion; que en nuestro caso el primer elemento de esta lista es el resultado de la matriz de coeficientes de aproximacion, mientras que los otros elementos de esta lista son matrices de coeficientes de detalle; una ves recopilados estos valores, se procede a realizar la reconstrucción de las señal EEG para ello se empleo la función x = mdwtrec(dec), a continuación mostraremos un pequeño ejemplo de codigo aplicando este toolbox.

```
%% Aplicando la DWT
% Descomponiendo la señal en 6 niveles
dec = mdwtdec('r',X,6,'db2');
%normalización de los coeficientes
decBIS = chgwdeccfs(dec,'cd',0,1);
%reconsruccion de la señal
Xbis = mdwtrec(decBIS);
figure;
plot(Xbis(1:5,:)','r');   hold on
plot(Xbis(21:25,:)','b');
plot(Xbis(31:35,:)','g')
grid; set(gca,'Xlim',[1,96])
%calculando la energía de c/u de las señales
[E,PEC,PECFS] = wdecenergy(dec);
```

Figura 3.15: Ejemplo de aplicacion del Wavelet Toolbox de MATLAB.[The MathWorks, 2019]

Es preciso destacar, que para los fines de esta investigacion se usaran 6 niveles de

descomposicion en cada una de las Wavelet Madre (Daubechies de segundo orden). Otra mencion que es importante destacar es, que hasta este punto, se trabajo con los archivos correspondientes a cada uno de los pacientes, es decir, se trabajo de manera aislada cada parte del preprocesamiento de la señal EEG; y es hasta el análisis con la DWT que se unen los datos. Para llevar a cabo la union de los datos de cada uno de los pacientes, se implemento un pequeño script de Python en el cual, a través de la librería de Pandas; que es ampliamente usada para el manejo y analisis de datos, que pueden llegar a ser de grandes dimensiones; y aplicando el atributo Chunzise, el cual nos permite leer esta gran cantidad de datos al seccionarlos, esto al establecer una medida con la cual se seccionaran los datos; se logra unir cada una de la matrices por paciente, en un archivo CSV el cual se usara en el analisis anteriormente descrito.

Reducción de la dimensionalidad de los datos a través de la técnica de PCA

Para la reduccion de la dimensionalidad de las características, se emplear la técnica

de Analisis de Componentes Principales, el cual trata de explicar la estructura de las varianzas y covarianzas de un conjunto de variables X_i, mediante unas cuantas combinaciones lineales de ellas, llamadas componentes principales. Estos componentes principales, no están correlacionados entre sí, y cada uno maximiza su varianza. EL PCA (por sus siglas en inglés), aspira a reducir o simplificar los datos, para facilitar su analisis e interpretación [Rodríguez Hernandez, 1998].El autor explica, que dicha reduccion es posible por que la variabilidad de los datos se puede explicar por un número k menor de componentes principales, de tal manera que el conjunto de datos original, se reduciría de n dimensiones por p variables. En otras palabras, este analisis es un método descriptivo que nos permite obtener una representacion de los datos, en un nuevo espacio dimensional que es resultado de los componentes principales.

Para efectuar el análisis PCA, en este trabajo de investigacion recurrimos nuevamen-

te al lenguaje de programación Python; para ello seguiremos las siguientes etapas.

- Cargar los datos: Tomaremos los datos obtenidos del aprendizaje activo.
- Estandarizar los datos: Para obtener un rendimiento óptimo al momento de llevar a cabo la clasificación; es necesario estandarizar o normalizar los datos entre las escalas 0 (media) y 1 (varianza).
- Análisis de Componentes Principales: Una ves extraído el número óptimo de componentes, se llevar a cabo el PCA, para finalmente guardar la nueva matriz de datos que será aplicada en el proceso de clasificacion.

Etiquetado de datos a través de Active Learning

Una vez completado el proceso anterior de reduccion de la dimensionalidad de los datos en el cual se unieron los datos de cada uno de los pacientes en una sola matriz de datos, se proceder a realizar el etiquetado de los datos. Dicho proceso se dividirá en dos fases, ambas de gran importancia para este trabajo de tesis, la primera de ellas llamada ”Identificación de las clases (Clusterización de los datos)” consiste en el agrupamiento o clusterización de los datos de acuerdo a las ondas cerebrales que intervienen en el ”Proceso Cerebral del Lenguaje”, es importante mencionar que esta agrupacion o clusterizacion de datos será usada como anotador en el proceso de consulta de la técnica de Aprendizaje Activo (Active Learnig).

La segunda fase consta del modelado y aplicación de la técnica de Aprendizaje Activo (Active Learnig), la cual usaremos para el proceso de etiquetado con el que pretendemos identificar el ”Proceso Cerebral del Lenguaje.ª través de las ondas cerebrales. A continuación, se describir con mayor detalle cada una de estas fases que comprenden el proceso de ”Etiquetado de datos a través de Active Learning”.

Identificacion de las clases (Clusterizacion de los datos)

Es preciso mencionar, que dentro de la literatura correspondiente al Aprendizaje Activo (Active Learning), nos podemos valer de un conjunto de datos previamente etiquetados, es decir que tenga las clases previamente identificadas, con la finalidad de que estos asuman el rol de anotador para el proceso de consulta: como se menciona en el segundo capitulo de este trabajo de tesis, en la sección "Aprendizaje Activo(Active Learning)". Habiendo aclarado esto, enseguida describimos como se llevo a cabo la fase de clusterizacion de los datos para identificar las clases que serán usadas en el proceso de Aprendizaje Activo (Active Learnign) y que serán usados como anotador en el proceso de etiquetado llevado por la segunda fase de este procedimiento de Aprendizaje Activo (Active Learning).

Primeramente se seccionara el set de datos Child Mind Institute - Multimodal Resource for Studying Information Processing in the Developing Brain (MIPDB) [Simon P. Kelly, 2016] en el cual estamos usando los datos de 6 pacientes seleccionados al azar, es decir una matriz con dimensiones 768x 182177, la cual se segmentó en dos grupos de menor porcentaje. El primer grupo ser de un 50 % de los datos; dicho grupo se someter a un proceso de clusterizacion con el fin de generar las etiquetas o clases tomando en cuenta el papel lingüístico que tienen en el proceso cerebral del lenguaje (ver tabla 3.3) y que seran empleadas posteriormente en la estrategia de Aprendizaje Activo (Active Learning); dicho proceso de clusterización ser ejecutado en el Wavelet Toolbox de Matlab, para realizar la clusterización de los datos, se uso el método de agrupacion jerarquica (Ascending Hierarchical Clustering) y para calcular la distancia entre los clusters, se empleo la técnica Distancia media ponderada (WPGMA) como método de vinculacion (linkage) sobre los valores obtenidos de las descomposición Wavelet Daubechies, en específico, la energía relativa obtenida del tercer nivel de detalle de la descomposición (D3), el cual resultó el mejor en el proceso de clusterizacion, como se muestra en la figura 3.16.

Este proceso de identificacion de clases (clusterizacion), se definió para obtener cuatro clases, los cuales englobaran aquellos datos de interés para esta investigacion, es decir, estas clases fueron formados por los resultados obtenidos de la aplicación de la trasformada Wavelet Daubechies, tomando como principal característica la energía de la Wavelet; dentro de estos resultados se engloban las características lingüísticas que se pueden encontrar en el "Proceso cerebral del lenguaje" que origina en el hemisferio izquierdo del cerebro y en donde pueden intervenir las frecuencias Alfa, Delta, Teta y Beta (ver tabla 3.3).Dichos clases, fueron sometidas a un analisis, con la finalidad de observar las características presentes dentro de cada una de ellas; en este proceso de identificacion de las clases se pudieron determinar 4 (cuatro) clases principales como se listan a continuacion (ver tabla 3.2), así como se describe el impacto que genera la presciencia de cada una de estas clases.

Clases identificadas en el proceso de Clusterizacion		
Clase	Identificador de la clase	Descripción del impacto
Procesamiento de frases	1	Clasificacion de las palabras y/o frases en categorías gramaticales.
Retencion de palabras	2	Generacion de un vocabulario amplio.
Comprension del léxico	3	Habilidad para el entendimiento o compresion de una lengua.
Ordenamiento de palabras	4	Habilidad para la construccion sintactica dele lenguaje(Ordenamiento de palabras, construccion de oraciones, construccion de frases).

Tabla 3.2: Clases identificadas en el proceso de Clusterizacion, que seran usadas en le proceso de Aprendizaje Activo (Active Learning).

Ondas cerebrales en el Proceso del Lenguaje	
Onda	Papel Lingüístico
Delta	Procesamiento de las frases. Posible deferenciacion de las palabras en categorías gramaticales.
Theta	Involucrado en los procesos de análisis de la memoria.
Alfa	Interconecta partes del encéfalo y participa en el acceso léxico.
Beta	Participa en la clasificacion sintactica de las palabras y la retencion de las palabras para construir estructuras sintacticas.
Gamma	Realiza reconstrucciones sintacticas. Se involucra en la multitud de procesos lingüísticos.

Tabla 3.3: Ondas cerebrales que se analizaron en este trabajo de Investigación.[Morales, 2020]

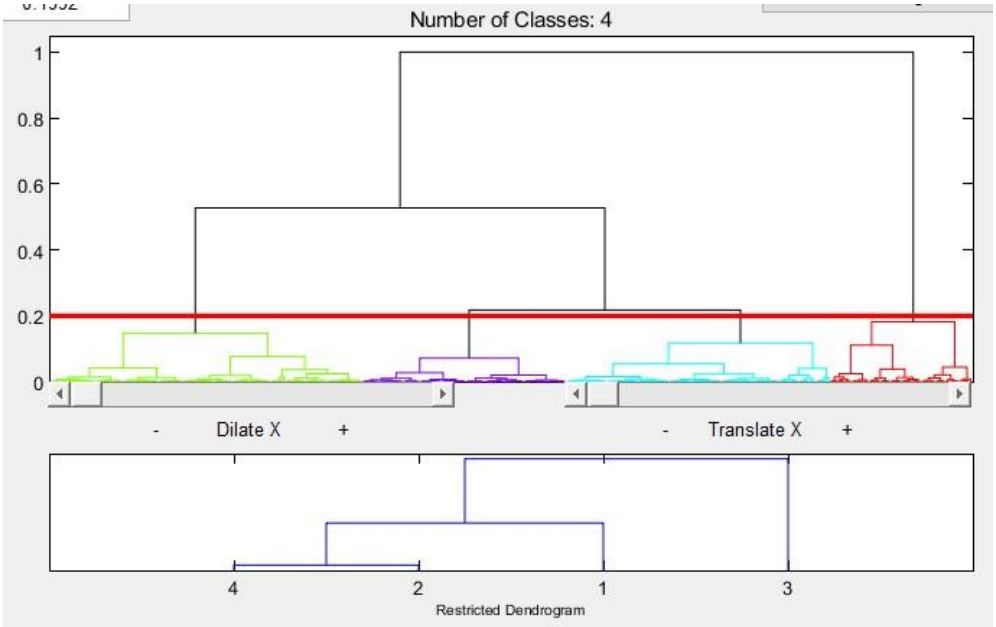

Figura 3.16: Clusterización de los datos.

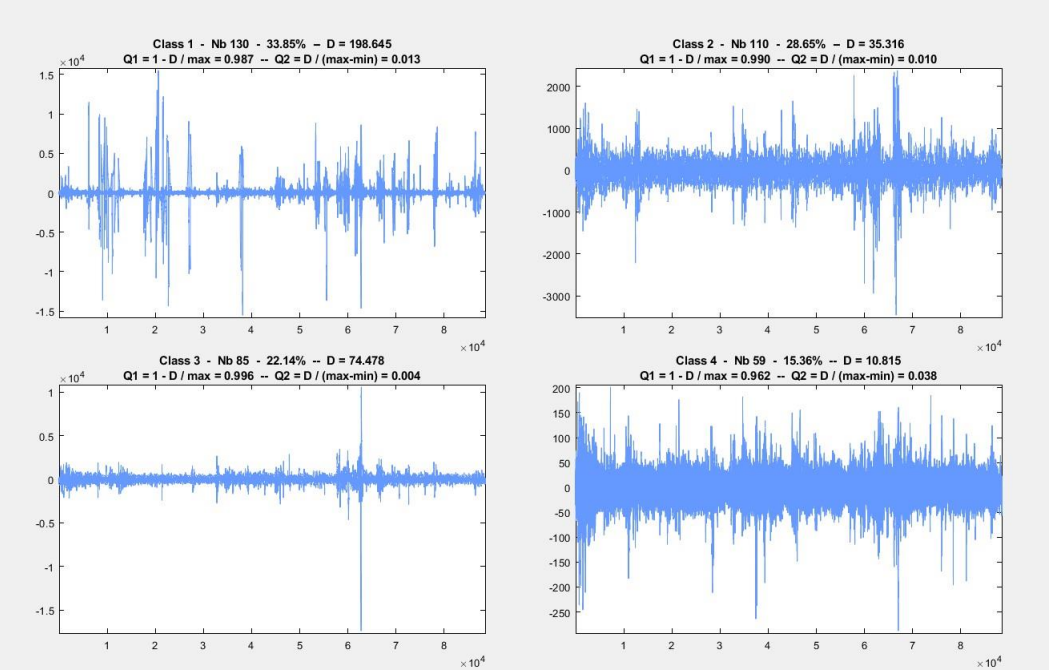

Figura 3.17: Analisis de cada uno de los clusters.

En la figura 3.17 podemos observar la agrupacion de los datos para el proceso de identificacion de clases, que representan el 50 % de los datos de la muestra total, es decir, un total 384 vectores con una longitud de 182177 (dimensiones de la matriz 384X182177). Dicha agrupación se dio tras buscar las cuatro clases principales antes mencionadas (ver tabla 3.2), que pueden ser desencadenadas por la aplicación de dos paradigmas incluidos dentro del estudio que origino el set de datos Child Mind Institute - Multimodal Resource for Studying Information Processing in the Developing Brain (MIPDB) [Simon P. Kelly, 2016], el primer paradigma es Sequence Learning, en el cual se incluye una secuencia visual de aprendizaje para memorizar localizaciones y los pacientes daban reportes de lo aprendido de manera regular. El segundo paradigma que se emplea es Naturalistic Viewing, en el cual se les mostraban a los pacientes cortometrajes de acuerdo a su rango de edad, esto con la finalidad de provocar un estimulo audio-visual de alto nivel; en la figura 3.1 se pueden observar dichos paradigmas. El segundo grupo de datos, que corresponde al 50 % restante no será sometido a ningún tipo de proceso, por lo que permanecer sin etiquetar y ser nuestro grupo de datos sin etiquetar, los cuales seran usados para

el proceso de etiquetado de los datos que se describe a continuación.

Aprendizaje Activo (Active Learning)

Uno de los principales objetivos del Aprendizaje Activo (Active Learning) es lograr obtener resultados similares a los obtenido con técnicas tradicionales de aprendizaje supervisado; sin la necesidad de tener un conjunto de datos etiquetados muy grande. Teniendo lo anterior en cuenta, se tomara un 50 % de los datos como Pool, que será empleado en el proceso de consulta del Aprendizaje Activo y el 50 % de los datos seran datos etiquetados previamente como se describe en la seccion "Identificacion de las Clases (Clusterizaciíon de los datos)" del tercer capitulo de este trabajo de tesis y como se observa en la tabla 3.2. Para la etapa de proceso de etiquetado a través de la tecnica de Aprendizaje Activo (Active Learning), se uso el FrameWork modAL para Pyhton [Danka and Horvath, 2018]; el cual fue creado sobre la librería scikit-learn, lo que nos permite crear rapidamente flujos de trabajo de Aprendizaje Activo (Active Learning) de manera modular, flexible y extensible. Como se mencion en la sección "Aprendizaje Activo (Active Learning)" del marco teorico de este trabajo de tesis; esta técnica de aprendizaje semi-supervisado tiene como objetivo acelerar el proceso de aprendizaje, especialmente si no se cuenta con un conjunto de datos muy extenso como para aplicar métodos de aprendizaje supervisado tradicionales.

En la figura 3.18, podemos observar con mayor detalle el proceso de etiquetado a través del Aprendizaje Activo (Active Learning) que sera ejecutado en este trabajo de investigacion.

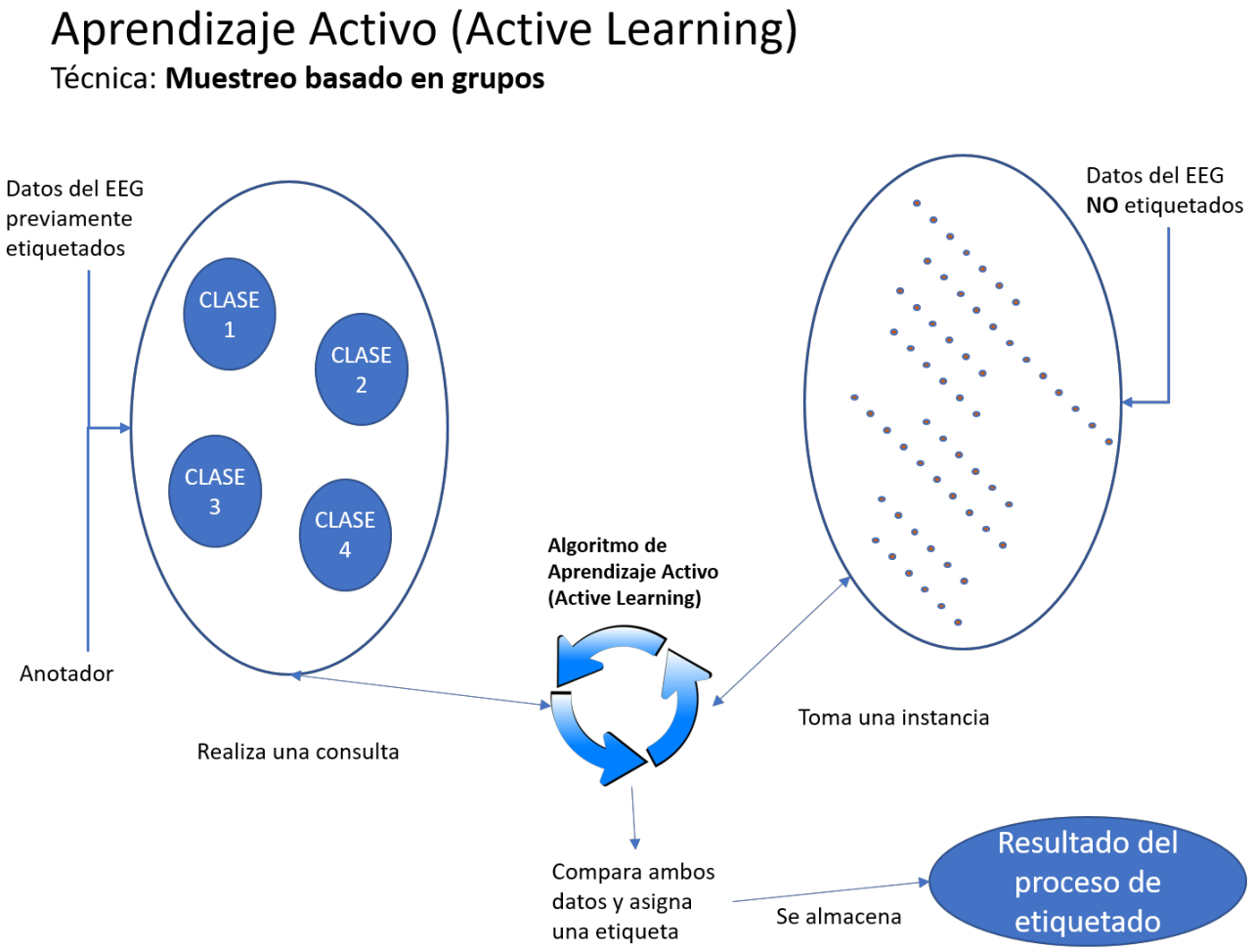

Figura 3.18: Proceso de Aprendizaje Activo que será empleado en este trabajo de investigación.

Se plantea entonces aplicar la técnica de Muestreo basado en grupos (Pool-base Sampling); dicha estrategia supone un pequeño grupo de datos etiquetados y un gran grupo de datos sin etiquetar, de tal forma las consultas se extraen selectivamente del grupo, como se describe en la seccion ”Aprendizaje Activo (Active Learning)” del marco teorico de este trabajo de tesis.

Para llevar a cabo esta estrategia de Aprendizaje Activo (Active Learning), crearemos algunos escenarios de experimentación, donde emplearemos como estimador principal las métricas de clasificacion (distancia, método de votacion, etc) empleadas por los clasificadores K-nearest neighbor y Random Forest; cabe destacar que en el proceso del algoritmo de Aprendizaje Activo (Active Learning), es necesario emplear una métrica de estimacion que puede ser un modelo matematico o la aplicación de un clasificador ya conocido donde a través de sus métricas de clasificacion se realiza el proceso de estimación, que en el caso de esta investigación, se opt por elegir a los dos clasificadores antes mencionados, y que para los fines de esta investigacion y en la redaccion de este documento, se referir a estos clasificadores como estimadores, ya que en la literatura se hace referencia de esta manera y se explicara más a detalle en los resultados de este trabajo. Y como estrategia de consulta, emplearemos el muestreo de incertidumbre (Uncertainty sampling); esta estrategia de consulta, calcula la utilidad de la predicción por cada ejemplo y selecciona una instancia en funcion de la utilidad. Para aplicar todo lo anterior, nos situaremos dentro del framework modAl de Python, que estar dividido en las siguientes etapas:

Cargando los datos

Primeramente, procederemos a cargar los datos a utilizar; para ello cargaremos los datos obtenidos en el proceso de identificación de clases donde se determinaron las clases y/o etiquetas necesarias que requiere el proceso de Aprendizaje Activo (Active Learning), ya que como se menciona en la literatura, podemos usar un segmento de datos previamente etiquetados que serviran como anotador en el proceso de consulta como se describe en en la seccion "Aprendizaje Activo (Active Learning)" del marco teorico de este trabajo de tesis. Es importante mencionar que, en conjunto a estos datos que se mencionan se encuentra un segmento de datos que no poseen ninguna etiqueta y que solo han sido procesados, por lo tanto, tenemos una matriz de datos de dimensiones 768x182177. Posteriormente, dividiremos el conjunto de datos en dos; que corresponden al conjunto de entrenamiento que constara del segmento de

datos previamente etiquetados que representa el 50 % de la muestra; así como el conjunto de Pool de datos sin etiquetar que constara del 50 % de los datos restantes, como se muestra en la figura 3.19.

```
#CARGANDO EL SET DE DATOS
filename = 'EEGdata_f_s1-6.csv'
dataset=pd.read_csv(filename,sep=',').values
#SEPARANDO LAS CARACTERISTICAS DE LOS TARGETS
Xdata = dataset[0:383,:-1] #todas las columnas excepto la ultima(datos etiquetados)
ydata = dataset[0:383,len(dataset[0])-1] #solamente la ultima columna(targets)

##GENERANDO EL POOL DE DATOS
X_pool = dataset[384:690,:-1] #datos sin etiquetar
```

Figura 3.19: Segmentación de los datos de la matriz Inicial (768X182177) en dos grupos.

Después de realizar esta primera segmentacion del set de datos, procederemos a realizar una segunda segmentacion en el conjunto de datos etiquetados para obtener dos subconjuntos de entrenamiento y prueba, dicha división la realizaremos empleando la librería ”train _test _split” como se muestra en la figura 3.20.

```
##SEGMENTANDO LOS DATOS EN ENTRENAMIENTO Y PRUEBA
from sklearn.model_selection import train_test_split #importando la libreria
X_train, X_test, y_train, y_test = train_test_split( Xdata, ydata, train_size=0.10)
print(X_train.shape, X_test.shape, y_train.shape, y_test.shape)
```

Figura 3.20: Segmentación de los datos en subconjuntos de entrenamiento y prueba.

Iniciando el Aprendizaje Activo

Para iniciar el proceso de aprendizaje, en el framework de modAL se debe crear un objeto, que contenga tres variables necesarias; la primera de ellas es el estimador que en nuestro caso se trata del clasificador KNeighbors y posteriormente se ejecutara el clasificador Random Forest, los cuales son llamados desde la librería de scikit-learn. La segunda variable que se tiene que tener es la estrategia de consulta, que en nuestro caso sera el muestreo de incertidumbre (Uncertainty sampling). Y por ultimo, se pasaran los datos de entrenamiento y prueba iniciales con los que trabajara el framework. Hasta este punto, tenemos la corrida inicial del Aprendizaje Activo, por lo que es necesario actualizar nuestro modelo y así podamos pasar a la siguiente etapa.

Actualización del modelo de Aprendizaje Activo

En esta etapa el modelo aún sigue en aprendizaje, por lo que es necesario actualizarlo; esto se llevara a cabo a través de las consultas o querys, es en esta etapa donde entrar en accion la estrategia de consulta o muestreo de incertidumbre (Uncertainty sampling); cuyo objetivo es la reducción de la cantidad de incertidumbre en sus predicciones; con cada consulta, se eliminar la instancia del grupo y se registrara la precision del modelo.

Finalmente se evaluar el rendimiento del modelo de Aprendizaje Activo, para lo cual graficaremos las predicciones correctas e incorrectas obtenidas durante el proceso; con lo que se finaliza la etapa de Aprendizaje Activo para así continuar con el desarrollo de esta investigación.

Proceso de clasificación

Una vez terminado el proceso de etiquetado con la técnica de Aprendizaje activo (Active Learning) que se describe en la figura 3.18, procedemos a guardar los resultados de la prediccion de las etiquetas. Estos resultados seran almacenados en archivos CSV, como se muestra en la figura 3.21.

```
# Quitando la instancia consultada del pool de datos sin etiqueta
X_poolr = np.delete(X_pool, query_index, axis=0)
df1=pd.DataFrame(X_poolr)
df1.to_csv('X_dataC.csv')#guardando los datos
y_poolr = np.delete(y_pool, query_index)
df2=pd.DataFrame(y_poolr)
df2.to_csv('y_dataC.csv')#guardando los datos
```

Figura 3.21: Salvando los resultados del Aprendizaje Activo (Active Learning).

Es importante mencionar que, dentro de el archivo y_dataC.csv se encuentra la prediccion de las etiquetas de los datos evaluados durante el proceso de consulta que se llevo a cabo en cada escenario de experimentación del proceso de Aprendizaje Activo (Active Learning). Es en este punto, donde podríamos decir que ya hemos asignado una etiqueta o "target" perteneciente a una clase (ver tabla 3.2) al 50 % de los datos que estaban sin identificar y que se encuentran almacenados en el archivo X_dataC.csv, por lo que ahora tenemos un set de datos completamente etiquetado con sus respectivos "targets" de cada una de las clases determinadas en la sección "Proceso de identificación de clases (Clusterizacion de los datos)" como se describe en el tercer capitulo de esta tesis, con lo que obtendremos una matriz de 690X182177; por lo que ahora podemos someter a este set de datos etiquetado a procesos Machine Learnign tradicionales como técnicas de clasificacion supervisada. Habiendo dicho lo anterior, continuaremos describiendo el proceso de clasificación que se efecto en este trabajo de investigacion.

Como se menciono con anterioridad, lo primero que se realizó fue la union de los datos previamente etiquetados en el proceso de clusterización, como se describe en la seccion ”Identificación de clases (Clusterización de los datos)” del tercer capítulo de esta tesis; en union con los resultados del proceso de etiquetado con la técnica de Aprendizaje Activo (Active Learning). Dicha unión se llevo acabo a través de la union de matrices, usando el lenguaje de programacion MATLAB como se muestra en la figura 3.22.

```
%% unir las dos matrices de datos e indices
%DATOS
dataEEG_fr=[A;B];
%INDICES
Idx_fr = [IdxA;IdxBd];
%% unir la matriz final
EEG_final = [dataEEG_fr Idx_fr];
```

Figura 3.22: Uniendo la matriz de datos etiquetados con la matriz de resultado del proceso de Aprendizaje Activo (Active Learning)

Una vez que obtenemos la matriz de datos resultante de este proceso de union, se proceder a realizar la clasificacion, con la finalidad de detectar los patrones del

”Proceso Cerebral del Lenguaje” que se da en le hemisferio izquierdo del cerebro y del cual se obtuvieron las clases cuyas características nos permiten identificar dicho proceso (ver tabla 3.2). Para los fines de esta investigación, realizaremos una comparativa entre clasificadores, con lo que someteremos los datos obtenidos hasta el momento, resultado de los procesos anteriores del desarrollo experimental.

Para el desarrollo de dicha comparativa, efectuaremos la clasificación de los datos aplicando varios de los clasificadores más comunes como lo son el K-nn, Maquina de Vector Soporte, Naive Bayes y Multi Layer Perceptron. En cuanto a esta etapa de clasificacion de los datos, desarrollaremos el algoritmo en el lenguaje de programacion Python, donde aplicaremos la libería Scikit-learn, que es usada ampliamente en

Machine Learning, ya que incluye una variedad de algoritmos de clasificacion, regresion y analisis. Cabe destacar que, para proceder con el proceso de clasificación es necesario someter al set de datos a la técnica de Análisis de Componentes Principales (PCA, por sus siglas en Inglés); esto con la finalidad de reducir la dimension de los datos, ya que como recordaremos nuestra matriz tiene una dimension de 690x182177 y esto resulta en una dimensión muy grande para poder analizar durante el proceso de clasificacion y que podría resultar en un alto costo computacional. Es por ello, que al someter este set de datos ya etiquetado a la técnica de PCA, obtendremos una matriz de 690X21, siendo la ultima columna la que contenga la informacion de nuestros ”targets” los cuales corresponden a las clases que se determinaron en la seccion ”Proceso de identificacion de clases (Clusterizacion de los datos)” del tercer capitulo de este trabajo de tesis.

Así pues, ejecutaremos las siguientes etapas para el proceso de clasificación.

- Cargando los datos: Procederemos a leer la matriz de datos que se obtiene del proceso experimental.
- Seccionar los datos: se separan los datos de las etiquetas o ”targets”, con la finalidad de conformar los conjuntos de entrenamiento y prueba que seran empleados en el proceso de clasificacion.
- Clasificacion: implementación de los algoritmos a través de la librería de Scikit-learn.
- Resultados: graficar las matrices de confusión de cada uno de los algoritmos de clasificación.

En el siguiente capitulo, nos adentraremos en los resultados que se obtuvieron durante el desarrollo de este trabajo de investigacion.

Capítulo 4

Resultados

En relacion con la problematica expuesta en este trabajo de investigación, se planteó como objetivo el modelado de un sistema BCI capaz de etiquetar señales electroencefalograficas durante el proceso cerebral del lenguaje, esto a través del uso de técnicas de Aprendizaje Activo (AL, por sus siglas en inglés). Es por ello, que se plante la realizacion de dos escenarios de experimentación como se observa en la figura 4.5 y que efectuaremos usando el Framework de Python llamado modAL [Danka and Horvath, 2018], el cual fue desarrollado sobre la librería scikit-learn.

Es importante mencionar que el Framework de modAL, ya tiene implementado un proceso o función de aprendizaje dentro de el, donde se emplean clasificadores (p.e. K-nn) como estimadores y que a través de sus métricas se realiza el cálculo para la prediccion de las etiquetas; es por ello que a pesar de ser conocidos por ser clasificadores en la literatura del Apredizaje Activo se refieren a ellos como estimadores. Otro proceso que ya se encuentra dentro del Framework, es la funcion para la estrategia de consulta (Query strategy), la cual evalúa la informacion de las instancias o datos sin etiquetar y que de forma iterativa puede ser generada la consulta a partir de una distribucion determinada [Settles, 2010].

Siguiendo con el mismo orden de ideas, en este trabajo se efectuar el desarrollo experimental empleando la técnica y/o escenario de Aprendizaje Activo de Muestreo Basado en Grupos (Pool-based Sampling), que servirá como base y sobre esta técnica se le realizarán variaciones en el proceso de aprendizaje del algoritmo y la estrategia de consulta. Primeramente, en el proceso de aprendizaje del algoritmo se aplicar la función de aprendizaje ya establecida; usando como estimadores dos clasificadores, el primero de ellos es el clasificador k vecinos más cercanos (k-nearest neighbors, o K-nn); el cual es un método de clasificacion no perimétrico que estima la probabilidad a posteriori de una instancia o dato pertenezca a una clase; dicha estimacion se calcula a través de la funcion de densidad [Bishop, 2006]. El otro clasificador que será usado como estimador ser el clasificador de Bosques Aleatorios (Random forest), el cual consta de la combinación de árboles de decisión independientes entre ellos, y dicha independencia les permite escoger un subconjunto de datos; donde el voto mayoritario de un conjunto de árboles es el que asigna el dato a una determinada clase [AA.VV., 2018]. Es importante recalcar, que para los fines de esta investigacion y la redacción de este documento, se har referencia a estos clasificadores como los estimadores de la prediccion de la funcion de Aprendizaje Activo (AL), por lo que en párrafos siguientes los mencionaremos como 'Estimador K-nn' y 'Estimador Random Forest'.

La segunda variacion, que se efectuara en la técnica y/o escenario de AL, se aplicará en la estrategia de consulta que realiza el algoritmo durante el proceso de etiquetado de los datos, recordemos que dicha funcion ya esta integrada en el Framework de modAL. La primer estrategia de consulta que será usada lleva por nombre Muestreo por incertidumbre (Uncertanty sampling), dicha estrategia encuentra el ejemplo mas útil y lo expone para ser etiquetado; esto se calcula con la expresion.

$$U(x)=1-P(—x), \tag{4.1}$$

Donde x es la instancia o dato que se va a predecir y $\hat{x}$ es la predicción más probable;

al aplicar esta estrategia de consulta, se seleccionará la instancia o dato con mayor incertidumbre [Danka and Horvath, 2018]. Otra de las estrategias de consulta que seran aplicadas a este desarrollo experimental, se le conoce como Muestreo en modo de lote (Ranked batch-mode sampling); esta estrategia genera una consulta cuyo resultado es una lista optimizada de instancias o datos a etiquetar, esto en base a algunos criterios de calidad lo que permite que los lotes (batch) sean de tamaño arbitrariamente grande [Cardoso et al., 2017]. En esta estrategia de consulta, la puntuacion de cada instancia o dato se calcula a través de la expresión:

$$\text{score} = \alpha(1 - \Phi(x, X_{\text{labeled}})) + (1 - \alpha)U(x), \tag{4.2}$$

donde $\alpha = X_{\text{unlabeled}}/X_{\text{unlabeled}} + X_{\text{labeled}}$, X_{labeled} es la intancia o dato sin etiquetar, $U(x)$ es la incertidummbre de la predicción para x y Φ es llamada funcion de similitud, que mide que tan bien se explora el espacio de características cerca de x.

Con estas variaciones que se mencionaron anteriormente, se efectuar el desarrollo experimental de este trabajo de investigacion. A continuacion, describiremos el corpus de datos usado en esta etapa del trabajo de investigación. Es importante destacar, que se uso el mismo corpus de datos en ambas variaciones de los experimentos.

Durante la planificacion del desarrollo del trabajo de investigacion, se proyecto en primera instancia el tomar las muestras de 31 pacientes y/o voluntarios del set de datos Child Mind Institute - Multimodal Resource for Studying Information Processing in the Developing Brain (MIPDB) [Simon P. Kelly, 2016]. Pero tras los cambios efectuados en este trabajo de investigacion, y haciendo un analisis se determino el uso de solo un 20 % del total de los pacientes; quedando el corpus de datos de la siguiente manera (ver figura 4.1).

PACIENTES	MUESTRAS POR PACIENTE	LONGITUD DE LA MUESTRA
6	128	182177

Figura 4.1: Corpus de datos usado en la fase experimental.

Estos seis pacientes que se muestran en la tabla, fueron tomados al azar del conjunto inicial de 31 pacientes. Cabe destacar, que en este punto ya se ha completado la fase del preprocesamiento de la señal de cada uno de los pacientes; es decir, se ha realizado la limpieza de la señal EEG para trabajar con el menor ruido posible, como se describe en la sección ”Preprocesamiento de los datos”, del tercer capitulo de este trabajo de tesis. Posteriormente, se tomaron conjuntos de tres pacientes para construir una sola matriz de dimensiones 384x182177; donde cada matriz representa el 50 % de los datos totales del corpus. Una de estas matrices fue sometida a un proceso de agrupamiento (clusterización) para obtener las etiquetas, como se detalla en el tercer capitulo de este trabajo de tesis en la seccion ”Proceso de identificacion de clases (Clusterizacion de los datos)”; en este proceso se tom como base la energía de la Wavelet para aplicar un clusterizador jerarquico, dichas clases que se determinaron estan identificadas con los números 1, 2, 3 y 4 como se muestra en la tabla 3.2, las cuales corresponden a las ondas cerebrales que se presentan en el ”Proceso Cerebral del Lenguaje” y el 50 % restante de los datos se encuentra sin etiquetar. Finalmente, se unieron estas dos matrices para obtener el corpus de datos completos (dataframe), cuyas dimensiones son 768x182177.

Como se mencionó con anterioridad, para llevar a cabo la fase de experimentacion no apoyaremos de la técnica y/o escenario de Aprendizaje Activo Muestreo Basado en Grupos (Pool-based Sampling), que se encuentra integrada en el Framework modAL de Python. Dicha fase se dividirá en dos grandes escenarios de experimentacion, que seran agrupados de acuerdo a la estrategia de consulta del algoritmo; es decir, identificaremos a estos escenarios de experimentación por los nombres Uncertanty Sampling (US) y Ranked Batch-mode Sampling (RBMS), en la figura 4.5 se muestra la estructura del proceso experimental empleado en este trabajo de investigación. A continuacion, se describir con detalle cada uno de los escenarios de experimentacion.

En el primer escenario de experimentacion corresponde al Uncertanty Sampling (US); en este escenario de experimentación, se segmentaron los datos en dos grupos el primero de ellos abarca los datos etiquetados que corresponde al 50 %, mientras que el segundo grupo corresponde a los datos sin etiquetar, es decir, el 50 % restante. Del primer grupo, se realizó una segunda segmentación donde se separaron los datos de las etiquetas (targets), posteriormente de este grupo se tomaron los datos para crear los conjuntos de entrenamiento y las etiquetas de prueba, este proceso de segmentacion se realiz con el apoyo de la librería Sklearn a través de su complemento

train_test_split, donde se tomo el 10 % de los datos para el conjunto de entrenamiento y el 90 % restante para el conjunto de prueba, dichos datos son seleccionados de manera aleatoria; es importante destacar que dentro de este conjunto solo usaremos el correspondiente a las etiquetas (targets) el cual se usara como Pool para la etapa de consulta del algoritmo de AL. El segundo grupo, que es el 50 % restante pasa directamente al pool de datos que seran consultados; como se muestra en el anexo 4.2.

```
#CARGANDO EL SET DE DATOS
filename = 'EEGdata_f_s1-6.csv'
dataset=pd.read_csv(filename,sep=',').values
#SEPARANDO LAS CARACTERISTICAS DE LOS TARGETS
Xdata = dataset[0:383,:-1] #todas las columnas excepto la ultima(datos etiquetados)
ydata = dataset[0:383,len(dataset[0])-1] #solamente la ultima columna(targets)

##SEGMENTANDO LOS DATOS EN ENTRENAMIENTO Y PRUEBA
from sklearn.model_selection import train_test_split #importando la libreria
X_train, X_test, y_train, y_test = train_test_split( Xdata, ydata, train_size=0.10)
print(X_train.shape, X_test.shape, y_train.shape, y_test.shape)
```

Figura 4.2: Segmentación de los datos en los escenarios US y RBMS.

El proceso de aprendizaje se inicia con la variable learner, a la cual se le pasan los datos de entrenamiento obtenidos con anterioridad; cabe destacar que este proceso de aprendizaje corre en segundo plano; dicho proceso toma los argumentos como el estimador; es importante mencionar que en esta sección del aprendizaje del algoritmo aplicaremos dos variaciones en el estimador, donde usaremos las métricas de los clasificadores K-vecinos más cercanos (k-nn=4) y el clasificador de Bosques Aleatorios (Random forest), como podemos ver a manera de ejemplo en la figura 4.3.

```
##INICIANDO EL ACTIVE LEARNER
learner = ActiveLearner(
    estimator=KNeighborsClassifier(n_neighbors=4), ##estimador
    query_strategy = uncertainty_sampling,
    X_training=X_train, y_training=y_train
)
# pool-based sampling
n_queries = 100
performance_history = list()

for idx in range(n_queries):
    query_idx, query_instance = learner.query(X_pool)
    learner.teach(
        X=X_pool[query_idx].reshape(1, -1),
        y=y_pool[query_idx].reshape(1, )
    )
    # se remueve la instancia consultada del Pool de datos
    X_pool = np.delete(X_pool, query_idx, axis=0)
    y_pool = np.delete(y_pool, query_idx)
    model_accuracy = learner.score(Ex_data, Ey_data)#conjunto de evaluación
     # Guardamos el historial de la predicción.
    performance_history.append(model_accuracy)
    #performance_history = [model_accuracy]
    #print('Precisión después de la consulta no. %d: %f' % (idx+1, acc=model_accuracy))
    print('Precisión después de la consulta {n}: {acc:0.4f}'.format(n=idx + 1, acc=model_accuracy))
```

Figura 4.3: Escenario US proceso de aprendizaje activo.

Para el proceso principal del algoritmo de AL, el cual es el procedimiento de consulta basada en grupos (Pool-based Samplig), se aplicara como estrategia de consulta (Query strategy) el Muestreo por incertidumbre (uncertainty_sampling); el cual realiza la consulta del Pool de datos de uno a uno, es por ello que se decidi optar

realizar un total de 100 consultas (n_queries=100) en las cuales se tomarán los datos sin etiquetar (X_pool) y las etiquetas de prueba (y_pool). Para evaluar este proceso de consulta, se creo un conjunto de datos del corpus completo, donde se seleccionó de manera aleatoria tanto datos etiquetados como aquellos sin etiquetar, equivalente al 50 % de los datos. Siguiendo el mismo orden de ideas, se realiz una segunda

evaluación, pero esta vez se uso solamente el conjunto de datos etiquetados, es decir, se corrió el algoritmo dos veces por cada clasificador.

Durante este proceso de consulta, se buscan las N cantidad de instancias con la mayor informacion en los datos provisto, esto al llamar a la estrategia de consulta (uncertainty_sampling); la cual elige una instancia aleatoria de la cual solicita una

etiqueta. El valor que retorna dicho proceso es, el índice y/o etiqueta de la instancia del grupo elegido para ser etiquetado y la propia instancia; una vez que esta instancia ha sido elegida, se remueve del Pool de datos que son consultados en este proceso y se procesa el calculo de la precision de la prediccion con respecto a los etiquetados, dicha precision se medirá en cada una de las iteraciones de las consultas que realiza el algoritmo, ya que en cada iteracion se nos arroja el puntaje maximo de la evaluación que realiza la métrica del clasificador usado como estimador.

En el segundo escenario de experimentacion corresponde al Ranked Batch-mode Sampling (RBMS), en dicho escenario se repiti la misma segmentacion de los datos que en el escenarios anterior (ver figura 4.2). Una vez segmentados los datos, procedemos a establecer el proceso de aprendizaje, que para este algoritmo cambia un poco la estructura con respecto al anterior; primeramente estableceremos en una variable a nuestros clasificadores (K-nn y Random Forest), que tendrán la funcion de estimador, recordemos que, como en el escenario anterior, esto nos dar pauta para correr el algoritmo de manera independiente con cada uno de los clasificadores. Posteriormente, estableceremos el tamaño el batch (lote) para que este tome 20 muestras de datos por vez (BATCH_SIZE); así mismo, establecemos la estrategia de consulta uncertainty_batch_sampling dentro de la variable preset_batch. Después, estos parametros serán pasados a la variable learner, donde se llevar a cabo el proceso de aprendizaje en segundo plano.

Para el procedimiento de consulta, se aplicar como estrategia de consulta Muestreo de incertidumbre por lotes (uncertainty_batch_sampling); la cual evaluará lotes con 20 muestras por vez, ampliando el muestreo de consulta por incertidumbre, asignándole una calificación a cada una de muestras del lote, es decir, identifica que muestra del lote es mas importante y procede a su etiquetado. Como podemos observar en la figura 4.4, que nos ilustra a manera de ejemplo un fragmento de codigo.

```
# Estableciendo el clasificador que sera el estimador
knn = KNeighborsClassifier(n_neighbors=4)

# Estableciendo el tamaño del batch 20 samples por vez
BATCH_SIZE = 1
preset_batch = partial(uncertainty_batch_sampling, n_instances=BATCH_SIZE)

# Iniciando el model de active learning
learner = ActiveLearner(
    estimator=knn,
    X_training=X_train,
    y_training=y_train,
    query_strategy=preset_batch
)
# Aislando los datos para ser graficados
predictions = learner.predict(Xdata)
is_correct = (predictions == ydata)

# Obteniendo el score de los datos evaluados
unqueried_score = learner.score(Xdata, ydata)

# Pool-based sampling
N_RAW_SAMPLES = 20
N_QUERIES = N_RAW_SAMPLES // BATCH_SIZE

performance_history = [unqueried_score]
```

Figura 4.4: Escenario RBMS proceso de aprendizaje activo.

A continuacion, se describir con mayor detalle los resultados obtenidos de este proceso de etiquetado que tiene como finalidad identificar el ”Proceso cerebral del Lenguaje”.

Análisis estadístico de la fase experimental con la técnica de Aprendizaje Activo (Active Learning)

En la fase experimental de este trabajo de investigación, se desarrollaron un total de 32 experimentos entre ambos escenarios de experimentación que se explicaron con anterioridad; los cuales se dividen en 16 experimentos para el escenario de experimentacion Uncertainty Sampling y 16 experimentos para el escenario de experimentacion Ranked Batch-Mode Sampling, en la figura 4.5 podemos observar como esta estructurado este proceso experimental.

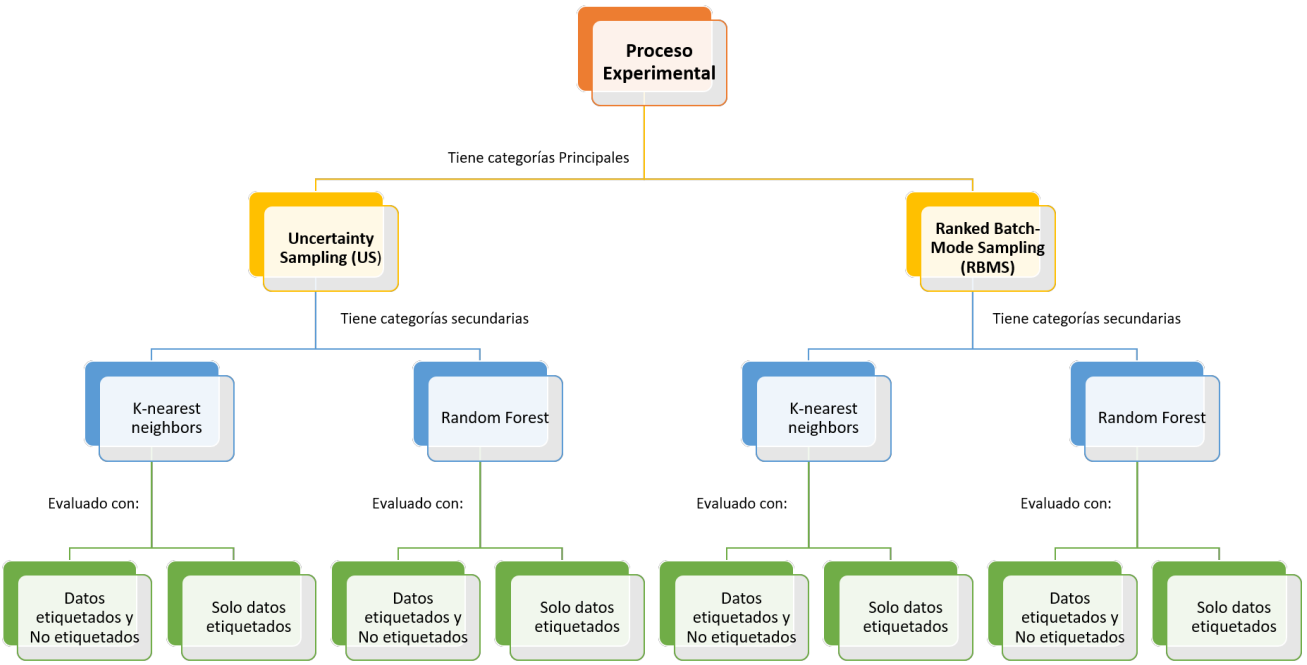

Figura 4.5: Estructura del proceso experimental.

A continuacion, detallaremos los resultados de la fase experimental de cada uno de los escenarios antes mencionados.

Escenario de experimentación:Uncertainty Sampling (US)

En primer lugar describiremos el analisis estadístico del escenario de experimentacion Uncertainty Sampling (US), el cual se encuentra dividido en dos categorías. La primer categoría que lleva por nombre K-nearest neighbors y emplea las métricas usadas por el clasificador K-nearest neighbors (K-nn) para realizar el cálculo de la estimación del proceso de etiquetado, es importante recordar que, según la literatura del Aprendizaje Activo (Active Learning), se puede hacer referencia a los clasificadores como ”Estimador o estimadores” y que son usados para el cálculo de la estimacion de la etiqueta durante el proceso del algoritmo. Dentro de esta categoría se someti a evaluacion el desempeño de este proceso de Aprendizaje Activo (Active Learnig) tomando en cuenta dos criterios; el primero de ellos realiza la evaluación al aplicar el conjunto de datos etiquetado y el conjunto de datos NO etiquetados, es decir, el set de datos completo. La segunda evaluación se realiza al aplicar solo el segmento de datos etiquetados; por lo tanto tendremos dos tablas con el análisis estadístico por cada uno de los conjuntos de datos que se aplicaron para la evaluacion del proceso del algoritmo de Aprendizaje Activo (Active Learning).

En este escenario de experimentacion, se realizaron corridas con 100 iteraciones donde con cada iteración se ejecutaba la estrategia de consulta Muestreo por incertidumbre (uncertainty_sampling), la cual nos devuelve un porcentaje de precision o ”Accuracy” por cada una de la iteraciones y que ser empleado para realizar el análisis estadístico correspondiente.En la primer corrida de cuatro experimentos, donde se evalu el desempeño con los datos etiquetados y No etiquetados, es decir, con el set de datos completo podemos observar que el primer experimento obtuvo el mejor desempeño con un puntaje del 27.15 % como se muestra en la tabla 4.1 y en la grafica de la figura 4.6 podemos observa con detalle los porcentajes de precisión de cada iteracion del primer experimento donde se obtuvo el mayor desempeño.

Escenario de Experimentacion: Uncertanty Sampling							
Estimador: K-nn=4							
Evaluación con respecto a los datos etiquetados y NO etiquetados							
Experimento 1		Experimento 2		Experimento 3		Experimento 4	
Media	0.257991	Media	0.212472	Media	0.254222	Media	0.207429
Error típico	0.00099156	Error típico	0.00144162	Error típico	0.00064489	Error típico	0.00060398
Mediana	0.2611	Mediana	0.2193	Mediana	0.2559	Mediana	0.2063
Moda	0.2637	Moda	0.2245	Moda	0.2559	Moda	0.2063
Desviación estándar	0.0099156	Desviación estándar	0.01441619	Desviación estándar	0.00644892	Desviación estándar	0.00603984
Varianza de la muestra	9.8319E-05	Varianza de la muestra	0.00020783	Varianza de la muestra	4.1589E-05	Varianza de la muestra	3.648E-05
Curtosis	1.8756389	Curtosis	-0.83440975	Curtosis	13.9993285	Curtosis	2.21766486
Coeficiente de asimetría	-1.43824398	Coeficiente de asimetría	-0.81339137	Coeficiente de asimetría	-3.15654608	Coeficiente de asimetría	0.43519112
Rango	0.0443	Rango	0.047	Rango	0.047	Rango	0.0391
Mínimo	0.2272	Mínimo	0.1828	Mínimo	0.2167	Mínimo	0.1854
Máximo	0.2715	Máximo	0.2298	Máximo	0.2637	Máximo	0.2245
Suma	25.7991	Suma	21.2472	Suma	25.4222	Suma	20.7429
Cuenta	100	Cuenta	100	Cuenta	100	Cuenta	100

Tabla 4.1: Analisis estadístico del experimento US con el estimador K-nn, evaluado con respecto a los datos etiquetados y NO etiquetados.

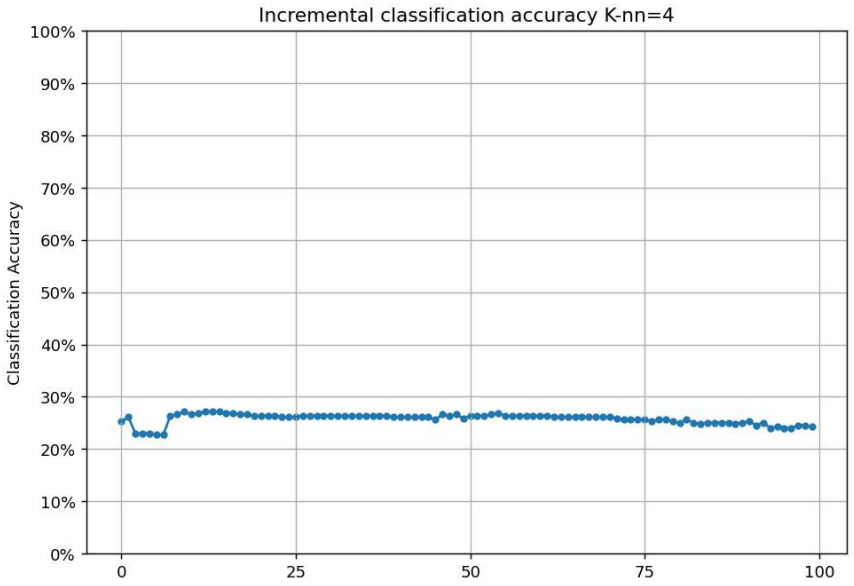

Figura 4.6: Grafica del Experimento 1 donde se obtiene el puntaje maximo del proceso de etiquetado en el Escenario US con el estimador K-nn, evaluado con respecto al conjunto de datos etiquetados y NO etiquetados.

En la segunda corrida de cuatro experimentos, donde se evaluó el desempeño solo con el conjunto de datos etiquetados, es decir el 50 % del set de datos, se obtuvo un puntaje maximo del 57.18 % que corresponde al primer experimento de la corrida como se muestra en la tabla 4.2.

Escenario de Experimentacion: Uncertanty Sampling							
Estimador: K-nn=4							
Evaluación con respecto a los datos etiquetados							
Experimento 1		Experimento 2		Experimento 3		Experimento 4	
Media	0.5718	Media	0.423	Media	0.448632	Media	0.436
Error típico	1.339E-16	Error típico	2.2316E-17	Error típico	0.00010039	Error típico	0
Mediana	0.5718	Mediana	0.423	Mediana	0.4491	Mediana	0.436
Moda	0.5718	Moda	0.423	Moda	0.4491	Moda	0.436
Desviación estándar	1.339E-15	Desviación estándar	2.2316E-16	Desviación estándar	0.00100392	Desviación estándar	0
Varianza de la muestra	1.7929E-30	Varianza de la muestra	4.9802E-32	Varianza de la muestra	1.0079E-06	Varianza de la muestra	0
Curtosis	-2.04123711	Curtosis	-2.04123711	Curtosis	0.87775115	Curtosis	#¡DIV/0!
Coeficiente de asimetría	-1.0152933	Coeficiente de asimetría	1.0152933	Coeficiente de asimetría	-1.69132986	Coeficiente de asimetría	#¡DIV/0!
Rango	0	Rango	0	Rango	0.0026	Rango	0
Mínimo	0.5718	Mínimo	0.423	Mínimo	0.4465	Mínimo	0.436
Máximo	0.5718	Máximo	0.423	Máximo	0.4491	Máximo	0.436
Suma	57.18	Suma	42.3	Suma	44.8632	Suma	43.6
No. Consultas	100	Cuenta	100	Cuenta	100	Cuenta	100

Tabla 4.2: Analisis estadístico del experimento US con el estimador K-nn, evaluado con respecto a los datos etiquetados.

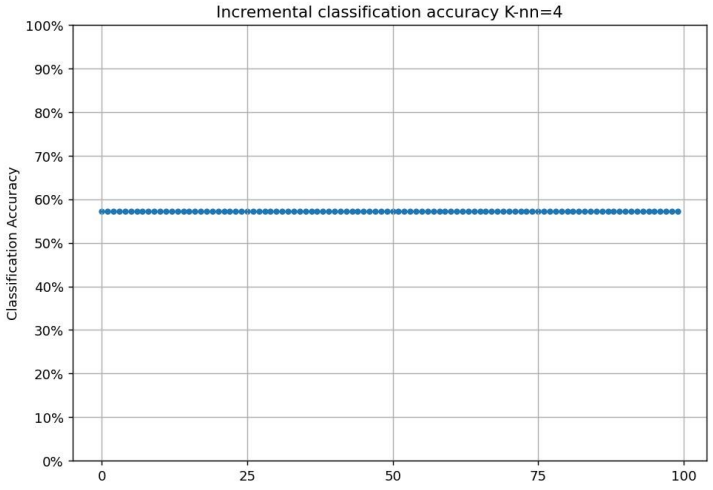

Figura 4.7: Grafica del Experimento 1 donde se obtiene el puntaje maximo del proceso de etiquetado en el Escenario US con el estimador K-nn, evaluado con respecto al conjunto de datos etiquetados.

La segunda categorías del escenario de experimentación Uncertanty Sampling lleva por nombre Random Forest y emplea las métricas del clasificador Random Forest para realizar el calculo de la estimación del proceso de etiquetado. En la primer corrida de cuatro experimentos de esta categoría, se evaluó el desempeño del proceso de etiquetado con respecto a los datos etiquetado y NO etiquetados, es decir con el set de datos completo. Se obtuvo un puntaje máximo del 33.42 % correspondiente al segundo experimento de la corrida, como se observa en la tabla 4.3.

Escenario de Experimentacion: Uncertanty Sampling							
Estimador: Ramndom Forest							
Evaluación con respecto a los datos etiquetados y NO etiquetados							
Experimento 1		Experimento 2		Experimento 3		Experimento 4	
Media	0.283947	Media	0.311858	Media	0.256603	Media	0.25402
Error típico	0.00135809	Error típico	0.00091735	Error típico	0.00123434	Error típico	0.00135081
Mediana	0.2846	Mediana	0.3107	Mediana	0.2585	Mediana	0.2533
Moda	0.2898	Moda	0.3081	Moda	0.2585	Moda	0.2533
Desviación estándar	0.01358091	Desviación estándar	0.00917353	Desviación estándar	0.01234337	Desviación estándar	0.01350808
Varianza de la muestra	0.00018444	Varianza de la muestra	8.4154E-05	Varianza de la muestra	0.00015236	Varianza de la muestra	0.00018247
Curtosis	-0.56449944	Curtosis	-0.28279978	Curtosis	0.4223632	Curtosis	0.63508027
Coeficiente de asimetría	-0.07228938	Coeficiente de asimetría	0.26474064	Coeficiente de asimetría	-0.00309694	Coeficiente de asimetría	-0.63132293
Rango	0.0574	Rango	0.0418	Rango	0.0705	Rango	0.0731
Mínimo	0.2559	Mínimo	0.2924	Mínimo	0.2219	Mínimo	0.2115
Máximo	0.3133	Máximo	0.3342	Máximo	0.2924	Máximo	0.2846
Suma	28.3947	Suma	31.1858	Suma	25.6603	Suma	25.402
Cuenta	100	Cuenta	100	Cuenta	100	Cuenta	100

Tabla 4.3: Analisis estadístico del experimento US con el estimador Random Forest, evaluado con respecto a los datos etiquetados y NO etiquetados.

En al figura 4.8 podemos observar con mayor detalle los porcentajes de precisión en cada una de las iteraciones que corresponden al experimento 2, el cual tiene el puntaje máximo de desempeño en el proceso de etiquetado.

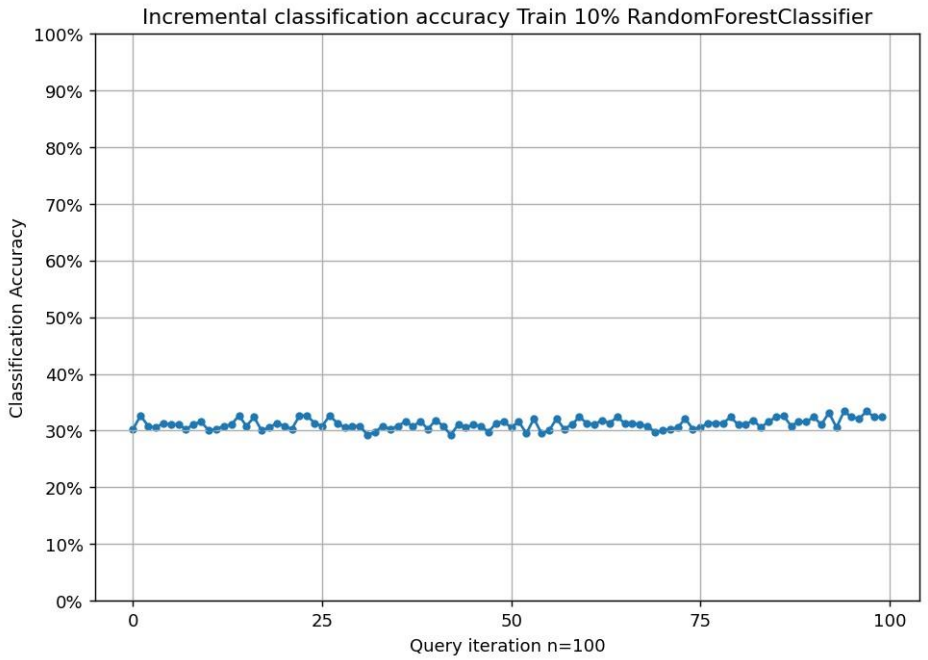

Figura 4.8: Grafica del Experimento 2 donde se obtiene el puntaje maximo del proceso de etiquetado en el Escenario US con el estimador Random Forest, evaluado con respecto al conjunto datos etiquetados y NO etiquetados.

En la segunda corrida de cuatro experimentos de la categoría Ramdom Forest, se evalu el desempeño con el segmento de datos etiquetados, es decir el 50 % de los datos y se obtuvo un puntaje maximo del 69.71 % de precision correspondiente al experimento 4, como se muestra en la tabla 4.4. Mientras que en la figura 4.9 podemos observar el porcentaje de precisión de cada iteracion correspondiente al experimento 4, el cual obtuvo el mayo puntaje de desempeño.

Escenario de Experimentacion: Uncertanty Sampling							
Estimador: Ramndom Forest							
Evaluación con respecto a los datos etiquetados							
Experimento 1		Experimento 2		Experimento 3		Experimento 4	
Media	0.657834	Media	0.61068	Media	0.585032	Media	0.652193
Error típico	0.00121657	Error típico	0.0018916	Error típico	0.00250194	Error típico	0.00197696
Mediana	0.658	Mediana	0.6084	Mediana	0.5927	Mediana	0.6501
Moda	0.6632	Moda	0.6031	Moda	0.5953	Moda	0.6475
Desviación estándar	0.01216568	Desviación estándar	0.01891595	Desviación estándar	0.02501944	Desviación estándar	0.01976957
Varianza de la muestra	0.000148	Varianza de la muestra	0.00035781	Varianza de la muestra	0.00062597	Varianza de la muestra	0.00039084
Curtosis	0.00282917	Curtosis	-0.05945445	Curtosis	1.31835701	Curtosis	-0.22012455
Coeficiente de asimetría	-0.15746949	Coeficiente de asimetría	0.532293	Coeficiente de asimetría	-1.18393854	Coeficiente de asimetría	-0.07341851
Rango	0.0601	Rango	0.0888	Rango	0.1253	Rango	0.094
Mínimo	0.6266	Mínimo	0.5744	Mínimo	0.5013	Mínimo	0.6031
Máximo	0.6867	Máximo	0.6632	Máximo	0.6266	Máximo	0.6971
Suma	65.7834	Suma	61.068	Suma	58.5032	Suma	65.2193
Cuenta	100	Cuenta	100	Cuenta	100	Cuenta	100

Tabla 4.4: Analisis estadístico del experimento US con el estimador Random Forest, evaluado con respecto a los datos etiquetados.

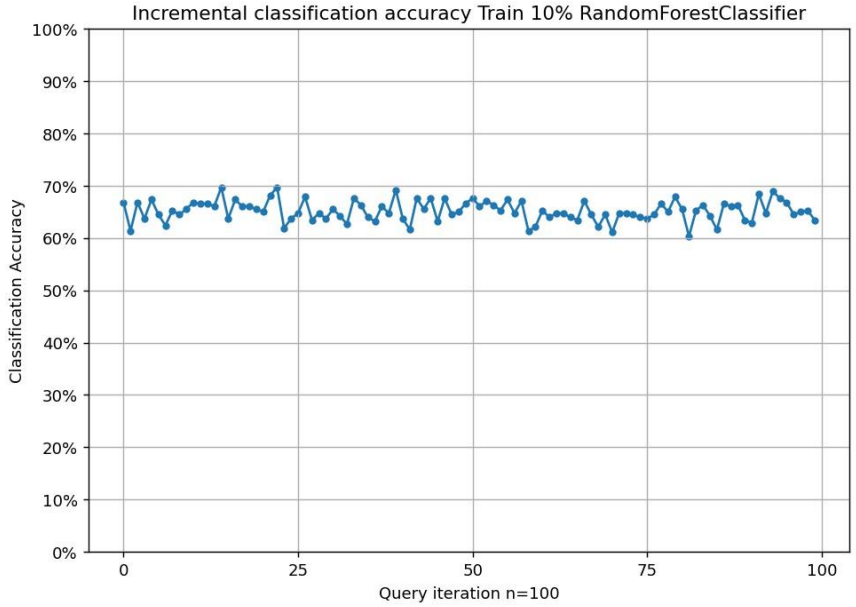

Figura 4.9: Grafica del Experimento 4 donde se obtiene el puntaje maximo del proceso de etiquetado en el Escenario US con el estimador Random Forest, evaluado con respecto al conjunto de datos etiquetados.

Escenario de experimentación:Ranked Batch-Mode Sampling (RBMS)

Enseguida, detallaremos en analisis estadístico del escenario de experimentacion Ranked Batch-Mode Sampling (RBMS), que como en el escenario anterior, también se encuentra dividido en dos categorías y que se pueden identificar por el mismo nombre. La primera categoría también lleva por nombre K-nearest neighbors y emplea las métricas del clasificador K-nearest neighbors (K-nn) para realizar el calculo de la estimación del proceso de etiquetado.

A diferencia del escenario de experimentacion anterior, en el escenario de experimentacion RBMS se ejecuto la estrategia de consulta Muestreo de incertidumbre por lotes (uncertainty_batch_sampling) a manera de lotes o "Batch", es decir, se tomaran 20 muestras de datos por vez para ser analizadas. Así mismo, este escenario de experimentacion se someti a dos evaluaciones de desempeño; la primera de ellas se tomo en cuenta los datos etiquetados y No etiquetados, es decir el set de datos completo para evaluar el desempeño. Posteriormente, en la segunda evaluación solo se tomaron los datos etiquetados, es decir el 50 % del set de datos. En la primer corrida de cuatro experimentos, donde se evalu con el set de datos completo (datos etiquetados y No etiquetados) se obtuvo un puntaje maximo del 30.55 % que corresponde al primer experimento, como se observa en la tabla 4.6. Mientras que en la figura 4.10, podemos observar con mayor detalle el desempeño del experimento 1.

Escenario de Experimentacion: Ranked bartch-mode Samplig							
Estimador: K-nn=4							
Evaluación con respecto a los datos etiquetados y NO etiquetados							
Experimento 1		Experimento 2		Experimento 3		Experimento 4	
Media	0.29046	Media	0.25915	Media	0.22229	Media	0.18054
Error típico	0.00121134	Error típico	0.00062203	Error típico	0.0002845	Error típico	0.0008791
Mediana	0.2898	Mediana	0.2598	Mediana	0.2219	Mediana	0.17885
Moda	0.2924	Moda	0.2611	Moda	0.2219	Moda	0.1775
Desviación estándar	0.00541726	Desviación estándar	0.0027818	Desviación estándar	0.00127234	Desviación estándar	0.00393144
Varianza de la muestra	2.9347E-05	Varianza de la muestra	7.7384E-06	Varianza de la muestra	1.6188E-06	Varianza de la muestra	1.5456E-05
Curtosis	2.09419166	Curtosis	-1.42286261	Curtosis	1.30376411	Curtosis	-0.28685789
Coeficiente de asimetría	1.25362841	Coeficiente de asimetría	0.0179051	Coeficiente de asimetría	0.44163193	Coeficiente de asimetría	1.06079834
Rango	0.0209	Rango	0.0078	Rango	0.0052	Rango	0.0105
Mínimo	0.2846	Mínimo	0.2559	Mínimo	0.2193	Mínimo	0.1775
Máximo	0.3055	Máximo	0.2637	Máximo	0.2245	Máximo	0.188
Suma	5.8092	Suma	5.183	Suma	4.4458	Suma	3.6108
Cuenta	20	Cuenta	20	Cuenta	20	Cuenta	20

Tabla 4.5: Analisis estadístico del experimento RBMS con el estimador K-nn, evaluado con respecto a los datos etiquetados y NO etiquetados.

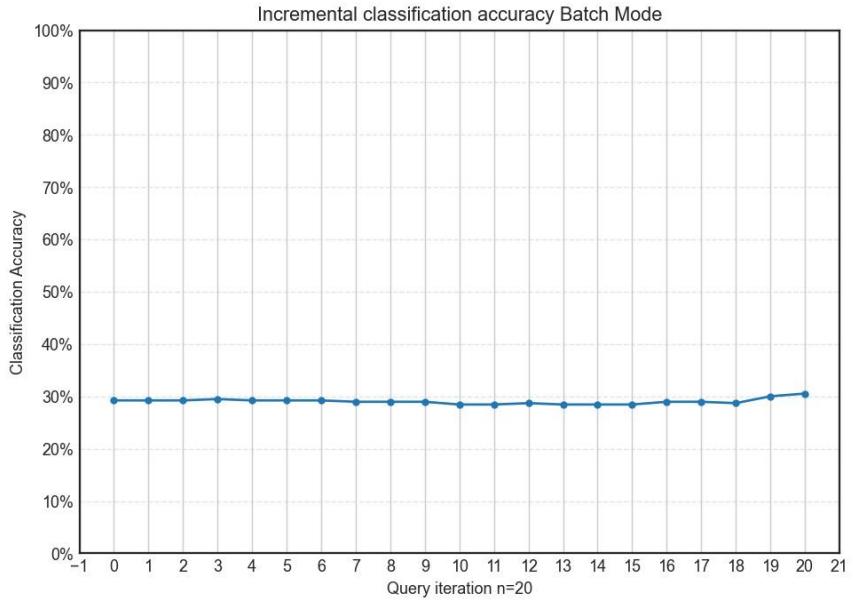

Figura 4.10: Grafica del Experimento 1 donde se obtiene el puntaje maximo del proceso de etiquetado en el escenario RBMS con el estimador K-nn, evaluado con respecto al conjunto de datos etiquetados y NO etiquetados.

En la segunda corrida de cuatro experimentos de la categoría K-nearest neighbors,

del escenario de experimentacion RBMS, donde se evaluó el desempeño del proceso de etiquetado con solo el 50 % de los datos que corresponde al segmento de datos etiquetados, se obtuvo un porcentaje maximo del 61.62 % correspondiente al primer experimeto de esta corrido como lo podemos observar en la tabla 4.6 y en la figura 4.11.

Escenario de Experimentacion: Ranked bartch-mode Samplig							
Estimador: K-nn=4							
Evaluación con respecto a los datos etiquetados							
Experimento 1		Experimento 2		Experimento 3		Experimento 4	
Media	0.6162	Media	0.5587	Media	0.4804	Media	0.4909
Error típico	5.0941E-17	Error típico	0	Error típico	1.2735E-17	Error típico	0
Mediana	0.6162	Mediana	0.5587	Mediana	0.4804	Mediana	0.4909
Moda	0.6162	Moda	0.5587	Moda	0.4804	Moda	0.4909
Desviación estándar	2.2781E-16	Desviación estándar	0	Desviación estándar	5.6953E-17	Desviación estándar	0
Varianza de la muestra	5.1899E-32	Varianza de la muestra	0	Varianza de la muestra	3.2437E-33	Varianza de la muestra	0
Curtosis	-2.23529412	Curtosis	#¡DIV/0!	Curtosis	-2.23529412	Curtosis	#¡DIV/0!
Coeficiente de asimetría	1.08297715	Coeficiente de asimetría	#¡DIV/0!	Coeficiente de asimetría	-1.08297715	Coeficiente de asimetría	#¡DIV/0!
Rango	0	Rango	0	Rango	0	Rango	0
Mínimo	0.6162	Mínimo	0.5587	Mínimo	0.4804	Mínimo	0.4909
Máximo	0.6162	Máximo	0.5587	Máximo	0.4804	Máximo	0.4909
Suma	12.324	Suma	11.174	Suma	9.608	Suma	9.818
Cuenta	20	Cuenta	20	Cuenta	20	Cuenta	20

Tabla 4.6: Analisis estadístico del experimento RBMS con el estimador K-nn, evaluado con respecto a los datos etiquetados.

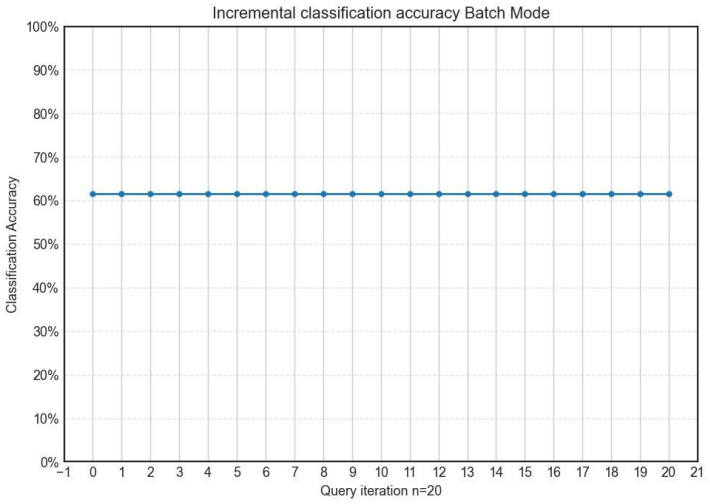

Figura 4.11: Grafica del Experimento 1 donde se obtiene el puntaje maximo del proceso de etiquetado en el escenario RBMS con el estimador K-nn, evaluado con respecto al conjunto de datos etiquetados.

La segunda categoría del escenario de experimentacion RBMS lleva por nombre Random Fores, la cual emplea las métricas del clasificador Random Forest para realizar el cálculo de la estimación del proceso de etiquetado. Al igual que la categoría anterior, se empleara el Muestreo de incertidumbre por lotes (uncertainty_batch_sampling) a manera de lotes o ”Batch”, es decir, se tomaran 20 muestras de datos por vez para ser analizadas. En la primer corrida de cuatro experimentos, donde se evaluó el desempeño del proceso de etiquetado el set de datos completo, es decirlos datos etiquetados y NO etiquetados, se obtuvo un puntaje maximo de 30.29 % correspondiente al experimento 3 de esta corrida; como lo podemos observar en la tabla 4.7, mientras que en la figura 4.12 podemos observar con mayor detalle el comportamiento del experimento 3.

Escenario de Experimentacion: Ranked bartch-mode Samplig							
Estimador: Random Forest							
Evaluación con respecto a los datos etiquetados y NO etiquetados							
Experimento 1		Experimento 2		Experimento 3		Experimento 4	
Media	0.278345	Media	0.252885	Media	0.2752	Media	0.254835
Error típico	0.00295459	Error típico	0.00298275	Error típico	0.00293238	Error típico	0.00259961
Mediana	0.2794	Mediana	0.2559	Mediana	0.2768	Mediana	0.2559
Moda	0.2768	Moda	0.2559	Moda	0.2663	Moda	0.2559
Desviación estándar	0.01321333	Desviación estándar	0.01333927	Desviación estándar	0.01311399	Desviación estándar	0.0116258
Varianza de la muestra	0.00017459	Varianza de la muestra	0.00017794	Varianza de la muestra	0.00017198	Varianza de la muestra	0.00013516
Curtosis	-0.14996974	Curtosis	-0.67714561	Curtosis	0.29389025	Curtosis	0.67487619
Coeficiente de asimetría	-0.44167789	Coeficiente de asimetría	-0.21745303	Coeficiente de asimetría	-0.07279326	Coeficiente de asimetría	0.26593704
Rango	0.0496	Rango	0.047	Rango	0.0549	Rango	0.0496
Mínimo	0.2507	Mínimo	0.2272	Mínimo	0.248	Mínimo	0.2324
Máximo	0.3003	Máximo	0.2742	Máximo	0.3029	Máximo	0.282
Suma	5.5669	Suma	5.0577	Suma	5.504	Suma	5.0967
Cuenta	20	Cuenta	20	Cuenta	20	Cuenta	20

Tabla 4.7: Analisis estadístico del experimento RBMS con el estimador Random Forest, evaluado con respecto a los datos etiquetados y NO etiquetados.

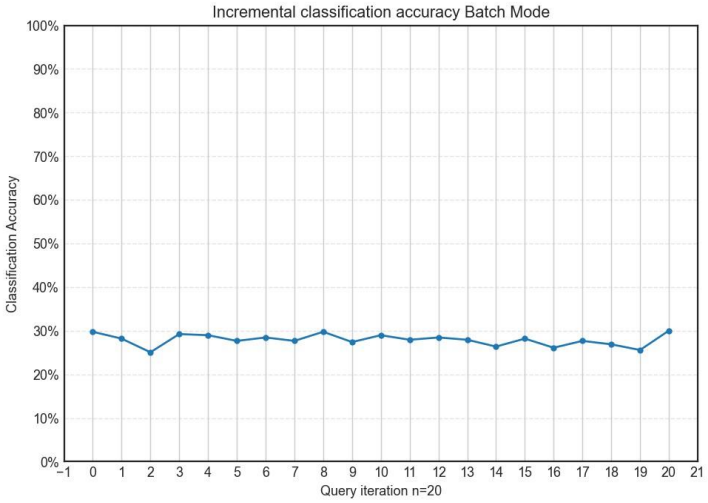

Figura 4.12: Grafica del Experimento 3 donde se obtiene el puntaje maximo del proceso etiquetado en el escenario RBMS con el estimador Random Forest, evaluado con respecto al conjunto de datos etiquetados y NO etiquetados.

En la segunda corrida de cuatro experimentos del al categoría Random Fores, donde se evaluó con el 50 % de los datos, es decir el segmento de los datos etiquetados. Se obtuvo un puntaje máximo del 70.5 % de precision correspondiente al segundo experimento de esta corrida, como se muestra en la tabla 4.8.

Escenario de Experimentacion: Ranked bartch-mode Samplig							
Estimador: Random Forest							
Evaluación con respecto a los datos etiquetados							
Experimento 1		Experimento 2		Experimento 3		Experimento 4	
Media	0.65458	Media	0.669335	Media	0.58407	Media	0.63525
Error típico	0.00293065	Error típico	0.00376806	Error típico	0.00364814	Error típico	0.00325333
Mediana	0.6554	Mediana	0.6723	Mediana	0.5796	Mediana	0.6358
Moda	0.6632	Moda	0.6554	Moda	0.5666	Moda	0.6475
Desviación estándar	0.01310627	Desviación estándar	0.01685128	Desviación estándar	0.01631499	Desviación estándar	0.01454932
Varianza de la muestra	0.00017177	Varianza de la muestra	0.00028397	Varianza de la muestra	0.00026618	Varianza de la muestra	0.00021168
Curtosis	-0.2322128	Curtosis	-0.50951679	Curtosis	-0.82292387	Curtosis	-1.07827881
Coeficiente de asimetría	-0.36222709	Coeficiente de asimetría	0.06127677	Coeficiente de asimetría	0.59528488	Coeficiente de asimetría	-0.18943354
Rango	0.0523	Rango	0.0653	Rango	0.0496	Rango	0.047
Mínimo	0.6266	Mínimo	0.6397	Mínimo	0.5666	Mínimo	0.611
Máximo	0.6789	Máximo	0.705	Máximo	0.6162	Máximo	0.658
Suma	13.0916	Suma	13.3867	Suma	11.6814	Suma	12.705
Cuenta	20	Cuenta	20	Cuenta	20	Cuenta	20

Tabla 4.8: Analisis estadístico del experimento RBMS con el estimador Random Forestt, evaluado con respecto a los datos etiquetados.

Mientras que en la figura 4.13, podemos observar con mayor detalle el desempeño del experimento dos el cual obtuvo el mayo puntaje de la corrida.

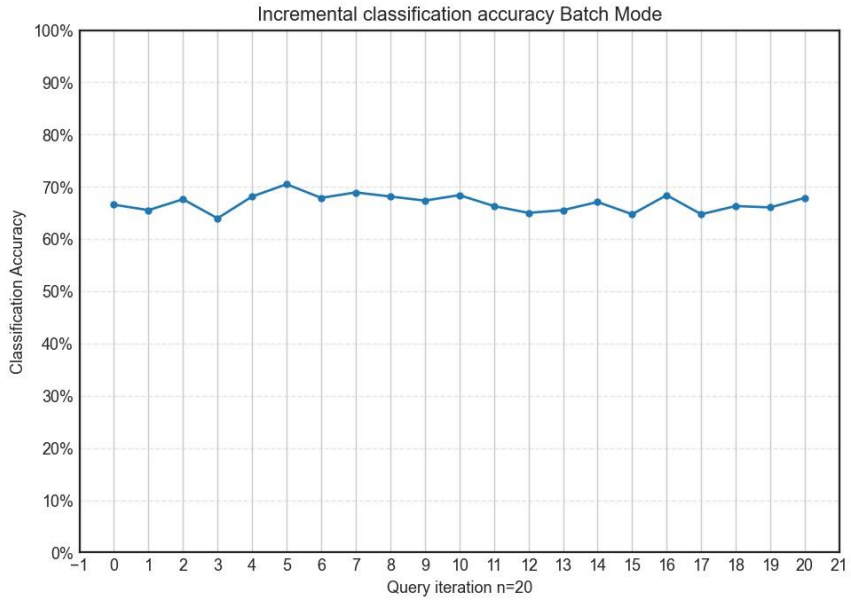

Figura 4.13: Grafica del Experimento 2 donde se obtiene el puntaje maximo del proceso etiquetado en el escenario RBMS con el estimador Random Forest, evaluado con respecto al conjunto de datos etiquetados.

A continuacion, describiremos con mayor detalle los resultados obtenidos del proceso de clasificación, posterior al proceso de etiquetado con el Aprendizaje Activo (Active Learning)

Proceso de Clasificación

Habiendo terminado el análisis estadístico del proceso de Aprendizaje Activo (Active Learning), se seleccionar el experimento con el mejor desempeño de cada uno

de los escenarios de experimentacion (Uncertainty Sampling y Ranked Batch-Mode Sampling), de dicho experimento seleccionado como aquel de mejor desempeño se le sometió a la técnica de Analisis de Componentes Principales (PCA, por siglas en Inglés), con la finalidad de reducir la dimensión del set de datos y tener un mejor rendimiento computacional. Cabe destacar que, en la aplicacion del proceso de reduccion de la dimension del set de datos con la técnica de PCA, se uso la uso la matriz completa donde se tiene los datos previamente etiquetados (ver tabla 3.2), como se explica en el tercer capítulo en la seccion "Proceso de clasificación" y en conjunto con los datos del resultado del proceso de etiquetado con el Aprendizaje Activo (Active Learning). Es importante mencionar que las clases previamente determinadas (ver tabla 3.2), seran identificadas en las distintas matrices de confusion con los numeros 0,1,2 y 3 como se muestra en la tabla 4.9.

Identificadores de las calses		
Clase	Identificador de la clase	Identificador en la matriz de confusión
Procesamiento de frases	1	0
Retencion de palabras	2	1
Comprension del léxico	3	2
Ordenamiento de palabras	4	3

Tabla 4.9: Identificadores de las clases, que serán usadas en le proceso de calsificacion.

Enseguida, detallaremos los resultados obtenidos en el proceso de clasificacion por cada experimento con el mejor desempeño del proceso de Aprendizaje Activo (Active Learning) para cada escenario de experimentación aplicado.

Resultados del proceso de clasificacion en el escenario Uncertainty Sampling(US).

Como se menciono con anterioridad, dentro de los 16 experimentos que se realizaron en este escenario de experimentacion se selecciono aquel con el mejor desempeño, dicho experimento corresponde al efectuado con el estimador Random Forest y el cual fue evaluado con respecto a los datos etiquetados previamente, como se muestra en la figura 4.18. De este experimento se tomaron los resultados del proceso de etiquetado con el Aprendizaje Activo (Active Learning), el cual se uni con el segmento de datos previamente etiquetados para formar una matriz completa, como se explica en el tercer capítulo en la seccion "Proceso de clasificación"; posteriormente se someti esta matriz de datos a la técnica de PCA para la reduccion de su dimensión.

Habiendo dicho lo anterior, en esta reducción de la dimension se obtuvo una matriz de datos con una nueva dimension de 690x21, siendo la ultima columna la que almacena las etiquetas o "targets" que seran útiles para el proceso de clasificacion como se muestra en la tabla 4.9. Esta nueva matriz de datos fue sometida a los siguientes clasificadores:

- Gaussian Naive Bayes(GNB).
- K-nearest neighbor(KNN).
- Multilayer Perceptron (MLP)
- Support Vector Machine(SVM).

A continuacion, detallaremos los resultados obtenidos durante este proceso de clasificacion; comenzaremos por el clasificador Gaussian Naive Bayes(GNB), en el cual segmentamos el set de datos en 80 % para el entrenamiento y el 20 % restante lo empleamos para la prueba o ”test” de este algoritmo de clasificacion, de dicho algoritmo se realizaron varias corridas de prueba en las cuales se obtuvo como mejor porcentaje de precision o ”Accuracy” del 51 %, como se muestra en la tabla 4.10.

	precision	recall	f1-score	support
1.0	0.65	0.31	0.42	42
2.0	0.57	0.59	0.58	39
3.0	0.41	0.87	0.56	30
4.0	0.53	0.30	0.38	27
accuracy			0.51	138
macro avg	0.54	0.52	0.49	138
weighted avg	0.55	0.51	0.49	138

Tabla 4.10: Reporte de precision del algoritmo de clasificacion Gaussian Naive Bayes(GNB).

Así mismo en la figura 4.14, podemos observar la matriz de confusión del proceso de clasificación del algoritmo GNB, donde se muestran los porcentajes de presion de cada una de las cuatro clases con las que cuenta el set de datos.

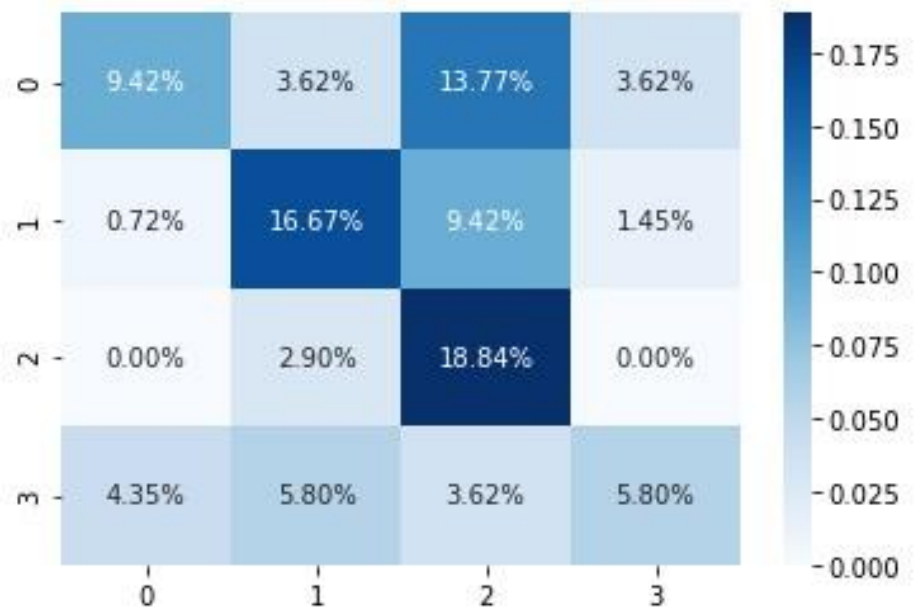

Figura 4.14: Matriz de confusión del algoritmo de clasificacion Gaussian Naive Bayes(GNB).

El siguiente algoritmo de clasificación corresponde a K-nearest neighbor(KNN), de igual manera en este proceso se segmentaron los datos los datos en 80 % para el entrenamiento y el 20 % restante lo empleamos para la prueba o ”test” y donde se evalu de manera iterativa el numero de ”vecinos” o K en un rango de 1 a 25,

esto con la finalidad de analizar un numero considerable de vecinos para observar su desempeño; donde el mayor porcentaje de precisión que se obtuvo fue del 46 % como se muestra en la tabla 4.11.

	precision	recall	f1-score	support
1.0	0.47	0.67	0.55	43
2.0	0.43	0.48	0.45	42
3.0	0.79	0.44	0.56	25
4.0	0.27	0.14	0.19	28
accuracy			0.46	138
macro avg	0.49	0.43	0.44	138
weighted avg	0.47	0.46	0.45	138

Tabla 4.11: Reporte de precision del algoritmo de clasificación K-nearest neighbor(KNN).

Así mismo, durante este proceso de clasificación se obtuvo la matriz de confusión del ”vecino” con el mejor desempeño, donde se observa los porcentajes de precision de cada una de las clases, como se muestra en la figura 4.15.

Figura 4.15: Matriz de confusion del algoritmo de clasificacion K-nearest neighbor(KNN).

Siguiendo con el proceso de clasificación del escenario de experimentación Uncertainty Sampling, continuamos con el clasificador Multilayer Perceptron (MLP), que de igual manera se hizo una segmentacion de datos en 80 % para el entrenamiento y el 20 % restante lo empleamos para la prueba o ”test” y se establecieron 3 capas ocultas para este algoritmo, de las diferentes corridas que se ejecutaron se obtuvo un porcentaje de precisión del 46 %, como se muestra en la tabla 4.12.

	precision	recall	f1-score	support
1.0	0.64	0.65	0.65	52
2.0	0.39	0.60	0.47	42
3.0	0.00	0.00	0.00	17
4.0	0.25	0.19	0.21	27
accuracy			0.46	138
macro avg	0.32	0.36	0.33	138
weighted avg	0.41	0.46	0.43	138

Tabla 4.12: Reporte de precision del algoritmo de clasificación Multilayer Perceptron (MLP).

Donde se obtuvo la matriz de confusion, como se muestra en la figura 4.16, que nos indica el porcentaje de precisión por cada una de las clases.

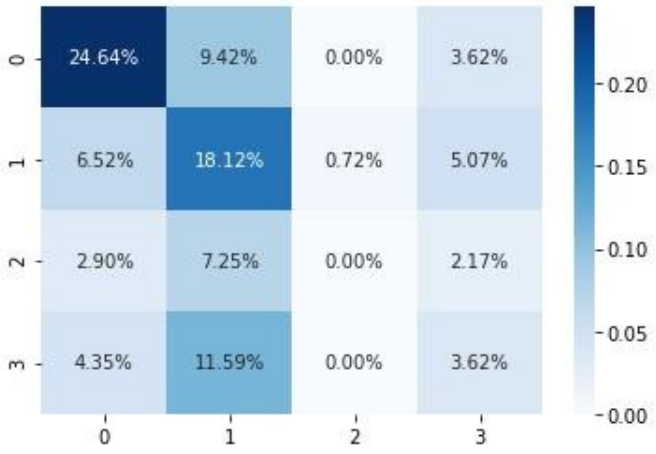

Figura 4.16: Matriz de confusión del algoritmo de clasificacion Multilayer Perceptron (MLP)).

Para terminar con el proceso de clasificacion del escenario de experimentacion Uncertainty Sampling, se aplico el algoritmo de clasificacion Support Vector Machine(SVM) en el cual se segmento el set de datos en 80 % para el entrenamiento y el 20 % restante lo empleamos para la prueba o ”test” y en las diferentes corridas realizadas se obtuvo un mayor porcentaje de precisión del 43 % como se muestra en la tabla 4.13.

	precision	recall	f1-score	support
1.0	0.62	0.54	0.58	39
2.0	0.38	0.78	0.51	45
3.0	0.10	0.05	0.07	20
4.0	1.00	0.06	0.11	34
accuracy			0.43	138
macro avg	0.52	0.36	0.32	138
weighted avg	0.56	0.43	0.37	138

Tabla 4.13: Reporte de precision del algoritmo de clasificación Support Vector Machine(SVM).

En la figura 4.17, podemos observar la matriz de confusion obtenida de la corrida con el mejor desempeño y se nos muestra los porcentajes de precision por cada una de las clases.

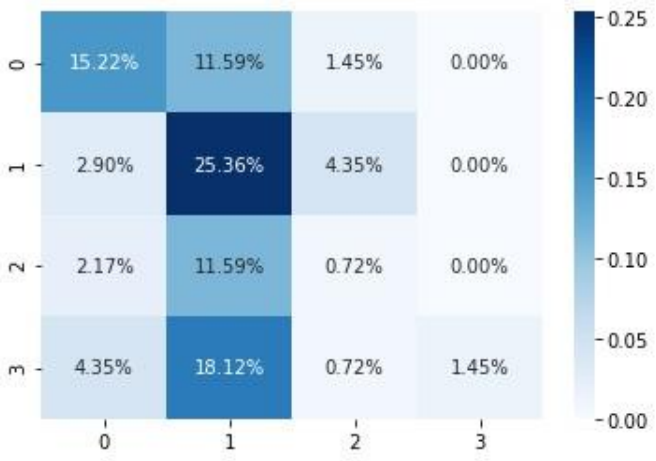

Figura 4.17: Matriz de confusión del algoritmo de clasificacion Support Vector Machine(SVM).

Resultados del proceso de clasificacion en el escenario Ranked Batch-Mode Sampling(RBMS).

Como en el escenario de experimentacion anterior, se procedi a elegir el experimento con el mejor desempeño de los 16 que se efectuaron en esta etapa. Posteriormente se ejecuto el mismo proceso de la unión de la matriz de datos previamente etiquetado con la matriz resultante del proceso de etiquetado con el Aprendizaje Activo (Ative Learnig) y seguidamente se redujo la dimension de este set de datos con la técnica PCA, como se explico con anterioridad. De esta reduccion se obtuvo una nueva matriz de dimension 690x21, siendo la ultima columna la que almacena las etiquetas o ”targets” que seran útiles para el proceso de clasificacion como se muestra en la tabla 4.9. Esta nueva matriz de datos fue sometida a los siguientes clasificadores:

- Gaussian Naive Bayes(GNB).
- K-nearest neighbor(KNN).
- Multilayer Perceptron (MLP).
- Support Vector Machine(SVM).

En este escenario de experimentacion se comenzó con la ejecucion del algoritmo de clasificacion Gaussian Naive Bayes(GNB, donde segmentamos los datos en 80 % para el entrenamiento y el 20 % restante lo empleamos para la prueba o ”test”, de la varias corridas que se realizaron en este proceso de clasificación el porcentaje mayor de precisión fue del 40 % como se muestra en la tabla 4.14.

	precision	recall	f1-score	support
1.0	0.89	0.35	0.50	46
2.0	0.38	0.67	0.48	36
3.0	0.25	0.81	0.38	16
4.0	0.50	0.05	0.09	40
accuracy			0.40	138
macro avg	0.50	0.47	0.36	138
weighted avg	0.57	0.40	0.36	138

Tabla 4.14: Reporte de precision del algoritmo de clasificacion Gaussian Naive Bayes(GNB).

En la figura 4.18, nos muestra la matriz de confusión donde podemos observar los porcentajes de clasificacion de cada una de las clases.

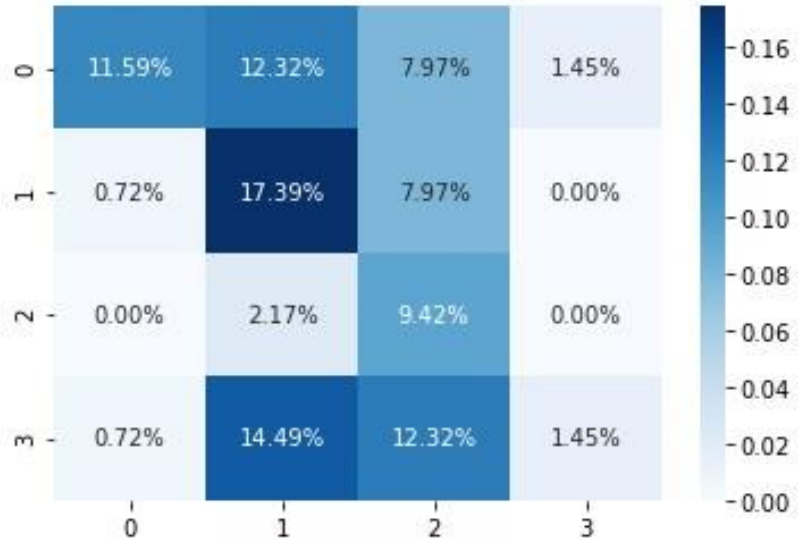

Figura 4.18: Matriz de confusión del algoritmo de clasificacion Gaussian Naive Bayes(GNB.

Continuaremos con el algoritmo de clasificacion K-nearest neighbor(KNN) que al igual que en el proceso de clasificación del escenario de experimentación anterior, se segmentaron los datos los datos en 80 % para el entrenamiento y el 20 % restante lo empleamos para la prueba o ”test” y donde se evalu de manera iterativa el numero de ”vecinos” o K en un rango de 1 a 25, esto con la finalidad de analizar un numero considerable de vecinos para observar su desempeño; donde el mayor porcentaje de precisión que se obtuvo fue del 51 % como se muestra en la tabla 4.15.

	precision	recall	f1-score	support
1.0	0.54	0.62	0.58	53
2.0	0.48	0.54	0.51	37
3.0	0.62	0.53	0.57	15
4.0	0.41	0.27	0.33	33
accuracy			0.51	138
macro avg	0.51	0.49	0.50	138
weighted avg	0.50	0.51	0.50	138

Tabla 4.15: Reporte de precision del algoritmo de clasificación K-nearest neighbor(KNN).

De esta corrida de clasificación, se obtuvo la matriz de confusion que se presenta en la figura 4.19, donde podemos observar el porcentaje de precision de cada clase.

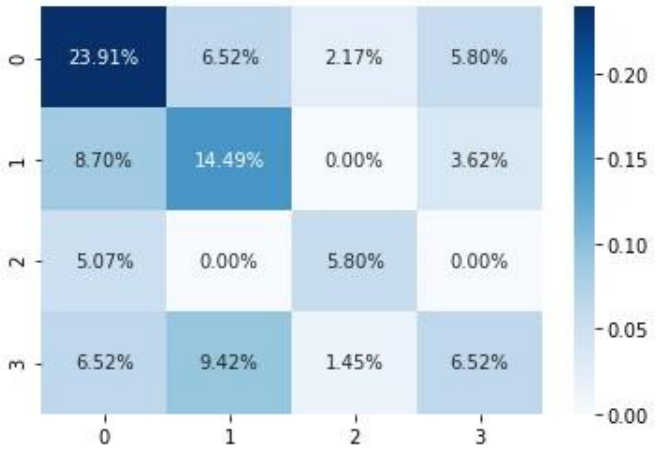

Figura 4.19: Matriz de confusion del algoritmo de clasificacion K-nearest neighbor(KNN).

Siguiendo con el proceso de clasificación de este escenario de experimentacion, continuaremos con la ejecucion del algoritmo de clasificación Multilayer Perceptron (MLP) que al igual que en el escenario anterior, una segmentación de datos en 80 % para el entrenamiento y el 20 % restante lo empleamos para la prueba o "test" y se establecieron 3 capas ocultas para este algoritmo, de las diferentes corridas que se ejecutaron se obtuvo un porcentaje de precisión del 41 %, como se muestra en la tabla 4.16.

	precision	recall	f1-score	support
1.0	0.38	0.62	0.47	37
2.0	0.45	0.44	0.44	50
3.0	0.39	0.55	0.46	20
4.0	1.00	0.03	0.06	31
accuracy			0.41	138
macro avg	0.56	0.41	0.36	138
weighted avg	0.55	0.41	0.37	138

Tabla 4.16: Reporte de precision del algoritmo de clasificación Multilayer Perceptron (MLP).

En la figura 4.20, se presenta la matriz de confusion obtenida de la corrida con el mejor desempeño, donde se muestran los porcentajes de clasificacion de cada clase.

Figura 4.20: Matriz de confusión del algoritmo de clasificacion Multilayer Perceptron (MLP).

Para terminar con el proceso de clasificacion de este escenario de experimentacion, detallaremos la aplicacion del algoritmo de clasificacion Support Vector Machine(SVM) que al igual que en el escenario anterior, se aplicó la segmentacion el set de datos en 80 % para el entrenamiento y el 20 % restante lo empleamos para la prueba o ”test” y en las diferentes corridas realizadas se obtuvo un mayor porcentaje de precisión del 49 % como se muestra en la tabla 4.17.

	precision	recall	f1-score	support
1.0	0.83	0.51	0.63	59
2.0	0.41	0.74	0.53	38
3.0	0.26	0.36	0.30	14
4.0	0.27	0.15	0.19	27
accuracy			0.49	138
macro avg	0.44	0.44	0.41	138
weighted avg	0.55	0.49	0.48	138

Tabla 4.17: Reporte de precision del algoritmo de clasificación Support Vector Machine(SVM).

Por ultimo, en la figura 4.21 se nos presenta la matriz de confusion obtenida de la corrida con el mejor desempeño, donde se puede observar los porcentajes de precision de cada clase.

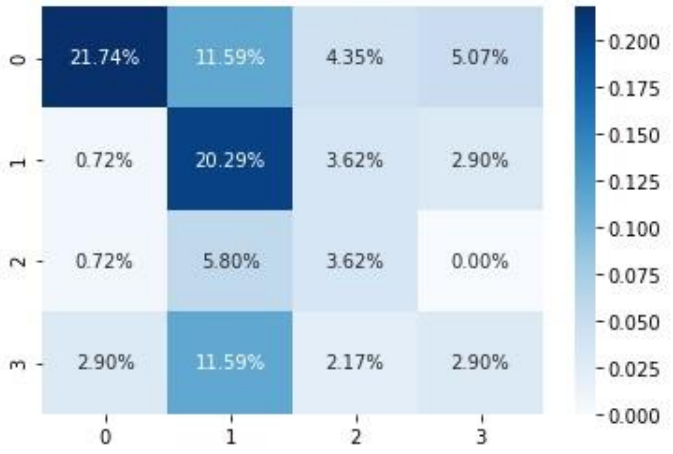

Figura 4.21: Matriz de confusión del algoritmo de clasificacion Support Vector Machine(SVM).

Conclusiones

El objetivo inicial de este trabajo de investigación era el preprocesamiento de la señal Electroencefalografica y la aplicación de técnicas de inteligencia artificial, que nos permitiría obtener informacion relevante para identificar los trastornos de lenguaje en niños de edad escolar. Pero debido a la pandemia de COVID-19, suceso que afecto a nivel mundial y que ha quedo grabado en los libros de historia, nos provoco que nos replanteáramos el rumbo de trabajo de esta investigacion ya que por las medidas de prevencion de contagios y al largo periodo de cuarentena nos fue imposible continuar con el trabajo de campo planeado, en el cual obtendríamos la señal Electroencefalografica de niños en edad escolar. Esto provoco un giro de 180 grados en el trabajo de investigacion, donde se acoto el objetivo a la identificación de el "Proceso Cerebral del Lenguaje", a cambiar el trabajo de campo para lo obtención de la señal Electroencefalografica y sustituirla por un set de datos de un estudio anterior mencionado en el tercer capítulo de este trabajo de tesis en la página 55.

La aplicacion de la técnica de Aprendizaje Activo (Active Learning) en el proceso de etiquetado de señales Electroencefalográficas, resulta una técnica de bajo costo, no invasiva, portable y un campo poco explorado que, poco a poco a manifestado ser una herramienta de diagnostico potencialmente útil demostrando prometedores resultados para esta tarea. Si bien, el tratamiento y clasificación de señales electroencefalograficas ya resulta todo un reto; ya que estas señales poseen gran cantidad de ruido, la ejecución de la técnica de Aprendizaje Activo (Active Learning) nos permitiría reducir el tiempo y costo computacional en un gran porcentaje.

Con el cambio en el rumbo de trabajo de esta investigacion se planteo un nuevo objetivo, donde nos propusimos modelar un sistema cerebro computadora (BCI, por sus siglas en ingles), capaz de identificar características del "Proceso Cerebral del Lenguaje" proveniente de la señal EEG entre individuos de edad escolar (ver tabla 3.2). Con respecto a lo anterior, se puede concluir que el proyecto de investigación

alcanzo su objetivo general, ya que se diseño un proceso integral el cual se compone del preprocesamiento de la señal EEG, la extraccion y analisis de características, etiquetado de la señal EEG a través del Aprendizaje Activo (AL, por sus siglas en Inglés) y clasificación; siendo la etapa de Aprendizaje Activo la de mayor importancia para este trabajo de investigacion, donde se pudo estudiar la aplicación de la técnica de AL, para el etiquetado se señales electroencefalográficas.

En particular en este trabajo de investigacion se estudio la técnica y/o escenario de Aprendizaje Activo (Active Learning) llamada Muestreo Basado en Grupos (Pool-based Sampling), dicha técnica fue implementada en un set de datos del dominio publico llamado Child Mind Institute - Multimodal Resource for Studying Information Processing in the Developing Brain (MIPDB) [Simon P. Kelly, 2016], destinado a el avance científico en el área de la neuropatología, donde se ejecut el proceso y la metodología propuesta en el tercer capitulo de esta tesis de investigacion y que se aplico en un segmento de datos del set de datos antes mencionado, en la cual llevamos a cabo dos escenarios de experimentación principales identificados con los nombres Uncertainty Sampling y Ranked Bach-Mode Sampling, como se describe en la página 95 de esta tesis .

En cada uno de estos escenarios de experimentacion se realizaron diferentes variaciones, una de ellas es la manera de evaluar el proceso de etiquetado del algoritmo de Aprendizaje Activo (Active Learning); donde se tomaron en cuenta los siguientes criterios:

- Evaluación con respecto a los datos etiquetados y NO etiquetados, donde se crea un conjunto de datos de manera aleatoria entre los datos etiquetados y los no etiquetados.
- Evaluación con respecto a los datos etiquetados, donde se toma solamente el conjunto de datos etiquetados de nuestro corpus de datos.

La segunda variación aplicada a cada uno de los dos escenarios principales de experimentacion, corresponde al estimador usado en el proceso de aprendizaje del algoritmo de AL; donde se emplearon los clasificadores K-vecinos más cercanos (K-nn) y Bosques Aleatorios (Random Forest). Tomando en cuenta lo anterior, se tuvieron en total 32 experimentos entre ambos escenarios de experimentación, los cuales se dividen en 16 experimentos para el escenario de experimentacion Uncertainty Sampling y 16 experimentos para el escenario de experimentación Ranked Batch-Mode Sampling. A continuacion, mostraremos la compendio del análisis estadístico resultado de la fase experimental de este trabajo de investigacion que comprende un total de 16 experimentos por cada uno de los escenarios de experimentacion y donde sintetizaremos los mejores resultados por cada uno de los escenarios de experimentacion.

	Escenario de Experimentacion: Uncertanty Sampling			
	Estimador: K-nn=4		Estimador:Random Forest	
	Evaluación datos etiquetados y NO etiquetados	Evaluación datos etiquetados	Evaluación datos etiquetados y NO etiquetados	Evaluación datos etiquetados
Mínimo	0.2272	0.5718	0.2924	0.6031
Máximo	0.2715	0.5718	0.3342	0.6971
Desviación estándar	0.009915604	1.33898E-15	0.009173526	0.019769573
Varianza de la muestra	9.83192E-05	1.79287E-30	8.41536E-05	0.000390836
Curtosis	1.875638904	-2.041237113	-0.282799783	-0.220124547
Coeficiente de asimetría	-1.438243981	-1.015293303	0.264740641	-0.073418509
No. Consultas	100	100	100	100

Tabla 4.18: Síntesis de los experimentos donde se muestran los puntajes maximos de precisión del escenario de experimentacion Uncertainty Sampling.

	Escenario de Experimentacion: Ranked Batch-Mode Samplig			
	Estimador: K-nn=4		Estimador:Random Forest	
	Evaluación datos etiquetados y NO etiquetados	Evaluación datos etiquetados	Evaluación datos etiquetados y NO etiquetados	Evaluación datos etiquetados
Mínimo	0.2846	0.6162	0.248	0.6397
Máximo	0.3055	0.6162	0.3029	0.705
Desviación estándar	0.005417263	2.27813E-16	0.013113994	0.016851277
Varianza de la muestra	2.93467E-05	5.18987E-32	0.000171977	0.000283966
Curtosis	2.094191665	-2.235294118	0.293890253	-0.509516792
Coeficiente de asimetría	1.253628411	1.082977149	-0.072793259	0.06127677
No. Consultas	20	20	20	20

Tabla 4.19: Síntesis de los experimentos donde se muestran los puntajes maximos de precisión del escenario de experimentacion Ranked Batch-Mode Sampling.

Tomando como referencia las tablas 4.18 y 4.19, en conjunto con los porcentajes de clasificacion mostrados en el escenario de experimentación Ranked Batch-Mode Sampling, es importante destacar el buen desempeño obtenido empleando el conjunto de 'Evaluación datos etiquetados' en combinación con el clasificador Random Forest que fue usado como estimador en ambos escenarios de experimentacion, ya que para el tipo de datos usados durante esta fase, que son las señales electroencefalograficas previamente tratadas como se menciona en el tercer capítulo de este trabajo de investigacion "Desarrollo"; las métricas empleadas por este clasificador donde se evaluaron los datos presenta mayor eficacia; ya que dicha evaluacion es a través de árboles de decisión independientes, donde el voto mayoritario decide a que clase pertenece la instancia o dato evaluado.

Dichos datos del proceso de etiquetado a través de la técnica de Aprendizaje Activo (Active Learnig) se refuerzan con los resultados obtenidos del proceso de clasificacion por métodos tradicionales de Machine Learning, en el cual pudimos comparar cuatro clasificadores diferentes obteniendo resultados prometedores en este trabajo de investigación, sentando una buena base para la continuación de este trabajo de investigación.

Tomando en cuenta lo anterior, podemos concluir que la técnica Muestreo basado en grupos del proceso de Aprendizaje Activo (Active Learnig) en conjunto con la ejecucion del clasificador Random Forest a manera de ”Estimador” durante el proceso de etiquetado a través de la técnica de Aprendizaje Activo (Active Learning) y realizando las consultas en el proceso del algoritmo de Aprendizaje Activo (Active Learning) mediante los datos previamente etiquetados, nos permitió desarrollar un algoritmo de etiquetado de datos para señales electroencefalograficas empleando el Aprendizaje Activo (Active Learning), todo este proceso que se explico anteriormente le hemos bautizado como ”AP-Learning” en honor a dos pacientes (Alexa y Paulina) que inspiraron este trabajo de investigación y que resulta una muy buena estrategia a usar para el etiquetado de señales electroencefalográficas, la cual podemos definir como la primera version de una herramienta de diagnostico de trastornos de lenguaje donde empleamos la técnica Muestreo basado en grupos del proceso de Aprendizaje Activo (Active Learnig).

Trabajo a Futuro

El presente trabajo de investigacion se presenta como la primera piedra que servirá como cimiento a futuras investigaciones; algunas de ellas encaminadas directamente a continuar el rumbo de este trabajo de tesis por un servidor, así como algunas otras líneas de investigaciones futuras que podrán ser retomadas por algún otro investigador.

A continuacion, se presentaran los trabajos de investigacion futura que seran cubiertos por un servidor en el grado doctoral, y que por los sucesos ocurridos por la gran afectacion provocada por la pandemia de COVID-19 y por exceder el alcance de esta tesis, no lograron ser tratados a profundidad; teniendo como objetivo el adaptar y mejorar el modelo propuesto; entre los trabajos a futuro se destacan:

- Realizar la recolección de la señal Electroencefalografica de voluntarios en edad escolar como se tenia previsto en un principio.
- Verificar y adaptar la estructura del preprocesamiento de la señal.
- Profundizar en el estudio y aplicacion de las técnicas de Aprendizaje activo, para someterlas a un mayor volumen de datos.
- Encaminar los métodos y procesos de la investigación para la deteccion y clasificacion de trastornos del lenguaje.

Anexos

Consentimiento Informado

CARTA DE CONSENTIMIENTO

Nombre del proyecto (tesis) Adquisición De Datos Mediante Estudio De Electroencefalografía

Nombre de los investigadores:

Ing. Eugenio Salvador Martínez Velázquez Estudiante del Programa de la Maestría en Ciencias de la Computación del Instituto Tecnológico de León

Dra. María del Rosario Baltazar Flores. Profesora Investigadora de la División de Estudios de Posgrado e Investigación del Instituto Tecnológico de León.

Propósito del estudio:

Dentro de la formación de la División de Estudios de Posgrados e Investigación del Instituto Tecnológico de León, se considera de gran importancia el avance Científico y Tecnológico de nuestro país.

El objetivo de este proyecto es recolectar la señal EEG (Electroencefalograma) de una muestra de pacientes previamente establecida por el equipo de investigación y someterla a una primera fase de tratamiento a través de la limpieza de la señal EEG mediante filtrado digital y posteriormente, realizar el procesamiento de la señal para la extracción de las características y, por último, clasificar las señales mediante técnicas computacionales de inteligencia artificial para demostrar que es posible establecer diferencias en el comportamiento de las señales del lenguaje de los sujetos de prueba.

Procedimiento del estudio

El presente estudio es aplicado con fines de investigación científica, cuyo objetivo es contribuir en el avance e innovación tecnológica de nuestro país, a través de la adquisición de datos provistos por la señal de electroencefalografía. Como se describe a continuación.

- Para favorecer la concentración de los voluntarios, el estudio se deberá de llevar a cabo en un lugar sin mucho ruido y de fácil acceso.
- Antes de comenzar con el estudio, se deberá validar que el voluntario se encuentra debidamente hidratado y que tendrá los medios para hidratarse durante la prueba.
- Se validará que el voluntario no presente ningún malestar físico o emocional, así como validar que se encuentra en una posición cómoda, que no le genere cansancio o fatiga durante el estudio.

Figura 4.22: Consentimiento Informado página 1

- Una vez validados los puntos anteriores, se le colocará la diadema al voluntario, verificando que la diadema sea colocada correctamente y que no presente ninguna molestia.
- Durante la toma de la señal EGG los familiares, docentes y directivos podrán permanecer en la habitación, siempre y cuando no sea un motivo de distracción o interfieran durante el estudio.

- Se le explicara al voluntario y al familiar, las instrucciones para realizar la toma de la señal EEG, así como la importancia de la prueba por su relevancia científica, con el fin de que se tome con seriedad la prueba.
- El estudio se dividirá en sesiones, con un tiempo de duración estimado de 30 minutos por cada paciente.
- Durante las sesiones del estudio, se le darán al voluntario periodos de descanso, en los cuales se sostendrá una pequeña conversación con el fin de despejar la mente.
- El estudio completo constara de varias sesiones para la toma de la señal, por lo que se organizara la logística necesaria para la realización de esta.

Una vez que se ha detallado los propósitos y procedimiento del presente estudio, le informamos que su hijo (a) cumple con los siguientes criterios de selección.

- Edad entre los 6 a 9 años de edad.
- Cuentan con un expediente dentro del centro educativo.
- Evaluación previa del docente.
- Niño o niña en etapa escolar sujeto a un proceso de adquisición y fortalecimiento de la lectura y escritura.

Riesgos

El presente estudio, no representa ningún riesgo para la salud.

Beneficios

Al contribuir en la realización de este estudio, usted aportara los datos que serán usados en la exploración de nuevas alternativas para la comunicación, los cuales son sumamente valiosos para los fines de la investigación, así como en el avance científico y tecnológico de nuestro país.

Participación voluntaria/Retiro del estudio: La participación de su hijo (a) es completamente voluntaria por lo que usted está en plena libertar de decidir, si sobre esta invitación. Por lo que, si así lo desea, podrá solicitar se le retire del estudio sin ningún tipo de consecuencias, de manera que le pedimos externar su decisión en la sección DECLARACIÓN PERSONAL DE CONSENTIMIENTO.

Figura 4.23: Consentimiento Informado pagina 2

Aviso de privacidad simplificado: Los investigadores a cargo del estudio, serán responsables del tratamiento y resguardo de los datos personales que nos proporcione, los cuales será protegidos conforme a lo dispuesto por la **Ley General de Protección de Datos Personales en Posesión de Sujetos Obligados.** Los datos personales que solicitaremos serán utilizados exclusivamente para las finalidades expuestas en este documento, por lo que usted podrá solicitar la corrección de sus datos o retirar el consentimiento para su uso.

DECLARACIÓN PERSONAL DE CONSENTIMIENTO

- Se me ha leído esta Carta de consentimiento.
- Me han explicado el estudio de investigación incluyendo el objetivo, los posibles riesgos y beneficios, y otros aspectos de la participación de mi hija (o) en el estudio.
- He podido hacer preguntas relacionadas a mi participación en el estudio, y me han respondido satisfactoriamente mis dudas.

Si usted entiende la información que le hemos dado en este formato, está de acuerdo en participar en este estudio, de manera total o parcial, y también está de acuerdo en permitir que su información de salud sea usada como se describió antes, entonces le pedimos que indique su consentimiento para participar en este estudio.

Registre su nombre y firma en este documento del cual le entregaremos una copia.

PARTICIPANTE y/o Familiar:
Nombre:
_____
Firma:
_____
Fecha/hora _____

TESTIGO 1
Nombre:
_____
Firma:
_____
Relación con el participante: _____
Fecha/hora: _____

Figura 4.24: Consentimiento Informado pagina 3

Bibliografía

[AA.VV., 2018] AA.VV. (2018). Tecnologías de la informacion geografica: Perspectivas multidisciplinares en la sociedad del conocimiento. Publicacions de la Universitat de València.

[Aguilar et al., 2015] Aguilar, F. L., Valdivia, I., Rodríguez Valdés, R., Gárate Sanchez, E., Morgade Fonte, R., and Castillo Yzquierdo, G. (2015). ”Hallazgos electroencefalograficos en los pacientes con trastorno específico del desarrollo del lenguaje

. Rev. Cubana Neurol Neurocir.

[Albesa and Ayala, 2017] Albesa, S. A. and Ayala, C. O. (2017). Trastornos del lenguaje.

[Alvarado and Batanero, 2008] Alvarado, H. and Batanero, C. (2008). Significado del teorema central del limite en textos universitarios de probabilidad y estadÍstica. Estudios Pedagogicos (Chile) Num.2 Vol.34, 34.

[Birbaumer et al., 1990] Birbaumer, N., Elbert, T., Canavan, A. G., and Rockstroh, B. (1990). Slow potentials of the cerebral cortex and behavior. Physiological Reviews, 70(1):1–41.

[Birbaumer et al., 2000] Birbaumer, N., Heetderks, W. J., McFarland, D. J., Peckham, P. H., Schalk, G., Donchin, E., Quatrano, L. A., and Robinson, C. J. (2000). Brain–Computer Interface Technology: A Review of the First International Meeting. IEEE Transactions on Rehabilitation Engineering, 8(2):164–173.

[Birbaumer et al., 2003] Birbaumer, N., Hinterberger, T., Kübler, A., and Neumann, N. (2003). The thought-translation device (TTD): Neurobehavioral mechanisms and clinical outcome. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 11(2):120–123.

[Bishop, 2006] Bishop, C. (2006). Pattern Recognition and Machine Learning. Information Science and Statistics. Springer New York.

[Bradley et al., 2005] Bradley, W., Daroff, R., Fenichel, G., and Jankovic, J. (2005). Neurologia Clinica:. Elsevier.

[Bradley et al., 2009] Bradley, W., Daroff, R., Fenichel, G., and Jankovic, J. (2009). Neurología Clínica, Vol 1. Diagnostico y tratamiento. Vol 2. Trastornos neurologicos. Elsevier Health Sciences Spain.

[Cardinali, 1991] Cardinali, D. (1991). Manual de neurofisiología. Díaz de Santos.

[Cardinali, 2007] Cardinali, D. (2007). Neurociencia aplicada: sus fundamentos. Editorial Médica Panamericana.

[Cardoso et al., 2017] Cardoso, T. N., Silva, R. M., Canuto, S., Moro, M. M., and Gonçalves, M. A. (2017). Ranked batch-mode active learning. Information Sciences, 379:313 – 337.

[Danka and Horvath, 2018] Danka, T. and Horvath, P. (2018). modAL: A modular active learning framework for Python. available on arXiv at `https://arxiv.org/abs/1805.00979`.

[Daubechies, 1992] Daubechies, I. (1992). Ten Lectures on Wavelets. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104).

[Eldar, 2015] Eldar, Y. (2015). Sampling Theory: Beyond Bandlimited Systems. Cambridge University Press.

[Elliott, 1988] Elliott, D. F. (1988). Handbook of Digital Signal Processing: Engineering Applications. Academic Press.

[Fazel-rezai, 2011] Fazel-rezai, R. (2011). RECENT ADVANCES IN BRAIN COMPUTER INTERFACE SYSTEMS Edited by Reza Fazel-Recent Advances in Brain-Computer Interface Systems.

[Florian, 2010] Florian, A. (2010). Cerebro, mente y conciencia. Internet Medical Plublishing.

[Freund and Simon, 1994] Freund, J. and Simon, G. (1994). Estadística elemental. Pearson Educacion.

[Kozhushko et al., 2018] Kozhushko, N. J., Nagornova, Z. V., Evdokimov, S. A., Shemyakina, N. V., Ponomarev, V. A., Tereshchenko, E. P., and Kropotov, J. D. (2018). Specificity of spontaneous EEG associated with different levels of cognitive and communicative dysfunctions in children. International Journal of Psychophysiology, 128:22–30.

[Langer et al., 2017] Langer, N., Ho, E. J., Alexander, L. M., Xu, H. Y., Jozanovic, R. K., Henin, S., Petroni, A., Cohen, S., Marcelle, E. T., Parra, L. C., and et al. (2017). A resource for assessing information processing in the developing brain using eeg and eye tracking. Scientific Data, 4(1).

[Lawhern et al., 2015] Lawhern, V., Slayback, D., Wu, D., and Lance, B. J. (2015). Efficient labeling of eeg signal artifacts using active learning. In 2015 IEEE International Conference on Systems, Man, and Cybernetics, pages 3217–3222.

[Lopez et al., 2008] Lopez, L., Pérez, S., and de la Torre, M. (2008). Neuroanatomía. Editorial Médica Panamericana S.A.

[Malik and Amin, 2017] Malik, A. and Amin, H. (2017). Designing EEG Experiments for Studying the Brain: Design Code and Example Datasets. Elsevier Science.

[Martínez Perez, 2009] Martínez Perez, L. A, P. (2009). Terapia regresiva reconstructiva: una luz en el laberinto. Un método para reparar el alma. Editorial Libros en Red.

[Mayor et al., 2013] Mayor, L., Burneo, J., and Ochoa, J. (2013). Manual de electroencefalografía: Handbook of Electroencephalography. Universidad de los Andes, Facultad de Medicina.

[Morales, 2020] Morales, D. (2020). Oscilopatología en trastornos del espectro autista. Areté, 20.

[Obler et al., 2001] Obler, L., Gjerlow, K., Méndez, E., and Tena, P. (2001). El lenguaje y el cerebro. Serie Lingüísitica / Cambridge University Press. Ediciones Akal.

[Pascual-Marqui, 2002] Pascual-Marqui, R. (2002). Standardized low resolution brain electromagnetic tomography (sloreta): Technical details. Methods and findings in experimental and clinical pharmacology, 24 Suppl D:5–12.

[Poeppel and Hickok, 2004] Poeppel, D. and Hickok, G. (2004). Towards a new functional anatomy of language. Cognition, 92(1):1 – 12. Towards a New Functional Anatomy of Language.

[Proakis and Manolakis, 1996] Proakis, J. and Manolakis, D. (1996). Digital signal Processing: Principles, Algorithms ,and Applications. Macmillan.

[Rangayyan, 2015] Rangayyan, R. (2015). Biomedical Signal Analysis. IEEE Press Series on Biomedical Engineering. Wiley.

[Rodríguez Hernandez, 1998] Rodríguez Hernandez, O. (1998). Temas de Analisis Estadístico Multivariado. Editorial Universidad de Costa Rica.

[Saltz and Stanton, 2017] Saltz, J. and Stanton, J. (2017). An Introduction to Data Science. SAGE Publications.

[Schreuder, 2014] Schreuder, M. (2014). Towards Efficient Auditory BCI Through Optimized Paradigms and Methods. epubli GmbH.

[Serna, 2018] Serna, B. (2018). 'Construcción de una interfaz Cerebro-Maquina por medio de un traductor de señales neuronales a palabras Dicotomicas textuales. Master's thesis.

[Settles, 2010] Settles, B. (2010). Active learning literature survey.

[Simon P. Kelly, 2016] Simon P. Kelly, PhD, M. M. M. P. N. L. P. L. P. P. S. C. M. (2016). Child mind institute - multimodal resource for studying information processing in the developing brain (mipdb). url http://fcon$_1$000.projects.nitrc.org/indi/cmi$_e$eg/index.html.

[Sundararajan, 2016] Sundararajan, D. (2016). Discrete Wavelet Transform: A Signal Processing Approach. CourseSmart Series. Wiley.

[The MathWorks, 2019] The MathWorks, I. (2019). Wavelet Toolbox. Natick, Massachusetts, United State.

[Tonin et al., 2018] Tonin, A., Birbaumer, N., and Chaudhary, U. (2018). A 20-questions-based binary spelling interface for communication systems. Brain Sciences, 8(7).

[Torres García et al., 2016] Torres García, A., Reyes García, C. A., Villaseñor Pineda, L., and García Aguilar, G. (2016). Analisis y clasificacion de electroencefalogramas (EEG) registrados durante el habla imaginada. Phd, Instituto Nacional de Astrofísica, Óptica y Electronica.

[Torres García et al., 2013] Torres García, A., Reyes García, C. A., Villaseñor Pineda, L., and Ramírez Cortés, J. (2013). Analisis de Señales Electroencefalograficas para la Clasificación de Habla Imaginada. Revista Mexicana de Ingenieria Biomedica, 34(1):23–39.

[Triola et al., 2012] Triola, M., Ayala, L., and Ramírez, R. (2012). Estadistica. Pearson Education.

[Usategui et al., 2007] Usategui, J., Martínez, I., and Blanca, M. (2007). Electronica digital y microprogramable. Electricidad-electronica. Paraninfo.

[Vega, 2012] Vega, F. (2012). Neurociencia del Lenguaje: Bases neurologicas e implicaciones clínicas. Editorial Medica Panamericana Sa de.

[Vicario, 1999] Vicario, C. (1999). Neurobiología de la vision. Politext Series. Edicions UPC.

[Vidal, 1973] Vidal, J. J. (1973). "Toward Direct Brain-Computer Communication . Annual review of Biophysics and Bioengineering, 2(7):157–180.

[Von Neumann et al., 1999] Von Neumann, J., Fontelles, J., and Mayeur, C. (1999). El ordenador y el cerebro. Antoni Bosch Editor, S.A.

[Zuleta, 2007] Zuleta, E. (2007). El sistema nervioso : desde las neuronas hasta el cerebro humano. Salud (Medellín, Colombia).: Interés general. Editorial Universidad de Antioquia.