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

ESTUDIO DEL IMPACTO DEL USO DE ELECTROENCEFALOGRAMA EN LA IDENTIFICACIÓN DE WEBSITE KEYOBJECTS

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

GINO ALESSANDRO SLANZI RODRÍGUEZ

PROFESOR GUÍA:
SR. JUAN DOMINGO VELÁSQUEZ SILVA

MIEMBROS DE LA COMISIÓN:
SR. PEDRO MALDONADO ARBOGAST
SR. PABLO LOYOLA HEUFEMANN

SANTIAGO DE CHILE
2014

Financiado por el Proyecto FONDEF CA12I10061 - AKORI

Resumen Ejecutivo

RESUMEN DE LA MEMORIA
PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL INDUSTRIAL
POR : GINO SLANZI RODRÍGUEZ
FECHA: 19/11/2014
PROF. GUIA: SR. JUAN VELÁSQUEZ

El presente Trabajo de Título tiene como objetivo principal conocer el impacto que significa la incorporación de una nueva fuente de información a la metodología de identificación de *Website Keyobjects*. Esta nueva fuente de información será la medida de la actividad bioeléctrica cerebral frente a los estímulos presentados en una página web. Específicamente, se busca diseñar e implementar un módulo con algoritmos de Data Mining para clasificar los objetos relevantes presentes en un sitio web según variables de actividad cerebral.

El trabajo se enmarca en el proyecto Fondef "Plataforma informática basada en web-intelligence y herramientas de análisis de exploración visual para la mejora de la estructura y contenido de sitios web (AKORI: *Advanced Kernel for Ocular Research and Web Intelligence*)". Este proyecto está siendo llevado a cabo en conjunto por el Laboratorio de Neurosistemas y el Departamento de Ingeniería Industrial de la Universidad de Chile.

La hipótesis que se valida en este trabajo es: *Es posible realizar una clasificación de los objetos relevantes de un sitio web según variables que caractericen la actividad bioelétrica cerebral*.

Para la validación de esta hipótesis se realiza un experimento en que 20 usuarios navegan libremente por un sitio web de estudio, mientras que un dispositivo de *Eye Tracking* guarda el posicionamiento de los ojos en la pantalla y los cambios en el tamaño de las pupilas y un electroencefalograma graba los potenciales eléctricos en la corteza cerebral en tiempo real.

El análisis de los datos obtenidos se realiza mediante el proceso KDD, en un principio se realiza para la dilatación pupilar con el fin de obtener una línea base de estudio. Luego se aplica a los datos del EEG para realizar una comparación a lo obtenido de las señales oculares. Para ambos tipos de data se seleccionan 19 sujetos, cuyos datos son preprocesados y limpiados, luego transformados según distintas características a los que se les aplican diversos algoritmos de clasificación. Con esto se determina como resultado una lista de objetos relevantes dentro del sitio de estudio.

Los resultados obtenidos indican que utilizando variables obtenidas de las señales eléctricas producidas en la corteza cerebral es posible clasificar *Website Keyobjects* con un 90% de precisión, mediante el algoritmo *K-Means*. Esta clasificación es en base a la línea base obtenida, donde de un total de 20 objetos, 18 fueron clasificados correctamente, y fue la mejor que se obtuvo dentro de todas las combinaciones de variables y clusterizaciones realizadas.

Finalmente se concluye que el trabajo fue exitoso y se proponen diversos trabajos futuros que aportan al proyecto AKORI y la metodología. Además se entrega una reflexión final correspondiente al uso de metodologías, algoritmos y conocimientos similares en otras áreas de estudio que generen valor para la sociedad.

A mis abuelos Juan, Coca, Gino y María

Agradecimientos

Al terminar esta larga etapa universitaria me gustaría agradecer a todos los que han estado conmigo durante estos años.

En primer lugar quisiera agradecer a mi familia por el apoyo durante toda mi vida. A mis padres por ser siempre los mejores. Por motivarme y darme fuerza en todo momento. A mis hermanas por estar en todas conmigo, quererme y valorarme, por entregarme palabras precisas y enseñanzas.

A Laura por estar siempre, por motivarme y comprenderme.

A mis amigos por los buenos momentos.

A los profesores Juan y Pedro por la ayuda en este difícil trabajo, sus palabras y comentarios. También a la gente de la salita, que siempre me ayudaron en lo que pudieron. A todos aquellos que aportaron siendo sujetos de experimentación y a la gente del Laboratorio de Neurosistemas.

Muchas gracias a todos los que han recorrido a mi lado este largo camino.

Tabla de Contenido

Índice de Tablas

Índice de Figuras

1 Introducción

En este capítulo se introduce al lector al proyecto de Memoria. En primer lugar, se contextualiza y entrega una idea general de la situación en la que se desarrolla el Trabajo de Título. Posteriormente, se describe el proyecto en detalle, luego de ser propuesta una hipótesis de investigación: se declaran los objetivos; la metodología de trabajo; los resultados esperados y los alcances. Finalmente, se entrega el plan de trabajo y la estructura de este informe.

1.1 Antecedentes Generales

Desde su nacimiento y a medida que transcurren los años, Internet ha ido creciendo explosivamente en su uso y penetración. Actualmente, cerca del 34% de la población mundial tiene acceso a la Web y, particularmente, en Chile el 58% puede acceder a su utilización [5].

Por otro lado, las instituciones ven en la Web una vitrina disponible para mostrarse al mundo. Las empresas exponen sus productos y servicios en sus sitios web, llevando la competencia hasta el nivel *online*, donde intentan atraer nuevos clientes y retenerlos.

Debido a lo anterior es que se hace necesario el estudio del comportamiento de los usuarios web, sus preferencias e intereses. Es así como ha nacido el área de investigación denominada *Web Mining*, la cual busca extraer conocimiento de los datos generados en la Web, es decir, del contenido del sitio web, su estructura y el uso. Con esto, los *webmasters* y administradores pueden generar nuevas y mejoras estrategias de creación de valor.

En ese sentido, variados estudios se han realizado entregando a la comunidad resultados interesantes. En 2004, el profesor Juan Velázquez junto a otros investigadores desarrollaron una metodología que permitía encontrar las palabras más importantes dentro de un sitio Web. Las denominaron *Website Keywords* y las definieron como "*una palabra o un conjunto de palabras que son utilizadas en su proceso de búsqueda y que caracterizan el contenido de una página o sitio web dado*" [6]. Limitado solo a texto, este estudio fue mejorado posteriormente, extendiendo el análisis a contenido de texto y multimedia presente en los sitios Web [1].

En [1] se definen los *Web Objects* como "*un conjunto de palabras estructuradas o un*

recurso multimedia presente dentro de una página que posee metadatos que describan su contenido". Con la definición anterior, es posible definir los *Website Keyobjects*, que son "*objetos o conjuntos de objetos que atraen la atención del usuario y que caracterizan el contenido de un sitio web o página dada*". En este trabajo, se diseña e implementa una metodología que permite identificar los *Website Keyobjects* por medio del análisis de los datos guardados en los *Web Logs* y encuestas.

Posteriormente, se aumenta el grado de precisión de la metodología, añadiendo información empírica del usuario. En [7] se utiliza una técnica de *eye tracking*, para medir el tiempo de permanencia de los usuarios en los objetos.

En [4] se agrega el análisis de dilatación pupilar como variable de estudio en la identificación de los *Website Keyobjects*, encontrando como resultado que por sí sola esta variable no adquiere mucha relevancia, puesto que se necesita conocer la relación entre el estímulo y la emocionalidad. Es ahí donde nace la posibilidad de realizar este Trabajo de Memoria, para incluir la actividad bioeléctrica cerebral a la metodología, con el fin de obtener resultados más precisos y robustos.

1.2 Descripción del Proyecto y Justificación

Como se mostró anteriormente, el Trabajo de Memoria buscará mejorar la metodología para la identificación de *Website Keyobjects*, por medio de la inclusión de una nueva fuente de información, la actividad bioeléctrica cerebral. Esta información se obtendrá con la utilización de un electroencefalograma además de los equipos usados en los estudios anteriores.

Este trabajo está inmerso en el proyecto FONDEF titulado "Plataforma informática basada en web-intelligence y herramientas de análisis de exploración visual para la mejora de la estructura y contenido de sitios web (AKORI: *Advanced Kernel for Ocular Research and Web Intelligence*)". Este proyecto está siendo llevado a cabo en conjunto por la Escuela de Medicina de la Facultad de Medicina y el Departamento de Ingeniería Industrial de la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile.

El proyecto AKORI busca generar un servicio tecnológico de apoyo a la toma de decisiones de estructura, contenido y diseño de sitios web, destinado a las organizaciones.

Las metodologías implementadas para conocer los *Website Keyobjects* se van haciendo cada vez más precisas. El último estudio [4], muestra que la dilatación pupilar por sí sola no agrega mucho valor, por lo que se hace necesario analizar la emocionalidad subyacente en el comportamiento del usuario frente a los estímulos presentados en el sitio Web.

Se busca obtener conocimiento empírico del tipo de emociones que sienten los usuarios al verse enfrentados a los sitios Web, de manera de contar con un respaldo para la toma de decisiones que optimicen el diseño y la efectividad de los sitios Web.

1.3 Hipótesis de Investigación

La hipótesis de investigación de este trabajo plantea la posibilidad de conocer el estado emocional que experimentan los usuarios frente a estímulos visuales al navegar por un sitio web, reflejado en la importancia que tienen los objetos que lo componen. En otras palabras, lo que se quiere verificar es lo siguiente:

Es posible realizar una clasificación de los objetos relevantes de un sitio web según variables de caracterización de la actividad bioeléctrica cerebral.

1.4 Objetivos

Los objetivos que se han planteado para este trabajo son los siguientes:

1.4.1 Objetivo General

Diseñar e implementar un módulo con algoritmos de Data Mining para clasificar los objetos presentes en un sitio web según variables de actividad bioeléctrica cerebral.

1.4.2 Objetivos Específicos

1. Realizar un análisis del estado del arte sobre eye tracking, análisis de dilatación pupilar y EEG, con el fin de conocer técnicas e indicadores para usar en el trabajo.
2. Diseñar e implementar un experimento con el fin de obtener los datos necesarios para el análisis.
3. Caracterizar los datos obtenidos por el *eye tracker* y el EEG en el experimento.
4. Desarrollar técnicas para analizar los datos caracterizados.
5. Evaluar el impacto del uso del EEG, comparándolo con resultados anteriores.

1.5 Metodología

De modo de cumplir con los objetivos presentados anteriormente, se plantea una metodología de trabajo dividida en cuatro pasos.

La primera parte del trabajo consiste en un estudio profundo de todos los contenidos que son vistos en el proyecto, vale decir, técnicas de *eye tracking*, análisis de dilatación pupilar, electroencefalograma, *web mining*, *web objects*, etc.

La segunda fase está basada en el experimento. En esta etapa es importante lograr un buen trabajo junto a la gente de Medicina, para llevar a cabo un experimento que arroje resultados satisfactorios en cuanto a los datos que se quieren obtener. Este último estará basado en el experimento utilizado en [4], con la consideración que ahora se deberá incorporar en el sujeto experimental, el uso de un dispositivo de EEG. Además se tiene que recolectar la data Web necesaria, es decir, el sitio de estudio, sus objetos y los *web server logs*.

Posteriormente se analizan los datos obtenidos como resultado del experimento. Para esto se lleva a cabo el proceso KDD (*Knowledge Discovery in Databases*), en el que se aplican técnicas de *Data Mining* para conocer los objetos relevantes de un sitio web de estudio y clasificarlos según variables de emocionalidad.

Finalmente se concluirá en relación a la hipótesis de investigación y los resultados obtenidos. Se confeccionará un listado de los elementos principales del sitio web, considerando sus valores de las características empleadas.

1.6 Resultados Esperados y Alcances

Se espera generar una propuesta metodológica que incluya la bioactividad cerebral como nueva fuente de información para identificar *Website Keyobjects*. En este sentido, se contará con un módulo para clasificar estos objetos y finalmente con un listado de aquellos *Website Keyobjects* encontrados, clasificados según sus valencias y excitaciones emocionales respectivas.

El alcance de este Trabajo de Memoria es agregar una fuente de información al modelo de identificación de *Website Keobjects* generado en estudios previos, que ha ido mejorando de esta misma manera. Esta fuente de información vendrá de la actividad bioeléctrica cerebral obtenida desde un electroencefalograma.

Con este nuevo parámetro se evaluará la emocionalidad que existe detrás de las respuestas

fisiológicas vistas en estudios anteriores como la dilatación pupilar y la trayectoria del ojo. Así se integrará esta medida a la metodología y se analizará el aporte que significa para el estudio.

Por otro lado, el trabajo se centrará en el área web de estructura y contenido. Utilizará sólo páginas estáticas y se basará principalmente en técnicas de estudio de *Eye Tracking* y EEG.

1.7 Estructura de la Memoria

El presente informe está organizado en seis capítulos como se describe a continuación.

El primer capítulo describe la introducción al trabajo de Memoria, entregando antecedentes y descripciones generales del proyecto. Se plantea la hipótesis de investigación y se fijan los objetivos y metodología de trabajo. Además, se detallan los resultados esperados y alcances para posteriormente mostrar el plan de trabajo y la estructura del informe.

El segundo capítulo corresponde al marco teórico, que tiene como principal finalidad proveer al lector de la información necesaria para el entendimiento de los conceptos utilizados en el desarrollo del proyecto. En este capítulo se presentan los conceptos de Web e Internet, *Web Intelligence*, dilatación pupilar, bioactividad cerebral, emocionalidad, proceso KDD, y *Website Keyobject*.

El tercer capítulo muestra la propuesta de investigación sobre la que se fundamenta la hipótesis. Se explica detalladamente el procedimiento a seguir para validar la hipótesis planteada. Es decir, se profundiza en la metodología, explicando específicamente qué tipo de datos es necesario obtener para el análisis y cuál es el tipo de experimento que se requiere. Además se señala cuáles son los modelos y algoritmos que se utilizan y por qué.

En el cuarto capítulo se describe el experimento realizado para la obtención de datos. Se detalla su diseño e implementación por separado, utilizando tópicos como grupo experimental, instrumentación, sitio web, protocolo y resultados esperados.

El quinto capítulo exhibe la parte más importante de este trabajo, el análisis de los datos y los resultados obtenidos en base a la implementación del proceso KDD.

Finalmente, el sexto capítulo concluye sobre los principales resultados obtenidos y la hipótesis de investigación planteada. Se agregan discusiones y se proponen distintas aristas de trabajo futuro.

2 Marco Teórico

En este capítulo se introduce al lector los conceptos sobre los cuales se enmarca el trabajo de memoria, con el fin de entregar la información básica y necesaria para el entendimiento de los temas a tratar en el proyecto.

Los temas que se tratan tienen relación con la Web e Internet, *Web Intelligence*, dilatación pupilar, bioactividad cerebral, emocionalidad, proceso KDD y *Website Keyobjects*.

2.1 Web e Internet

Es común para la gente referirse a Internet y a la Web como un solo concepto sin diferencia alguna, pero es importante mencionar que son dos términos completamente distintos. Por un lado, Internet es la red de redes que permite la interconexión entre dispositivos, a través del envío y recepción de datos que viajan en paquetes. En cambio, la Web es el conjunto de páginas y objetos relacionados que se vinculan entre sí por medio de hipervínculos [8].

La arquitectura de la Web está basada en tres conceptos principales :

1. Hypertext Transfer Protocol (HTTP) [9]: Corresponde a un protocolo de comunicación para transferir datos entre dispositivos en una red. Las especificaciones de este protocolo, son mantenidas por el World Wide Web Consortium (W3C).
2. Hypertext Markup Languaje (HTML) [10]: Es un lenguaje estándar de hipertexto que sirve para describir la estructura y contenido de los sitios web. Además permite entrelazar información mediante vínculos o links. Mediante este lenguaje se crean códigos que son interpretados por los navegadores web, que a su vez muestran los sitios a los usuarios.
3. Uniform Resource Locator (URL) [11]: Es un localizador de recursos dentro de la Web. Generalmente contiene el protocolo usado, el nombre del dominio y el archivo que el usuario revisa en particular. Por ejemplo en https://correo.cec.uchile.cl/login.php, el protocolo es https//; el dominio es correo.cec.uchile.cl, y el archivo es login.php.

La información generada en la Web puede ser clasificada dentro de tres ámbitos [12]:

1. Contenido: Se refiere a los objetos en las páginas Web, que pueden ser textos, imágenes, sonidos o vídeos. El texto es fácilmente representado, mientras que el contenido multimedia necesita de metadatos para ser descrito.
2. Estructura: Son los *links* entre las páginas. Generalmente, cuando hay un *link* entre páginas, éstas están relacionadas por su contenido. Si un conjunto de páginas están vinculadas entre sí, se crea una comunidad de información común.
3. Usabilidad: Son los datos generados por los usuarios en su navegación, ya que los servidores Web almacenan todas las peticiones realizadas por los usuarios en archivos denominados *web logs*.

2.2 Web Mining

Es la aplicación de técnicas de minería de datos a la información producida en la Web. Como esta información se divide en tres ámbitos, a su vez el Web Mining puede ser separado bajo tres áreas [13], [14]:

1. Web Content Mining: Estudia el contenido de las páginas web con el fin de encontrar información, por medio del análisis de texto, imágenes y video.
2. Web Structure Mining: Se refiere a la estructura de los sitios y la Web. Específicamente, las páginas y los *links* están modelados como los nodos y arcos en un grafo direccionado. Un arco parte en el nodo que representa la página que tiene el *link* y termina en la página que está siendo llamada.
3. Web Usage Mining: Se centra en la identificación de patrones que pueden predecir comportamientos o formas de interactuar de los usuarios al momento de utilizar la Web.

2.3 Dilatación pupilar

Para estudiar la dilatación pupilar, antes es necesario analizar el contexto en el que está inmerso todo este sistema, es decir, el sistema visual. El sistema visual es el encargado de proveer al reino animal de uno de los sentidos más importantes, que es la visión.

2.3.1 Sistema Visual

Los principales componentes del sistema visual son los siguientes [15]:

1. Globo Ocular: De forma irregularmente esferoidal, está formado por tres capas, la túnica externa (córnea y esclerótica); túnica media o vascular (úvea, formada por el iris, cuerpo ciliar y coroides); túnica interna (retina). Además, en su interior limitan los compartimientos o cámaras anterior, posterior y vítrea.
2. Córnea: Estructura transparente que proporciona gran parte del poder refractivo necesario para enfocar la luz en la retina. También funciona como estructura de tejidos y humores intraoculares.
3. Esclerótica: Membrana fibrosa, muy resistente que protege los tejidos intraoculares, soporta la tensión de los músculos intraoculares y contribuye a mantener la forma y tono ocular.
4. Iris: Porción más anterior de la úvea. Presenta la forma de un disco perforado en su centro por un orificio circular, la pupila. Está inmerso en el humor acuoso.
5. Pupila: Orificio presente en el iris, encargado de administrar la cantidad de luz que ingresa y penetra en el ojo. Presenta un diámetro variable frente a diferentes estímulos.
6. Cuerpo Ciliar: Desempeña un papel importante en la acomodación, la nutrición del segmento anterior y la secreción del humor acuoso.
7. Coroides: Su riqueza en células pigmentarias le confiere un papel de pantalla a la luz y su naturaleza vascular la hace membrana nutricia del ojo.
8. Retina: Capa más interna del globo ocular. Aquí se inicia el proceso de la visión, siendo la parte especializada del sistema nervioso destinada a recoger, elaborar y transmitir las sensaciones visuales.
9. Vítreo: Gel transparente que ocupa la totalidad del espacio comprendido entre la superficie interna de retina, capa posterior del cristalino y cuerpo ciliar.
10. Cristalino: Lente biconvexa, con poder de convergencia variable.
11. Humor Acuoso: Gel transparente que se encuentra entre la córnea y el cristalino, que sirve para nutrir y oxigenar dichos componentes.

En la Figura 2.1, es posible ver con mayor claridad la ubicación de estos componentes en el ojo.

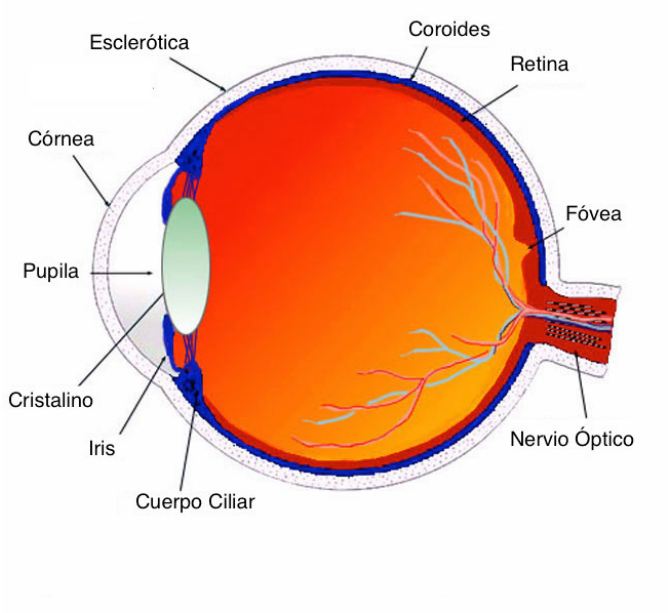

Figura 2.1: Sección Transversal del ojo humano.
Fuente: Imagen obtenida de [16].

2.3.2 Pupila

Es la abertura central en el iris que actúa de modo de diafragma, controla la cantidad de luz que entra en el ojo y es uno de los elementos oculares que mejora la calidad de la imagen que se forma en la retina [17]. El diámetro pupilar resulta del balance entre el músculo esfínter de la pupila y las fibras radiales del iris inervación autónoma. El reflejo pupilar es controlado por el sistema simpático (dilatación) y el parasimpático (contracción).

Por otro lado, con respecto a la dilatación pupilar, existe la pupilometría y la pupilografía. La pupilometría es la medición de los diámetros pupilares en condiciones basales y luego ante diferentes estímulos; la pupilografía es el análisis de dichas respuestas [17]. Para medir estos elementos existen técnicas como por ejemplo el *eye tracking*, utilizado en este trabajo.

Eye Tracking

Es una técnica con la que se miden los movimientos oculares de un individuo, con lo cual es posible conocer hacia qué lugares está dirigiendo su mirada, y la secuencia con que mueve sus ojos. Esta técnica entrega información importante al momento de analizar el comportamiento de los usuarios Web. Los principales métodos de *eye tracking* son los siguientes [12]:

- Electro-oculografía: Se basa en la medición de de la diferencia de potencial de la piel, mediante el uso de electrodos ubicados alrededor de los ojos. Esta técnica mide la

posición relativa de los ojos con respecto a la cabeza, por lo que no es apropiada para calcular el punto de atención.

- Lentes de Contacto Esclerales: Consiste en adjuntar una referencia óptica o mecánica a un lente de contacto para ser usado directamente en el ojo. Aunque es una de las técnicas más precisas, es la más invasiva, causando molestias cuando es usada, ya que se necesita un lente particularmente grande, debido a que debe cubrir la córnea y la esclerótica.

- Foto/Vídeo Oculografía: Consiste en una serie de imágenes o vídeos que guardan los movimientos oculares y que posteriormente son analizados de forma manual o automática.

- Reflejo de la córnea y centro de la pupila basado en vídeo: Esta técnica es la más usada actualmente. Radica en el uso de una cámara infrarroja montada bajo el monitor de un computador con un software de procesamiento de imágenes, para ubicar e identificar el reflejo de la córnea y el centro de la pupila. Gracias a esto, es posible separar los movimientos oculares de la posición de la cabeza, para calcular los focos de atención de los usuarios.

Existen técnicas similares para evaluar la pupilometría. Por ejemplo, se puede usar una cámara infrarroja al igual que en el caso descrito anteriormente, con la que se puede medir el diámetro de la pupila. De la misma manera se puede medir la dilatación pupilar por medio de análisis de imágenes con la ayuda de un software especializado.

2.4 Bioactividad Cerebral

Para hablar acerca de la actividad bioeléctrica cerebral, es necesario definir básicamente el funcionamiento del cerebro y las neuronas, y cómo se generan los potenciales eléctricos. Posteriormente se exhibe una técnica muy típica para medir estos potenciales que es el electroencefalograma (EEG) y cómo obtener información dadas estas mediciones.

2.4.1 El Cerebro y las Neuronas

El cerebro es el principal órgano del sistema nervioso [18]. Se encarga de llevar a cabo el procesamiento de la información que se recibe por medio de los sentidos dados distintos tipos de estímulos. La corteza cerebral está compuesta por cuatro lóbulos, frontal, parietal, temporal, occipital. Es en esa corteza donde las habilidades cognitivas ocurren. Cada lóbulo se asocia a un procesamiento específico. Además el cerebro está separado en dos hemisferios,

izquierdo y derecho. Por último, el cerebro está dividido en dos tipos de materias, la blanca y al gris. La materia blanca está compuesta de axones y la materia gris está compuesta de somas neuronales.

Por otro lado, las neuronas son las células diferenciadas del sistema nervioso central. Sus principales componentes son:

- Soma: Estructura correspondiente al cuerpo central de una neurona. Contiene el núcleo de esta célula y es donde se llevan a cabo los procesos internos.
- Dendritas: Son extensiones celulares con muchos brazos, donde se reciben señales desde otras neuronas.
- Axón: Es una proyección de la neurona, que sirve para llevar señales desde el soma hacia otras neuronas.
- Axón terminal: Árbol terminal del axón donde se produce la sinapsis con las neuronas post-sinápticas.

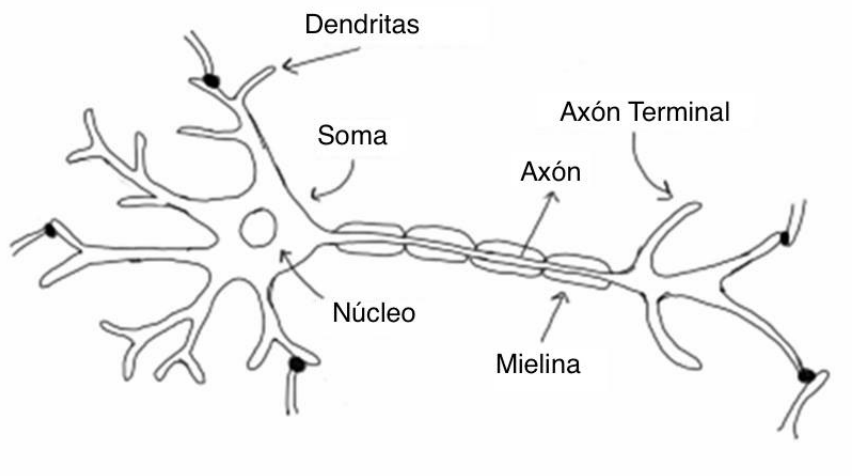

Figura 2.2: Estructura de una neurona.
Fuente: Imagen adaptada de [18].

2.4.2 Potenciales Eléctricos

Los potenciales eléctricos ocurren de la siguiente manera [19]: una neurona es activada por otras neuronas a través de potenciales de acción, produciendo potenciales post sinápticos excitatorios que van hacia las dendritas de aquella neurona. Estas dendritas empiezan a despolarizarse, quedando cargadas negativamente, en comparación con el soma de la neurona.

Como consecuencia de esta diferencia de potencial, la corriente fluye desde el soma no excitado hacia el árbol de dendritas y emerge de la superficie una polaridad negativa. En el caso opuesto, si el soma está excitado, la corriente fluye en caso contrario.

Estas diferencias de potencial y flujos de corriente bioeléctrica cerebral, pueden ser medidos desde la cabellera de los individuos, gracias a la orientación de las células ubicadas en la corteza cerebral. Las células son paralelas unas con otras y perpendiculares en la corteza, permitiendo la suma de potenciales y la propagación de los mismos a través de la superficie de la cabellera [20].

2.4.3 Electroencefalograma

Corresponde a una medición de las diferencias de potencial eléctricos sobre la corteza cerebral. Las mediciones se hacen a través de electrodos colocados en el cuero cabelludo y otras áreas sensibles, como cerca de los ojos y orejas. Cada electrodo mide la diferencia de potencial con respecto a otro electrodo de referencia [21].

Existe una distribución estándar de los electrodos que permite generar estudios comparativos [19]. En la Figura 2.3 se puede ver claramente la ubicación de los electrodos bajo el sistema internacional 10-20, junto a sus respectivos nombres para un sistema de 32 electrodos.

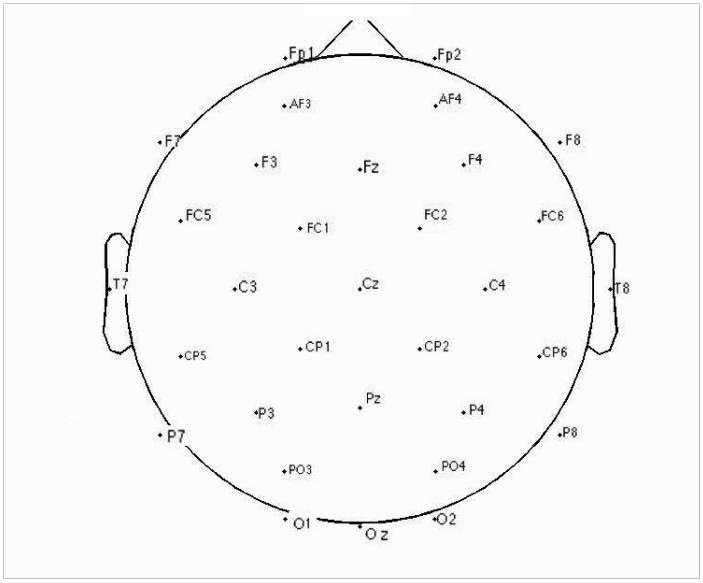

Figura 2.3: Distribución de electrodos según sistema internacional 10-20.
Fuente: Elaboración propia.

El resultado del EEG es una serie de oscilaciones, cada una con distinta frecuencia. A

continuación se describen las distintas bandas de frecuencias para un EEG [22]:

- Ondas Delta: Oscilaciones de menor frecuencia presentes en el EEG. Tienen un rango de 1 a 4 Hz. Se relacionan al sueño en humanos sanos y a patologías neurológicas.
- Ondas Theta: Oscilaciones cuyo rango va de 4 a 8 Hz. Predominan durante el sueño.
- Ondas Alpha: Ondas con rango entre los 8 y los 13 Hz. En adultos normales generalmente tienen amplitudes entre 10 y 45 μV y pueden ser medidas fácilmente durante estados de reposo.
- Ondas Beta: Oscilaciones de poca amplitud con rango de frecuencia de 13 a 30 Hz. Están presentes en actividades cognitivas con amplitudes de 10 a 20 μV.
- Ondas Gamma: Son las de mayor frecuencia en el EEG, de 36 a 44 Hz. Se relacionan con atención, excitación, reconocimiento de objetos y procesos sensoriales.

2.5 Emocionalidad

Las emociones humanas pueden ser creadas como un resultado del proceso de pensamiento interno relativo a la memoria o por estímulos al cerebro a través de los sentidos (visual, auditivo, olfativo, gusto, etcétera).

Existen diferentes clasificaciones para las emociones propuestas por los investigadores. Una buena forma para clasificarlas es bajo dos dimensiones, las de excitación y valencia emocional [23].

- Valencia Emocional: Representa el grado de emocionalidad que conlleva un estímulo, variando el nivel de placer producido, desde desagradable, pasando por neutro hasta agradable.
- Excitación Emocional: Corresponde a un nivel cuantitativo de activación que va desde no excitado a excitado.

Una excitación en particular no se corresponde con un estado emocional en particular, pero generalmente es mayor cuando se está en presencia de valencias emocionales muy bajas o muy altas.

En la literatura es posible encontrar variados estudios que intentan relacionar la bioactividad cerebral y la emocionalidad. Por un lado están los que lo hacen desde la perspectiva médica, y por otro, aquellos que utilizan herramientas de minería de datos para obtener conclusiones valiosas.

Se tienen por ejemplo, desde el ámbito médico, trabajos como [24], donde se muestra que cuando un grupo de sujetos es expuesto a estímulos visuales de distinta valencia emocional, sus potenciales positivos tardíos (300 - 700 ms) en ERP centro-parietales son más grandes para imágenes con contenido emocional positivo o negativa, mientras que para imágenes neutras son menores [19].

Otro estudio es [25], en el que se llega a resultados similares a [24] en cuanto a las diferencias en los potenciales positivos tardíos, aunque estas diferencias no son estadísticamente significativas. Además, en [26] se concluye que existen diferencias entre los ERP producidos por imágenes con contenido emocional positivo y negativo.

Desde el punto de vista de la minería de datos se han realizado variados estudios que pretenden clasificar estados de ánimo de sujetos mediante de la aplicación de algoritmos y técnicas a las señales de EEG. En la tabla 2.1 se presenta un resumen de algunos estudios y sus resultados.

Autor	Característica y Clasificación	Emoción	Resultado
Ishino y Hagiwara 2003 [27]	**Característica**: FFT, Transformada Wavelet, Varianza, promedio **Clasificación**: Redes neuronales	Alegría, tristeza, enojo y relajo	Alegría: 54.5% Enojo: 67.7% Tristeza: 59% Relajo: 62.9%
Takahashi 2004 [28]	**Característica**: Rasgos estadísticos **Clasificación**: SVM, Redes neuronales	Alegría, tristeza, enojo, miedo y realización	41.7% para 5 emociones, 66.7% para 3 emociones
Chanel et al. 2006 [29]	**Característica**: 6 bandas de frecuencias desde diferentes canales **Clasificación**: Naïve Bayes, Análisis Discriminante	3 grados de excitación emocional	58%
Chanel et al. 2009 [30]	**Característica**: Transformada de Fourier **Clasificación**: Análisis Discriminante, SVM	Positivo , neutro y negativo	63%
Lin et al. 2009 [31]	**Característica**: ASM 12 **Clasificación**: SVM	Alegría, tristeza, enojo y relajo	90.72%
Khalili y Moradi 2009 [32]	**Característica**: rasgos estadísticos y dimensiones de correlación **Clasificación**: Análisis discriminante cuadrático	Positivo, negativo y neutro	76.6%
Zhang y Lee 2009 [33]	**Característica**: PCA **Clasificación**: SVM con Kernel lineal, SVM con Kernel RBF	Negativo y positivo	73%
Liu y Sourina 2010 [**?**]	**Característica**: Dimensión Fractal **Clasificación**: SVM	valencia y excitación emocional	84.9% para valencias, 90% para excitación emocional
Petrantonakis y Hadjileontiadis 2010 [34]	**Característica**: rasgos estadísticos, Transformada Wavelet **Clasificación**: SVM, QDA, KNN	Alegría, tristeza, enojo, miedo, disgusto y sorpresa	62.3% para un canal, 83.3% para canales combinados

Tabla 2.1: Resumen estudios de emocionalidad con minería de datos.
Fuente: Tabla adaptada de [35]

2.6 Proceso *Knowledge Discovery in Databases* (KDD)

Proceso de extracción de información utilizando como fuente datos almacenados en repositorios de información [2]. Fue definido por Fayyad et al [36] como el "proceso no trivial de identificación de patrones válidos, originales, útiles y entendibles acerca de la data". Data se puede definir como un conjunto de hechos, y patrón es una expresión en algún lenguaje que describe un subconjunto de la data o un modelo aplicable al subconjunto.

El proceso KDD es una método interactivo e iterativo que consiste de 5 pasos, definidos a continuación:

1. Selección de datos: Consta de la extracción y selección de datos a utilizar desde una base de datos, la cual dependerá del tipo de problemática a resolver.
2. Pre-procesamiento de datos: Consiste en la limpieza de los datos, detectando los valores fuera de rango, faltantes o nulos.
3. Transformación de datos: Se trata de la modificación de los datos que se ingresarán al modelo.
4. Data Mining: Consiste en la selección apropiada de la tarea y de los algoritmos de minería de datos y su implementación para obtener los patrones subyacentes en la data que está siendo analizada.
5. Evaluación e interpretación: Es aquí donde se obtiene el conocimiento a partir de los datos luego de pasar por las etapas anteriores.

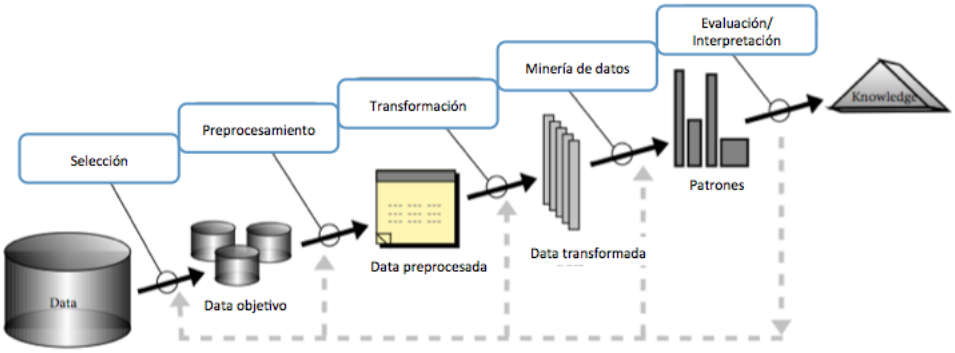

Figura 2.4: Diagrama del proceso KDD.
Fuente: Imagen adaptada de [36].

2.6.1 Técnicas de Minería de datos

Existen varios tipos de técnicas de minería de datos que se utilizan, dependiendo del propósito y los objetivos a los que se apunta. Básicamente se dividen en dos grandes áreas, las orientadas a verificación, que pretenden validar hipótesis, y las orientadas a descubrimiento, para encontrar patrones y reglas [2].

Los métodos orientados a descubrimiento identifican patrones de forma autónoma. Se dividen en métodos predictivos y métodos descriptivos. Los métodos descriptivos están orientados a la interpretación de la data, cuyo enfoque está en el entendimiento de la forma en que los datos se relacionan con sus partes. Por otro lado, los métodos predictivos intentan crear un modelo de comportamiento de forma automática, que obtenga nuevos ejemplos y sea capaz de predecir valores de una o más variables relacionadas con el ejemplo. Además genera patrones que facilitan el entendimiento del conocimiento descubierto. A su vez, los métodos predictivos se dividen en algoritmos de clasificación y regresión. Los algoritmos de clasificación tienen como objetivo clasificar casos futuros, mientras que los de regresión generan un pronóstico a partir de los datos.

Los métodos orientados a la verificación, tienen como misión validar hipótesis planteadas por un agente externo. Generalmente se ocupan métodos típicos de estadística tradicional, como bondad de ajuste, análisis de varianzas (ANOVA), y tests de hipótesis. Esta clasificación puede ser fácilmente entendida en la Figura 2.5.

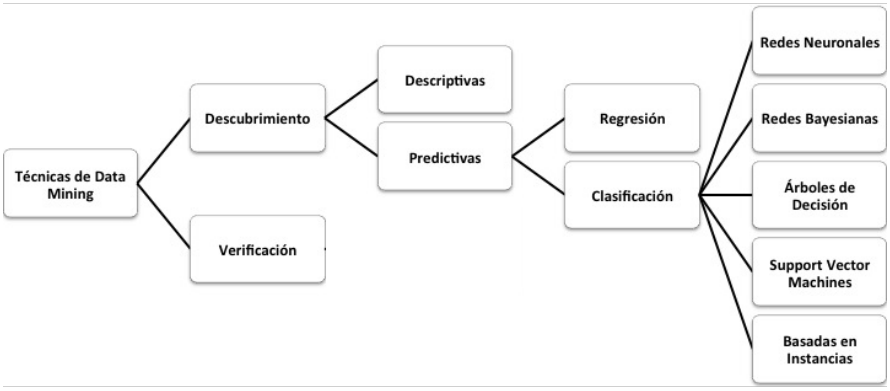

Figura 2.5: Técnicas de minería de datos.
Fuente: Imagen adaptada de [2].

Es posible también distinguir los algoritmos de minería de datos entre supervisados y no supervisados. Los métodos no supervisados modelan la distribución de instancias, donde estas no están previamente identificadas. Los algoritmos supervisados intentan descubrir las relaciones que existen entre variables o atributos independientes y una variable o atributo dependiente. Esta relación se modela como patrones de comportamiento que pueden predecir valores en base a entradas de datos [19].

2.6.2 Algoritmos de Minería de Datos

En esta sección se entrega una breve descripción de los modelos de minería de datos para clasificación.

***Support Vector Machine* [37], [19]**

SVM es un algoritmo supervisado de clasificación en forma binaria. Es uno de los más usados debido a su alto grado de precisión sumado a su fuerte base teórica.

Para su formulación se supone el set de entrenamiento $\{(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), ..., (\mathbf{x_n}, y_n)\}$, donde $\mathbf{x_i}$ son los vectores de datos e y_i es la etiqueta de clase, que toma el valor 1 o -1, dependiendo si pertenece o no a la clase en estudio. También puede llamarse la clase positiva o negativa respectivamente.

SVM busca la función lineal de la forma de la ecuación 2.1, de tal manera que si el vector $\mathbf{x_i}$ es de la clase positiva, entonces $f(\mathbf{x_i}) \geq 0$ y para la clase negativa el caso contrario, tal como lo muestra la ecuación 2.2. El parámetro $\mathbf{w}$ es denominado vector de pesos y b es llamado bias.

$$f(\mathbf{x}) = \langle \mathbf{w} \cdot \mathbf{x} \rangle + b \tag{2.1}$$

$$y_i = \begin{cases} 1 & \text{si } \langle \mathbf{w} \cdot \mathbf{x} \rangle + b \geq 0, \\ -1 & \text{si } \langle \mathbf{w} \cdot \mathbf{x} \rangle + b \leq 0. \end{cases} \tag{2.2}$$

En esencia, SVM busca un hiperplano que separe a los datos de entrenamiento de clases positivas y negativas. Este hiperplano es llamado la frontera de decisión y se puede visualizar en la figura 2.6.

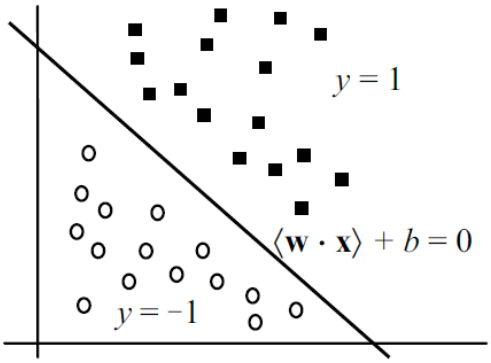

Figura 2.6: Frontera de decisión en SVM.
Fuente: Imagen adaptada de [37].

Para encontrar el superplano que separa los datos de entrenamiento en dos clases, el algoritmo maximiza el margen entre los datos de distinta clase. Esto puede ser expresado como el problema 2.3. Este problema de optimización es la condición ideal para SVM, ya que supone la existencia de un hiperplano que separe correctamente la data.

$$\begin{aligned} Min: \quad & \tfrac{\langle \mathbf{w} \cdot \mathbf{w} \rangle}{2} \\ s.a.: \quad & y_i(\langle \mathbf{w} \cdot \mathbf{x} \rangle + b) \geq 1 \end{aligned} \tag{2.3}$$

Si no existe tal hiperplano, el problema se relaja, quedando formulado como 2.4.

$$\begin{aligned} Min: \quad & \tfrac{\langle \mathbf{w} \cdot \mathbf{w} \rangle}{2} + C \sum_{i=1}^{n} \xi_i \\ s.a.: \quad & y_i(\langle \mathbf{w} \cdot \mathbf{x} \rangle + b) \geq 1 - \xi_i \\ & \xi_i \geq 0. \end{aligned} \tag{2.4}$$

Para problemas no lineales, el método SVM transforma el set de datos con funciones llamadas *Kernel* (Φ). La idea es convertir el espacio de entrada X en un espacio de características F como lo muestra la Figura 2.7.

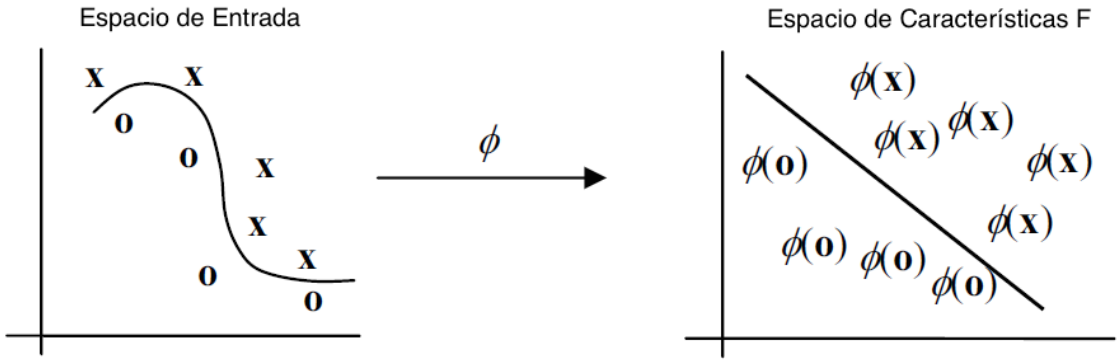

Figura 2.7: Transformación de un espacio de entrada a un espacio de características.
Fuente: Imagen adaptada de [37].

Con esta transformación del espacio, el problema queda formulado según 2.5.

$$\begin{aligned} Min: \quad & \tfrac{\langle \mathbf{w}\cdot\mathbf{w}\rangle}{2} + C\sum_{i=1}^{n}\xi_i \\ s.a.: \quad & y_i(\langle \mathbf{w}\cdot\Phi(\mathbf{x})\rangle + b) \geq 1-\xi_i \\ & \xi_i \geq 0. \end{aligned} \tag{2.5}$$

Si se desea que SVM clasifique para varias clases, se utiliza la técnica de uno versus el resto.

Redes Neuronales [38], [19]

Es un modelo matemático que genera predicciones basado en variables descriptivas de entrada. Se utiliza un conjunto de entrenamiento para generalizar las relaciones entre las variables de entrada y las variables de respuesta.

Una red neuronal comprende una serie de nodos independientes que están conectados a otros nodos y organizados en capas como se muestra en la figura 2.8. En este ejemplo se puede apreciar como los nodosse organizan en tres capas. La capa de entrada contiene el conjunto de nodos de la A a la F, y es donde se ingresan los datos. Luego viene la capa escondida con los nodos de la G a la J, donde se realizan los cálculos. Finalmente está la capa de salida (nodos K y L), que es donde se entrega el resultado final de la red. Cabe destacar que en el ejemplo se aprecia solo una capa escondida, pero en la práctica pueden existir múltiples. Además, en general, cada nodo está conectado a todos los nodos de las capas adyacentes a la suya.

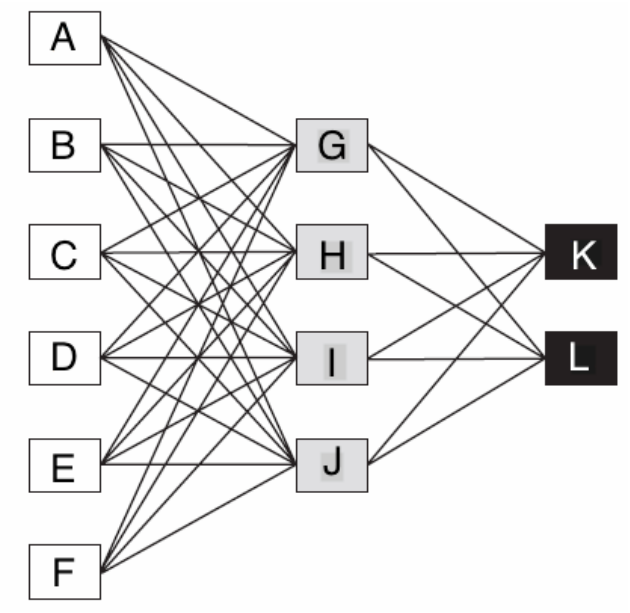

Figura 2.8: Ejemplo de una red neuronal.
Fuente: Imagen adaptada de [38].

Cada nodo utiliza los valores de la capa previa para realizar un cálculo y generar un nuevo valor que sirve de *input* para los nodos de la siguiente capa. La ecuación 2.6 presenta una forma típica para calcular los valores dentro de un nodo. La función g se denomina función de activación, siendo las más comunes las de las ecuaciones 2.7 y 2.8. Los parámetros w son otorgados por el entrenamiento de la red y cada nodo posee sus propios valores. Los valores I corresponden a los valores otorgados por los nodos previos.

$$Valor\ nodo = g(\sum_{j=1}^{n} I_j w_j) \tag{2.6}$$

$$g(x) = \frac{1}{1 + \exp^{-x}} \tag{2.7}$$

$$g(x) = \frac{\exp^{x} - \exp^{-x}}{\exp^{x} + \exp^{-x}} \tag{2.8}$$

Para entrenar los parámetros de los nodos de la red se sigue un proceso conocido como *backpropagation*. Este proceso consiste en entregar a la red observaciones del conjunto de entrenamiento seleccionadas al azar. Luego la red neuronal calcula la predicción para estas entradas y evalúa el error con respecto al valor que el set de entrenamiento entrega. Finalmente, los parámetros de los nodos se ajustan para disminuir el error de la red. Para generar un clasificador de muchas clases, basta con tener tantos nodos de salida como categorías se quieran clasificar.

Árboles de Decisión [38], [3]

Corresponden a un método de clasificación expresado como una serie de puntos de decisión a base de ciertas variables. Un árbol de decisión está compuesto por nodos y ramificaciones. Cada nodo representa un conjunto de observaciones que son clasificadas según algún criterio en dos o más subconjuntos que forman nuevos nodos. El nodo que es separado es el nodo-padre, y los subconjuntos son los nodos hijos. Un nodo que no tiene ramificaciones es un nodo hoja. La imagen 2.9 muestra un ejemplo simple de árbol de decisión.

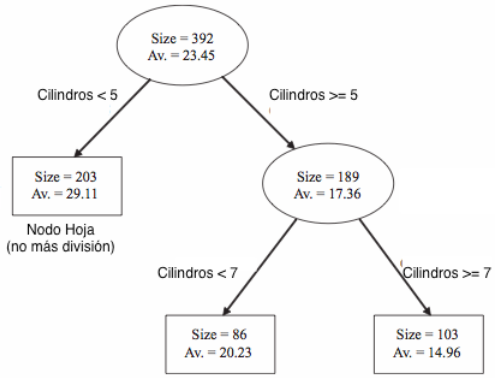

Figura 2.9: Ejemplo de árbol de decisión.
Fuente: Imagen adaptada de [38].

Un algoritmo bastante usado para el entrenamiento de modelos de árboles de decisión es el método del crecimiento y poda. Este método está construido de manera recursiva, donde en cada iteración el algoritmo considera el resultado de la función de los atributos de entrada, elige la mejor y selecciona una partición apropiada. Esto se repite hasta que las nuevas particiones no generen una ganancia o hasta que se dé un tipo de detención previamente definido por algún criterio.

Regresión Logística [39], [19]

Es un algoritmo predictivo de clasificación derivado de las regresiones lineales. Este tipo de regresiones estiman la probabilidad de que un caso sea positivo o negativo, es decir que pertenezca o no a la clase en estudio. Por lo tanto, el modelo que subyace a la regresión logística sólo permite valores entre 0 y 1.

Una regresión logística representada por h_θ puede ser formulada con las ecuaciones 2.9 y 2.10. Acá, x representa el caso particular que se quiere clasificar; θ representa los parámetros de la regresión logística; g es la función logística o sigmoidea, cuyos valores están entre 0 y 1.

$$h_\theta(x) = g(\theta^T x) \tag{2.9}$$

$$g(z) = \frac{1}{1 + \exp^{-z}} \tag{2.10}$$

Para el entrenamiento de los parámetros θ del modelo, se establece una función de costos, para que al ser minimizada se encuentren los parámetros que se ajusten al conjunto de datos de entrenamiento. La función de costos de una regresión logística para un set de entrenamiento $\{(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), ..., (\mathbf{x_m}, y_m)\}$ se define como la ecuación 2.11. $\mathbf{x_i}$ * son los vectores de datos e y_i es la etiqueta de clase que puede tomar los valores de 0 o 1, es decir, pertenece o no a la clase en estudio. La cantidad de parámetros θ se fija de acuerdo al problema.

$$J(\theta) = -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} log h_\theta(x^{(i)}) + (1 - y^{(i)}) log(1 - h_\theta(x^{(i)}))] \tag{2.11}$$

Es común que la regresión se encuentre sobre ajustada a los datos de entrenamiento, por lo que se utiliza un parámetro de regularización λ. La ecuación 2.12 muestra la función de costos incluyendo este parámetro.

$$J(\theta) = -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} log h_\theta(x^{(i)}) + (1 - y^{(i)}) log(1 - h_\theta(x^{(i)})) + \frac{\lambda}{2m} \sum_{j=1}^{n} \theta_j^2] \tag{2.12}$$

2.6.3 Evaluación de Clasificadores [2], [3]

Existen diversos métodos para evaluar los clasificadores mencionados anteriormente, entre los que se encuentran:

- *Accuracy*: Habilidad del modelo clasificador de predecir correctamente la clase en la que los datos nuevos son clasificados.
- Velocidad: Se refiere a los costos computacionales en procesamiento y tiempo para entrenar el modelo.
- Robustez: Es la habilidad del modelo para manejar ruido o valores faltantes en los datos y hacer predicciones correctas.

- Escalabilidad: La habilidad del modelo de entrenar el clasificar de forma eficiente para un conjunto de datos más grande.
- Interpretabilidad: Se refiere al nivel de entendimiento o información que entrega el modelo construido.

En algunos casos el *Accuracy* no entrega resultados satisfactorios por lo que existen otros indicadores como alternativa. Si se asume que el clasificador entrega valores como positivos y negativos, se pueden utilizar las siguientes alternativas:

- *Sensitivity*: Evalúa qué tan bien el clasificador reconoce a las muestras positivas y es definido como:

$$Sensitivity = \frac{verdaderospositivos}{positivos} \tag{2.13}$$

  donde *verdaderospositivos* corresponde al número de las muestras positivas verdaderas y *positivos* es el número de muestras positivas totales.
- *Specifity*: Mide qué tan bien reconoce las muestras negativas; está definido como:

$$Specifity = \frac{verdaderosnegativos}{negativos} \tag{2.14}$$

- *Precision*: Medida de evaluación del porcentaje de muestras clasificadas como positivas que realmente son positivas, se define como:

$$Precision = \frac{verdaderospositivos}{verdaderospositivos + verdaderosnegativos} \tag{2.15}$$

Con las definiciones de 2.13 y 2.14 se pueden relacionar con el *Accuracy* de la siguiente manera:

$$Accuracy = Sensitivity \cdot \frac{positivos}{positivos + negativos} + Specifity \cdot \frac{negativos}{positivos + negativos} \tag{2.16}$$

Existen casos en que los datos de muestra no pertenecen a solo una clase. Por lo que es necesario trabajar con clasificadores difusos, que entregan una probabilidad de pertenecer a cada clase.

2.7 Website Keyobject

El tema central de este trabajo de memoria es la identificación y clasificación de los *Website Keyobjects*, por lo que es necesario definirlos y mostrar el método desarrollado para encontrarlos.

2.7.1 Definición

Velásquez y Dujovne definieron en [1] un *Web Object* como ″un grupo estructurado de palabras o contenido multimedia, que está presente en una página Web y que posee metadatos que describen su contenido″. Los meta datos son fundamentales en la construcción del vector que representa la el contenido de la página y en la comparación de distintos archivos multimedia entre sí.

Además definieron los *Website Keyobjetcs* como "uno o un grupo de *Web Objects* que atraen la atención del usuario y que caracterizan el contenido de una página o sitio web". Con esta definición se da a conocer el contenido y formato que más interesan a los usuarios, por lo tanto, identificarlos puede resultar bastante útil para mejorar un sitio en su estructura y contenidos.

En la Figura 2.10 es posible ver una simple distinción entre los distintos objetos presentes (no todos) en la página de inicio del portal académico de U-Cursos. Claramente se distinguen el logo de U-Cursos, los formularios de inicio de sesión y algunas imágenes. Cada uno tiene diferentes formatos, por lo que para compararlos se requiere de la incorporación de meta datos para definir su contenido [4].

Figura 2.10: Ejemplo de algunos *Website Objects* presentes en el sitio web de U-Cursos. Fuente: Elaboración propia.

2.7.2 Representación

Los *Web Objects* deben ser descritos mediante meta datos, por lo que en [1] se asoció a cada objeto un documento XML[1], compuesto por meta datos que describen el contenido y la página a la que pertenecen. Además en la página Web también se establece relación entre el objeto y el documento XML, mediante tags de HTML [7].

Los meta datos fueron guardados con el siguiente formato:

- Identificador de la página
- Objeto: Identificador; formato; concepto(s).

2.7.3 Metodología para la identificación de Website Keyobjects

La metodología propuesta por Dujovne y Velásquez en [1], mejorada por González en [7] y posteriormente por Martínez en [4], para identificar los *Website Keyobjects*, se basa en dos procesos principales: Transformación de datos y aplicación de algoritmos de clustering [4].

Transformación de datos

Los datos utilizados provienen de diversas fuentes, por lo que es necesario llevar a cabo los siguientes procesos de transformación:

Sesionización

El objetivo de esta etapa es completar la secuencia de páginas visitadas por los distintos usuarios Web en el sitio. Adicionalmente, esta secuencia debe contener los tiempos de permanencia en las páginas por cada usuario, durante su sesión.

Incorporación de Metadatos

Primero se debe identificar los objetos que componen las páginas del sitio. Luego se tendrán que definir los conceptos que los describen, para finalmente almacenar esta información en una base de datos. El levantamiento de estos conjuntos debe hacerse junto al *webmaster*, para asegurar que los conceptos reflejen de buena manera el contenido de los objetos.

[1]XML: *Extensible Markup Language*, lenguaje estándar que define reglas para codificar documentos en un formato entendible para personas y máquinas.

Tiempo de permanencia en los objetos

Posterior a la definición de los objetos, en [1] se propone realizar una encuesta para conocer qué cantidad de tiempo de atención le otorgaba cada usuario a los objetos del sitio web.

En [14], se define una nueva forma de medición del tiempo de permanencia en cada objeto, utilizando técnicas de *eye tracking*. Debido a que cada objeto está relacionado con un grupo de pixeles, es posible determinar cuánto tiempo el usuario estuvo mirando cada uno de ellos. De esa manera, en lugar de hacer una encuesta, se lleva a cabo un experimento donde los usuarios recorren un sitio web, mientras que se mide el tiempo de permanencia en cada objeto presente.

Vector de comportamiento del usuario Finalmente, para cada sesión identificada, se seleccionan los n objetos que capturan en mayor medida la atención del usuario, definiendo el *Important Object Vector* (IOV), expresado según la ecuación 2.17.

$$v = [(o_1, t_1), ..., (o_n, t_n)] \tag{2.17}$$

Donde o_i representa el objeto y t_i representa el tiempo de permanencia en cada página.

Medidas de Similitud

Una vez que la limpieza y transformación de los datos está lista, el siguiente paso es clusterizar las sesiones de los usuarios, representadas por los IOVs. Para ejecutar los algoritmos es necesario definir una medida de distancia o similitud entre estos vectores.

En [1] se calcula la similitud entre dos IOV, α y β, usando la ecuación 2.18, donde *do* está definido como se muestra en la ecuación 2.19 y sirve para comparar dos objetos distintos y τ_k^α el tiempo empleado por un usuario α viendo el objeto k.

$$st(\alpha, \beta) = \frac{1}{i} * (\sum_{k=1}^{i} min(\frac{\tau_k^\alpha}{\tau_k^\beta}, \frac{\tau_k^\beta}{\tau_k^\alpha}) * do(o_k^\alpha, o_k^\beta)) \tag{2.18}$$

$$do(O_1, O_2) = 1 - \frac{L(O_1, O_2)}{max\{|O_1|, |O_2|\}} \tag{2.19}$$

Finalmente, con lo anterior es posible utilizar distintos algoritmos de clustering para encontrar los objetos más importantes. En [1], [7] y [4] se usa *Self Organizing Feature Maps*, *K-means* y *Association Rules*.

En el estudio más reciente [4], se obtiene como resultado una lista de *Website Keyobjects* clasificados con un 80% de precisión.

2.7.4 Algoritmos de Clustering

Como mencionado anteriormente, en los trabajos de Velásquez et al. se han utilizado diversos algoritmos de *clustering* para agrupar los vectores de comportamiento de los usuarios, entre los cuales se encuentran SOFM (*Self Organizing Feature Maps*), *K-Means* y *Association Rules*. A continuación se entrega una descripción de cada algoritmo y su relación y utilización en la metodología en los estudios de [1], [7] y [4].

SOFM **[4], [38]**

Los mapas autoorganizados de características corresponden a un tipo de red neuronal artificial que utiliza un algoritmo de aprendizaje no supervisado que permite la visualización de elementos con una alta dimensionalidad en una grilla de pocas variables (generalmente dos), llamado mapa. Se diferencian de las redes neuronales típicas en cuanto a la utilización de una función de vecindad para preservar las propiedades topológicas del espacio de entrada.

Cada neurona de la red está asociada a un vector de pesos de igual dimensión que el vector de entrenamiento y a una posición relativa dentro de la grilla. Además se le asigna un conjunto de neuronas llamado vecindad. Este conjunto define el tipo de topología de la red, siendo las rectangulares y hexagonales las tipologías más comunes. Un caso particular es el de las topologías toroidales, donde la primera neurona tiene como vecinos la neurona de la derecha, la de abajo, la última de su columna y la última de su fila (ver imagen 2.11). Los vecinos de las demás neuronas se definen de manera análoga.

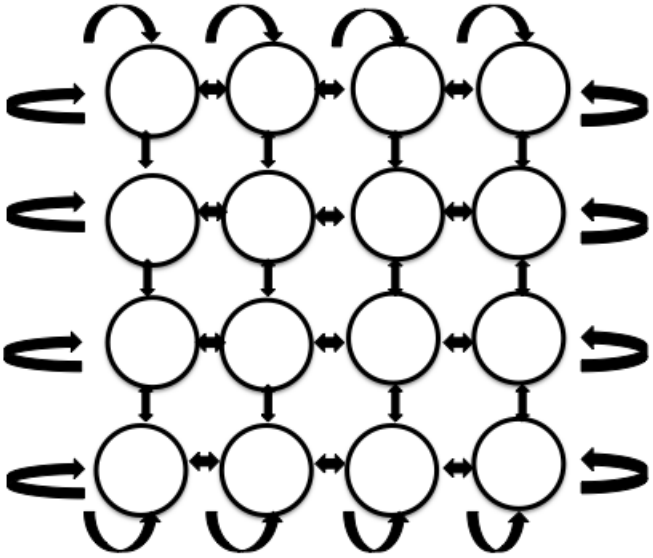

Figura 2.11: Ejemplo de una red con tipología toroidal.
Fuente: Imagen adaptada de [4].

Las vecindades definen una topología cuyo radio de influencia se va aminorando con cada iteración del algoritmo, lo que significa que al inicio del algoritmo los patrones generan muchos cambios a la neurona ganadora y a sus vecinos, pero a medida que el proceso siga su curso, las modificaciones se aplicarán solo a los vecinos más cercanos, para que al final del algoritmo se afectará solo a las neuronas ganadoras.

EL algoritmo de entrenamiento de la red tiene una naturaleza competitiva, pues todas las neuronas compiten entre sí para definir cuál es la más parecida al ejemplo. Cuando se encuentra a la ganadora, se arregla la red de manera que las neuronas ganadoras se vuelvan similares a los ejemplos presentados.

En el ámbito de los *Website Keyobjects* no es posible utilizar directamente este algoritmo para agrupar las sesiones, dada la naturaleza de la data. Para usarlo, cada neurona se define como un IOV, y para actualizarlas durante el entrenamiento, se utiliza la medida de similitud anteriormente descrita. Entonces, para cada IOV se busca la neurona más parecida y luego se actualiza el peso neto de acuerdo a las distancias calculadas. Este es un proceso iterativo que tiene fin una vez que los pesos netos son menores a ϵ.

***K-Means* [4], [38]**

El algoritmo K-Medias permite el agrupamiento de conjuntos de datos en K *clusters*, donde los objetos presentes en cada *cluster* son más similares a los de su mismo grupo que

a los elementos de los demás grupos. Es decir, son homogéneos con los de su grupo, pero heterógeneos con los otros.

Es un algoritmo con aprendizaje supervisado, pues el usuario debe elegir el número K de *clusters* que se crearán. Además, funciona con un enfoque *Top-Down*, en el que se comienza con un número de grupos dado y se van asignando los patrones a cada uno, donde cada objeto se asigna a solo un *cluster*. Es un algoritmo *Two-Phase*, ya que en cada iteración existe una fase inicial de asignación de elementos a los grupos y posteriormente se calculan los centroides correspondientes a cada uno.

A continuación se encuentran los pasos del algoritmo:

1. Iniciación de *clusters*: Tomando todos los elementos en el conjunto de datos como fuente, se seleccionan K aleatoriamente y se asignan a un *cluster*, fijándolos además como sus centroides. Posteriormente, los otros elementos se comparan con estos centroides y se asignan al *cluster* donde esté el centroide con la menor distancia a cada uno de ellos.
2. Cálculo de nuevos centroides: Para cada *cluster*, se compara la distancia entre el centroide y los demás elementos que hayan sido asignados a ese grupo. El que tenga la menor distancia será fijado como nuevo centroide.
3. Reasignación de elementos a *clusters*: Para cada elemento se compara la distancia a los nuevos centroides, asignándolos a nuevos grupos donde las distancias sean menores.

Este procedimiento se itera hasta que ya no existan movimientos de los elementos entre los *clusters*, o hasta que el error por iteración se mantenga estable, es decir, que más de una iteración el error no varíe.

***Association Rules* [4], [38]**

Las reglas de asociación corresponden a un método de agrupamiento no supervisado, que pretende entender o determinar los distintos *links* o asociaciones entre las variables y atributos del conjunto de observaciones. Las reglas de este método siguen el siguiente patrón:

SI variable1=a **Y** variable2=b **ENTONCES** variable3=c

Este método tiene variadas ventajas, entre ellas se encuentra la facilidad de interpretación, la posibilidad de acción frente a las reglas encontradas, y la capacidad para operar con grandes cantidades de datos.

Por otro lado, existen algunas limitaciones como permitir sólo el uso de variables categóricas, lo que puede ocasionar altos consumos de tiempo de procesamiento, y generar muchas reglas que deben ser priorizadas e interpretadas.

Utilizando este método, es posible crear valiosas reglas a base del agrupamiento de los datos, extracción de diversas reglas de los grupos, y luego priorizando aquellas reglas encontradas.

Para medir qué tan buena es una regla, se utilizan tres indicadores: *support*, *confidence* y *lift*.

- *Support*: Permite ver qué porcentaje de las observaciones se encuentran dentro del conjunto de datos.
- *Confidence*: Las reglas se pueden separar en dos partes, la condicional (SI e Y) y la causal (ENTONCES). En cada parte se puede medir el soporte que poseen. La confianza de la regla indica cuán predecible es. Se calcula mediante la división entre el soporte del grupo y el soporte de la parte condicional.
- *Lift*: Sirve para ver la fortaleza de la relación entre la parte condicional y la parte causal. Si es mayor que 1, indica una asociación positiva, en caso contrario, la asociación entre las partes es negativa. Se calcula dividiendo la confianza por el soporte de la parte causal.

Para encontrar reglas en el conjunto de datos, generalmente se utiliza el algoritmo *A priori*, definido en [40] y que busca encontrar subconjuntos que tengan un cierto nivel mínimo de confianza C, definido por el usuario.

El algoritmo utiliza un enfoque *Bottom-Up*, donde los subconjuntos son incrementados de a un elemento por cada iteración, y se prueba que cumplan con las reglas contrastándolos con los datos originales. Además, utiliza una estructura de árbol para almecenar para almacenar los candidatos lo que permite aplicar algoritmos de *pruning*, aumentando la eficiencia computacional del algoritmo.

3 Propuesta de Investigación

Este capítulo muestra al lector la propuesta formal de investigación con la que se busca validar la hipótesis planteada como base. Se pretende exhibir la justificación de cada uno de los supuestos e instrumentos utilizados en el presente trabajo. Esta propuesta está separada en cuatro secciones según [41]. En la primera se muestra el tipo de investigación al que corresponde este trabajo, luego se propone el diseño de la investigación propiamente tal, posteriormente se justifican procedimientos e instrumentos utilizados, y finalmente se habla sobre el manejo y análisis de los datos.

3.1 Tipo de Investigación

Para partir esta sección es necesario establecer que según los alcances y objetivos particulares a cada caso, es posible diferenciar ciertos tipos de investigaciones. Generalmente se consideran cuatro tipos: descriptivos, exploratorios, correlacionales y explicativos [41].

El proyecto AKORI busca generar una herramienta tecnológica para apoyar la toma de decisiones de estructura, contenido y diseño de sitios web, tomando como *input* variables fisiológicas que pueden describir las respuestas de los usuarios frente a diversos estímulos. El presente trabajo forma parte de las investigaciones del proyecto AKORI y pretende conocer el impacto que tiene la adición de una nueva variable (los neurodatos) a la metodología de identificación de *Website Keyobjets*.

En ese contexto se puede definir el presente trabajo como una investigación de tipo exploratorio. La base de este proyecto está en los resultados obtenidos en [4], donde se concluye que el uso de la dilatación pupilar como variable de clasificación de objetos en un sitio web, debiese ser complementada con el estudio de los procesos cerebrales que ocurren al mismo tiempo en los usuarios. Por lo tanto, este trabajo busca explorar si la inclusión de los neurodatos aporta información relevante acerca de las variables de emocionalidad del usuario y si es posible utilizarla como entrada para el clasificador de objetos. De esa manera se podrá comparar la precisión de ambos estudios y generar propuestas de acuerdo a los resultados.

3.2 Diseño de la Investigación

El diseño de la investigación consiste en la forma práctica y concreta de responder al problema de investigación y a lo planteado en los objetivos [41]. Es decir, en esta sección se define cuáles son las variables importantes de estudio y la manera como se deben obtener.

La principal variable de estudio es la actividad bioeléctrica cerebral o neurodatos generados por usuarios al momento de navegar por un sitio web. Además, se necesita conocer lo que está observando el usuario en el tiempo mientras realiza la navegación.

Para recolectar este tipo de datos es necesario llevar a cabo un experimento que permita al usuario navegar por un sitio web mientras se graba su electroencefalograma y a la vez se esté guardando información sobre sus ojos, como el posicionamiento y dilatación pupilar. Se diseña un experimento basado en el trabajo de [4], al que se le agrega un dispositivo de electroencefalografía en la instrumentación. El experimento se realiza en el Laboratorio de Neurosistemas de la Facultad de Medicina de la Universidad de Chile y se describe detalladamente en el capítulo 4 de este trabajo.

3.3 Selección de Instrumentos y Procedimientos

Como se mencionó anteriormente, el experimento es pieza fundamental para el desarrollo y término exitoso del Proyecto de Memoria. Por lo tanto, escoger instrumentos adecuados es clave.

Los instrumentos utilizados son los que están disponibles en la sala de registros del Laboratorio de Neurosistemas y corresponden a un distositivo de *Eye Tracking* y uno de EEG. Además se necesita un software que permita montar un experimento que incluya estos instrumentos y la presentación de estímulos visuales. Un punto importante es que la información producida por el *Eye Tracker* debe ser cruzada con la del EEG, por lo que es necesario que el software tenga la opción de enviar señales que marquen momentos importantes del experimento para el análisis y la sincronización.

El software utilizado es el *Experiment Builder*, que también está disponible en el Laboratorio y que permite realizar todo lo descrito anteriormente de una manera sencilla y amigable.

El procedimiento a seguir consiste en llevar diseñar e implementar el experimento para la obtención de datos y posteriormente aplicar el proceso KDD para el análisis.

En el capítulo 4 se entregan las características básicas de los instrumentos utilizados y el capítulo 5 muestra el análisis y los resultados.

3.4 Análisis de los Datos

El análisis de datos es ″un proceso continuo de examen de la información a medida que se obtiene, clasificándola, formulando preguntas adicionales, verificándola y desarrollando conclusiones″ [41].

En este trabajo el análisis de los datos se basa en el proceso KDD explicado en 2.6. Se trata principalmente de una serie de pasos que busca generar conocimiento a partir de un conjunto de datos.

Para hacer un buen análisis se debe hacer un tratamiento previo que consiste en seleccionar los datos y realizar un preprocesamiento, en el que se eliminan valores innecesarios para la investigación o que generan ruido. En este caso se eliminan los pestañeos, bostezos, tosidos, entre otros.

Posteriormente se aplican algoritmos de *clustering* y algoritmos de clasificación para identificar los objetos más importantes del sitio web en estudio.

Como se busca generar una comparación con el trabajo anterior [4], se utilizan los mismos modelos de *clustering*, que son *Association Rules* y *K-Means*. Es decir, se aplica la metodología anterior para determinar una línea base de objetos relevantes, pero utilizando los datos obtenidos en este trabajo. Esto permite además poder validar la metodología y tener etiquetas para los objetos en futuras clasificaciones.

Luego de lo anterior, es cuando se lleva a cabo la parte más importante del presente trabajo, que corresponde al estudio del uso del electroencefalograma en la identificación de *Website Keyobjects*.

Básicamente, se trata de caracterizar las curvas de las señales del EEG como una serie de parámetros que servirán de *inputs* para los modelos de clasificación que se emplean. Como un primer apronte se realiza un procedimiento análogo al de [4], utilizando como variables el tiempo, cambios en el tamaño de la pupila y otras características netamente relacionadas con la bioactividad cerebral. Los objetos encontrados en esta parte son comparados con los determinados anteriormente.

Finalmente se evalúan los resultados utilizando los indicadores de la sección 2.6.3. Con

esto es posible concluir sobre el impacto que genera el uso del EEG en la identificación de los *Website Keyobjects* y proponer recomendaciones y sugerencias de futuros trabajos.

4 Experimento

Este capítulo está destinado a describir detalladamente el experimento que permitirá contar con una base de datos para el análisis posterior. Lo primero es definir el diseño del experimento, se detalla la instrumentación, grupo experimental, protocolos y resultados esperados. Posteriormente se habla sobre la implementación del experimento.

4.1 Diseño del Experimento

El experimento se basa en el trabajo de Martínez [4], en el que se diseñó un programa de pruebas con el fin de conocer los objetos relevantes de un sitio web en particular por medio del análisis de posicionamiento ocular y dilatación pupilar. La principal diferencia con este trabajo, es la incorporación de la medida de actividad cerebral proporcionada por un dispositivo de electroencefalograma, con la que se reemplaza la dilatación pupilar para la identificación de los *Website Keyobjects*.

4.1.1 Instrumentación

Para la toma de muestras de los sujetos, se utilizan los siguientes instrumentos:

1. ***Eye Tracker***: Requerido para medir la exploración visual y dilatación pupilar del individuo. El Laboratorio de Neurosistemas cuenta con un equipo *EyeLink 1000* de la empresa *SR Research*. Este es el mismo equipo utilizado en [4] y [19].
   El equipo *EyeLink 1000* cuenta con una cámara infrarroja que detecta el reflejo de los ojos producido por luz infrarroja que envía un dispositivo anclado a la cámara. Este reflejo es procesado por la CPU de *EyeLink 1000* para determinar el desplazamiento de la exploración visual y la dilatación pupilar en cada momento.
   Las especificaciones técnicas del instrumento se encuentran en la tabla 4.1.

Tasa de Sampleo	2000 Hz Monocular / 1000 Hz Binocular
Precisión	0.25° - 0.5° precision promedio
Resolución	0.01° RMS, resolucion de micro sacadas de 0.05°
Acceso a Datos en tiempo real	1.4 msec (SD < 0.4 msec) @ 2000 Hz

Tabla 4.1: Especificaciones técnicas SR Research *Eyelink 1000*.
Fuente: Tabla obtenida de [19].

El equipo *EyeLink 1000* se conecta con su CPU donde se puede configurar a la medida de los requerimientos del usuario. Además se debe conectar con un computador que configure el experimento que será mostrado al sujeto. El software de configuración del experimento es *Experiment Builder* de la empresa *SR Research*.

Los tres elementos mencionados anteriormente conforman el sistema de *Eye Tracker*.

2. **Equipo de Electroencefalograma**: Instrumento necesario para medir la actividad bioelétrica cerebral, de ahora en adelante EEG. Para medir esta actividad se necesita un equipo que registre en ordenes de microvolts. Generalmente estos dispositivos cuentan con 32, 64 y hasta 128 electrodos para cumplir su misión.

   En esta oportunidad se contempla el uso del *Neuroheadset EPOC* desarrollado por la empresa *Emotiv*. Este dispositivo inalámbrico cuenta con 14 electrodos más dos de referencia y tiene como ventaja la facilidad en su uso y desarrollo. El tiempo que se emplea en su preparación también es considerablemente más bajo que un dispositivo de EEG convencional. En la siguiente figura se puede ver físicamente el *Emotiv EPOC*.

Figura 4.1: Equipo de Electroencefalografía *Emotiv EPOC Neuroheadset*.
Fuente: Imagen obtenida de [42].

En la tabla 4.2 es posible observar las especificaciones técnicas de este equipo.

Número de Canales	14 (más referencias CMS/DRL (posiciones P3/P4)
Canales (Sistema nternacional 10-20)	AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4
Método de sampleo	Sampleo secuencial.
Tasa de sampleo	128 SPS (2048 Hz)
Resolución	14 bits 1 USB = 0.51μV
Rango dinámico	8400 μV
Concectividad	Wireless, banda de 2.4GHz
Batería	12 horas

Tabla 4.2: Especificaciones técnicas *Emotiv EPOC Neuroheadset*.
Fuente: Tabla adaptada de [42].

4.1.2 Sitio Web

Al igual que en [4] se requiere un sitio web que cumpla con tres aspectos fundamentales: número de páginas, cantidad de objetos en cada una de ellas y cantidad de visitas al sitio.

El número de páginas es importante ya que determinará la complejidad en la identificación de los *Website Keyobjects*. Un sitio con una excesiva cantidad de páginas dificultará en demasía el análisis, mientras que uno con muy pocas no servirá en el estudio. En este experimento se requiere un sitio web que esté contenga entre treinta y cien páginas.

En cuanto a los objetos por cada página, es un análisis similar al anterior. Con pocos objetos por página el sitio web queda caracterizado fácilmente por ellos, y con demasiados objetos sería recomendable rediseñar el experimento en vez de testearlo en ese sitio. En este caso, el sitio debiese tener veinte objetos por página.

Finalmente, la cantidad de visitas al sitio es crucial para el proceso de análisis de datos. Las técnicas de *data mining* necesitan de una gran cantidad de datos para la identificación de patrones y generación de conocimiento. En este estudio se requiere al menos la cantidad suficiente de visitas para hacer análisis, junto con el acceso a los *web logs*.

4.1.3 Grupo Experimental

La elección del grupo experimental es una parte clave en el diseño de un experimento, ya que las respuestas de este grupo serán la base para el análisis posterior y las futuras conclusiones que se obtendrán del estudio. Utilizar un grupo erróneo conducirá a resultados erróneos.

Por otro lado, los participantes deben cumplir con un cierto perfil basado en los estudios anteriores del proyecto AKORI [19], [4], [3]. El experimento busca medir las respuestas fisiológicas de los sujetos a partir de la presentación de estímulos por medio de una página web. Por lo tanto se requiere que el grupo esté compuesto por personas cuya condición médica sea apta y además sepan utilizar un computador y un navegador web.

Las características específicas que debe cumplir el grupo experimental se mencionan a continuación:

- Personas de edad entre 18 y 50 años.
- Sujetos saludables, es decir, que no hayan sido diagnosticados con enfermedades que

puedan interferir en el estudio.

- No contar con historias de enfermedades neurológicas o psiquiátricas personalmente o en familia directa.
- Visión correcta. Se refiere a personas que puedan ver un sitio web sin usar anteojos ópticos.
- Sin historial de uso de drogas que afecten la dilatación pupilar.
- Consentimiento informado firmado.

El último punto señala que los participantes deben firmar un consentimiento informado. Este documento está aprobado por el Comité de Ética de la Facultad de Medicina de la Universidad de Chile y entrega la descripción del experimento en cuanto a objetivos, costos, riesgos, beneficios, entre otros aspectos del experimento.

4.1.4 Protocolo

El experimento está diseñado para registrar el posicionamiento ocular, dilatación pupilar y actividad cerebral de los sujetos al ser expuestos a estímulos visuales presentados en un sitio web.

El proceso de la tarea experimental es mostrar las distintas páginas que componen un sitio web en forma secuencial. De esta manera, los participantes pueden cambiar la página que está siendo visualizada cuando consideran que ya no hay nada más que les llame la atención o les interese. Además, no existirán tiempos mínimos ni máximos por página, otorgando el poder de acción a cada usuario.

Instrucción del Experimento

Para lograr lo anterior de buena manera, la instrucción del experimento debe permitir al usuario navegar libremente por las páginas presentadas como estímulos y cambiar entre las páginas cuando así lo decidan. Es por eso que se define la siguiente instrucción: *Navegue libremente por el sitio, sin límites de tiempo (mínimo ni máximo) por página. Utilice las flechas abajo y arriba para bajar o subir en la página. Cuando termine de explorar cada página presione la flecha derecha en el teclado para pasar a página siguiente.*

4.1.5 Resultados Esperados

Los resultados que se esperan del experimento son básicamente los que entregarán los dispositivos de *Eye Tracker* y EEG para cada sujeto en cada prueba.

En primer lugar, el *Eye Tracker* entrega un archivo EDF que contiene distintos parámetros que se registran mientras dura la prueba y permiten describir el posicionamiento ocular y la dilatación pupilar. Este archivo para ser visualizado debe convertirse a la codificación ASCII en extensión .asc. Los datos se registran de la siguiente forma:

```
<time> <xpl> <ypl> <psl> <xpr> <ypr> <psr> <xvl> <yvl> <xvr> <yvr> <xr> <yr>
```

Que corresponden a las siguientes variables:

- <**time**>: tiempo en milisegundos.
- <**xpl**>: posición ojo izquierdo en eje x.
- <**ypl**>: posición ojo izquierdo en eje y.
- <**psl**>: tamaño pupila izquierda (área or diámetro).
- <**xpr**>: posición ojo derecho en eje x.
- <**ypr**>: posición ojo derecho en eje y.
- <**psr**>: tamaño pupila derecha (área or diámetro).
- <**xvl**>: velocidad instantánea ojo izquierdo en eje x (grados/sec).
- <**yvl**>: velocidad instantánea ojo izquierdo en eje y (grados/sec).
- <**xvr**>: velocidad instantánea ojo derecho en eje x (grados/sec).
- <**yvr**>: velocidad instantánea ojo derecho en eje y (grados/sec).
- <**xr**>: resolución en eje x (unidades de posición/grados).
- <**yr**>: resolución en eje y (unidades de posición/grados).

En segundo lugar, el dispositivo de EEG entrega un archivo .bdf que contiene las señales grabadas por cuarenta electrodos posicionados en la cabeza del sujeto. Este archivo como tal no puede ser visualizado, pero con herramientas como *MATLAB* se puede expresar en diagramas que son entendidos por los investigadores.

4.2 Implementación

En esta sección se explica la implementación del experimento diseñado anteriormente. Se detallan los componentes escogidos para su puesta en marcha, es decir, el sitio web, el grupo experimental, el protocolo seguido en las pruebas y los resultados obtenidos. Además se destaca un cambio importante en la instrumentación y los motivos del cambio.

4.2.1 Instrumentación

El experimento fue diseñado para incluir el *Eyelink 1000* y el *Emotiv Epoc* en su instrumentación. Por los motivos que se describen a continuación, se decide cambiar el dispositivo de EEG por el mismo equipo utilizado en [19], es decir, el *Active Two System* de la compañía *BioSemi*.

El experimento debe ser capaz de generar archivos de *Eye Tracker* y EEG a la vez para poder cruzar y analizar en conjunto el posicionamiento ocular y la bioactividad cerebral. Es por eso que el archivo generado por el EEG debe contener marcadores o señales que permitan realizar el cruzamiento, es decir, debe ser posible de ver en qué partes del archivo ocurren los hechos que interesa analizar.

Por otro lado, el software en el que se programa el experimento, *Experiment Builder*, permite mandar un tipo de señales a dispositivos externos para lograr ese objetivo. Esto lo hace mediante la conexión por puertos que pueden ser en paralelo o por un puerto USB 1208HS.

El problema es que el software de grabación de señales de *Epoc*, denominado *Testbench*, soporta una conexión a través de solamente puertos en serie. Por lo tanto, no es posible enviar señales que marquen el EEG con los casos interesantes de análisis.

Además se descarta la instalación de otro software para programar experimentos, debido al protocolo existente en el Laboratorio de Neurosistemas.

Dado lo anterior, la mejor solución es utilizar el dispositivo de EEG presente en el Laboratorio cuya descripción es la siguiente.

El equipo es el *Active Two System* de *BioSemi* con capacidad para 256 electrodos para la cabellera, más 8 electrodos externos. En el Laboratorio de Neurosistemas se cuenta con 32 electrodos para la cabellera y 8 externos, formando un sistema de electroencefalografía de 40 electrodos.

La imagen 4.2 muestra el sistema de electroencefalografía utilizado en el presente proyecto.

Figura 4.2: Sistema de Electroencefalografía *Active Two System* de *BioSemi*.
Fuente: Imagen propia.

Además, es necesario un computador para procesar la información que entrega el EEG en tiempo real. En el computador está el software de *BioSemi* que se encarga de graficar las señales y guardarlas en los archivos correspondientes.

4.2.2 Sitio Web

De igual forma que en [4] el sitio web escogido es el del programa de MBA de Ingeniería Industrial de la Universidad de Chile[1]. El sitio está vigente desde 2011 y cuenta con la información acerca del programa, los profesores, el perfil de los estudiantes y egresados, etcétera.

El sitio elegido cuenta con 32 páginas y 359 objetos que aparecen 1014 veces en total, lo que indica que algunos objetos aparecen más de una vez en el sitio. En promedio, hay 31,9 objetos por página. Esta información junto con las bases de datos de los objetos y los *web logs* fue tomada directamente del trabajo de [4], por lo que no se entrará en mayor detalle respecto a este tema.

En la siguiente imagen es posible ver la primera página que se muestra en el experimento, que corresponde a la página de inicio del sitio.

[1] www.mbauchile.cl

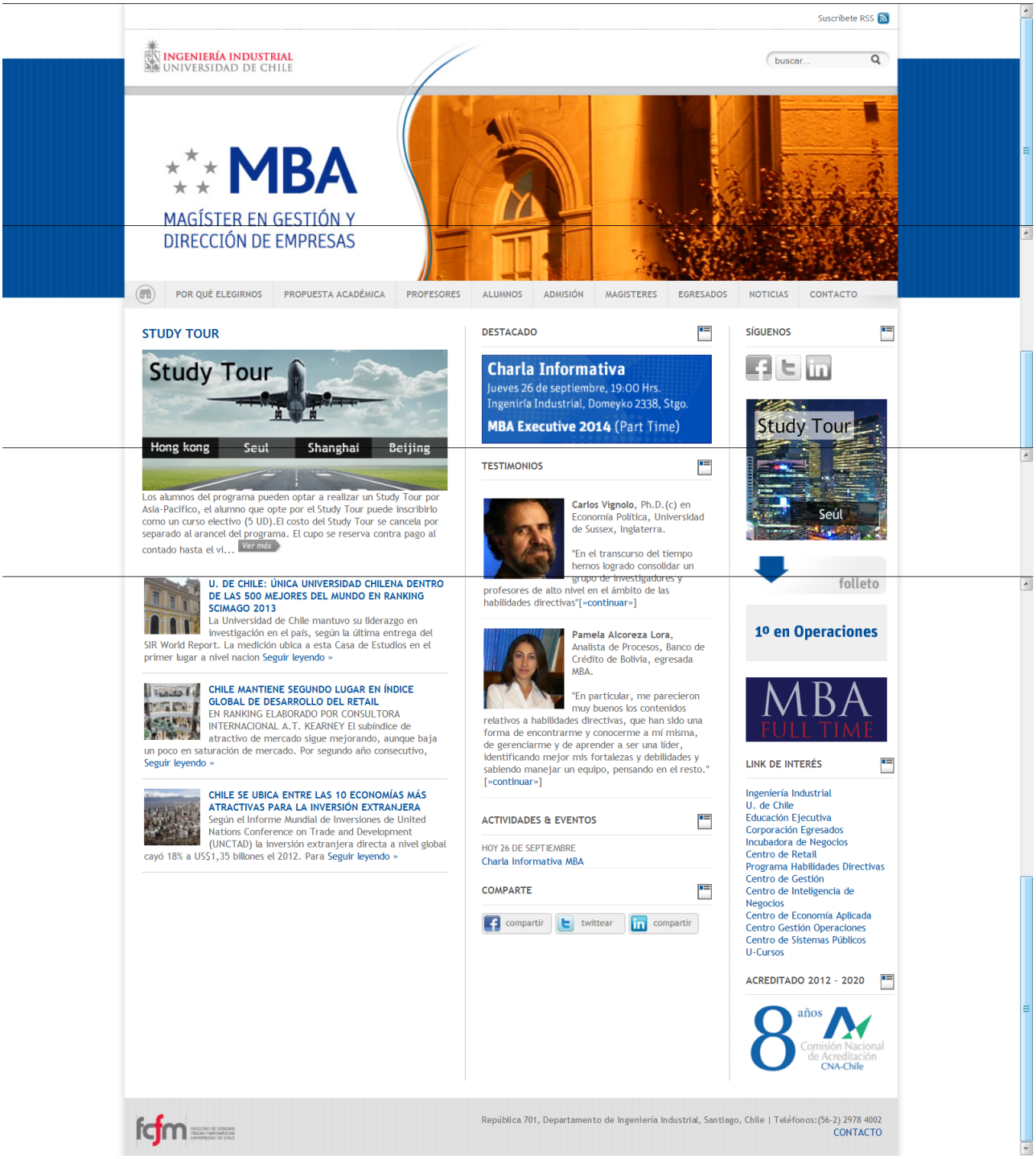

Figura 4.3: Página de inicio de www.mbauchile.cl y primera página mostrada en el experimento. Fuente: Imagen obtenida de [4].

4.2.3 Grupo Experimental

Se escogió un grupo experimental tomando en cuenta los requerimientos mostrados anteriormente. El grupo elegido consta de veinte personas de ambos sexos, diecisiete hombres y

tres mujeres. La edad promedio del grupo es de 24,2 años, presentando una varianza de 1,64 años.

El grupo está compuesto principalmente por estudiantes universitarios y profesionales de diversas áreas como medicina, ingeniería, fotografía, entre otras.

En la tabla 4.3 se pueden apreciar las principales características del grupo experimental.

Nº	Sexo	Edad	Actividad o Profesión
1	Femenino	22	Fotógrafa
2	Masculino	25	Odontología
3	Masculino	25	Médico
4	Masculino	25	Arquitecto
5	Masculino	25	Profesor
6	Masculino	24	Estudiante
7	Masculino	20	Estudiante
8	Masculino	25	Estudiante
9	Femenino	25	Estudiante
10	Masculino	25	Estudiante
11	Masculino	24	Ingeniero Civil Industrial
12	Masculino	25	Estudiante
13	Masculino	25	Estudiante
14	Masculino	24	Estudiante
15	Masculino	24	Estudiante
16	Femenino	24	Ingeniera Civil Industrial
17	Masculino	24	Estudiante
18	Masculino	24	Estudiante
19	Masculino	25	Estudiante
20	Masculino	25	Economista

Tabla 4.3: Características básicas del grupo experimental.
Fuente: Elaboración Propia.

4.2.4 Protocolo

Al momento de implementar el experimento es necesario seguir una serie de pasos para completar la tarea diseñada. Los pasos son los que se describen a continuación [19].

1. Preparar equipos: Es muy importante tener los equipos y dispositivos encendidos y listos para ser utilizados apenas llegue el sujeto de prueba.
2. Bienvenida y relleno de documentos: Al llegar a la sala de registros, se le da la bienvenida y agradecimientos al sujeto por su participación en el experimento. Además, se explica brevemente lo que se hará y se le pide que lea y firme el consentimiento informado.
3. Configurar EEG: Este paso es muy importante y presenta un alto grado de dificultad. Representa el mayor tiempo de preparación y debe seguir el siguiente proceso:

- Medición de las dimensiones de la cabeza del sujeto, siendo tres las distancias principales: nasion-inion (desde la frente a la nuca); trago-trago (desde una oreja a la otra); circunferencia. Con estas dimensiones se elegirá una gorra adecuada y se instalará de manera correcta en la cabeza del sujeto.
- Limpieza con alcohol las zonas donde se colocan los electrodos externos (alrededor del ojo y zona mastoidea), para reducir la impedancia de la piel y obtener señales más precisas y limpias.
- Colocar los 8 electrodos externos alrededor de los ojos y en las zonas mastoideas con adhesivos para electrodos y gel conductor que conecta el metal del electrodo a la piel.
- Escoger la gorra de EEG de 32 orificios de acuerdo al tamaño de la circunferencia de la cabellera del sujeto.
- Acomodar la gorra de EEG ajustando la posición Cz de la misma a la mitad de las distancias entre nasion-inion y trago-trago.
- Suministrar gel conductor en cada orificio cuidando que este en contacto con la piel de la cabellera.
- Ubicar la banda de 32 electrodos en la gorra de EEG respetando la posición correspondiente a cada uno. En la figura 4.4 se puede visualizar a un sujeto con la gorra y los electrodos conectados.
- Conectar los los electrodos al equipo para traspasar la información a un computador.
- Revisar en el programa *ActiView* en el computador de destino de las señales para comprobar la configuración ajustada.

Figura 4.4: Sujeto con gorra y electrodos posicionado para el experimento.
Fuente: Imagen propia.

4. Configurar *Eye Tracker*: Para cada usuario se debe configurar el dispositivo de acuerdo a sus propias características. El proceso es el siguiente:
   - Ajustar la cabeza del sujeto, según su altura, al soporte dedicado para este propósito.
   - Ajustar la altura de la pantalla de estímulo a la altura de los ojos del sujeto. La parte central de la pantalla debe quedar justo frente a los ojos del sujeto.
   - Ajustar la cámara infrarroja a la posición de la cabeza considerando que deben detectarse tanto el reflejo corneal y la pupila. Además se debe fijar el enfoque de la cámara.
   - Calibrar el *Eye Tracker* al ajuste realizado en la CPU exclusiva del *EyeLink 1000*.
   - Validar la calibración realizada.
5. Registro: Se lleva a cabo el experimento por medio de la tarea definida en 4.1.4, tal como se muestra en la imagen 4.5.
6. Fin del experimento: Al finalizar el experimento se deben realizar los siguientes pasos:
   - Detener el registro de los equipos y retirar los instrumentos del sujeto.
   - Lavar el cabello del sujeto para remover los residuos de gel conductor.
   - Finalmente se deben lavar los instrumentos utilizados, limpiar las zonas ocupadas y dejar en perfecto orden la sala de registro.

Figura 4.5: Sujeto realizando el experimento.
Fuente: Imagen propia.

4.2.5 Resultados

A partir de los experimentos realizados se obtuvo una gran cantidad de datos para cada sujeto, los que sirven para efectuar el análisis posterior. Los archivos generados por el *Eye Tracker* (edf) tienen en promedio un tamaño de 27 MB, mientras que los archivos del EEG (bdf) pesan en promedio 356 MB. El gran tamaño se debe a la precisión con la que se miden las respuestas y a la duración del experimento, que en promedio tardaba 30 minutos.

5 Análisis y Resultados

En este capítulo se muestra detalladamente el tratamiento de los datos obtenidos del experimento, su análisis mediante la metodología KDD y los resultados obtenidos.

El trabajo realizado se dividió en dos partes; primero se replicó la metodología de Martínez [4] para obtener una lista base de *Website Keyobjects* y objetos no relevantes, utilizando sólo las variables visuales, de tiempo y número de visitas. En segundo lugar, se emplea un procedimiento similar, utilizando las mismas varibles, pero esta vez en conjunto con la nueva fuente de información, es decir, la actividad bioeléctrica cerebral caracterizada de diferentes maneras.

A continuación se explica, por separado, el desarrollo del proceso KDD para la data generada por el ET y el EEG.

5.1 Proceso KDD

El presente trabajo considera la utilización del proceso KDD para el análisis de los datos. Este proceso consta de cinco etapas que serán descritas en detalle para los archivos generados por el *Eye Tracker* y el EEG.

Una vez finalizados los veinte experimentos, y como se menciona en 4.2.5, se contó con veinte archivos EDF originados por el *Eye Tracker*, más veinte BDF provenientes del EEG y una planilla de información de cada sujeto.

De los veinte sujetos, uno debió ser eliminado porque, debido a una falla técnica en el computador del EEG, la grabación de las señales del cerebro se vio interrumpida y no se terminó de buena manera su experimento. Los demás sujetos de estudio no presentaron mayores problemas en la toma de muestras, por lo que la selección de datos comprende la utilzación de dos archivos para cada una de las 19 personas.

Es importante destacar que la mayor parte del análisis fue realizado con el *software Matlab*, que es un programa diseñado para el trabajo matemático utilizando matrices. Además tiene una gran cantidad de funciones y algoritmos implementados que facilitan el trabajo de investigación.

5.1.1 *Eye Tracker*

Selección

Se seleccionan 19 de 20 archivos EDF para el desarrollo del trabajo. Como se declara anteriormente, estos archivos no pueden ser utilizados de otra manera que no sea con el *software Experiment Builder*, por lo que son convertidos a la extensión de texto ASCII. Los archivos de texto poseen columnas con valores para cada una de las variables posibles de estudio que contiene el archivo EDF. Las variables son las mencionadas en la parte 4.1.5.

En la figura 5.1 se muestra la dilatación pupilar sin procesar para un ojo en un intervalo de tiempo dado. Es posible notar que es una señal con mucho ruido, por lo tanto, en la etapa de preprocesamiento se deben eliminan los siguientes los siguientes elementos que generan ruidos, como son los parpadeos, las sacadas y las altas frecuencias.

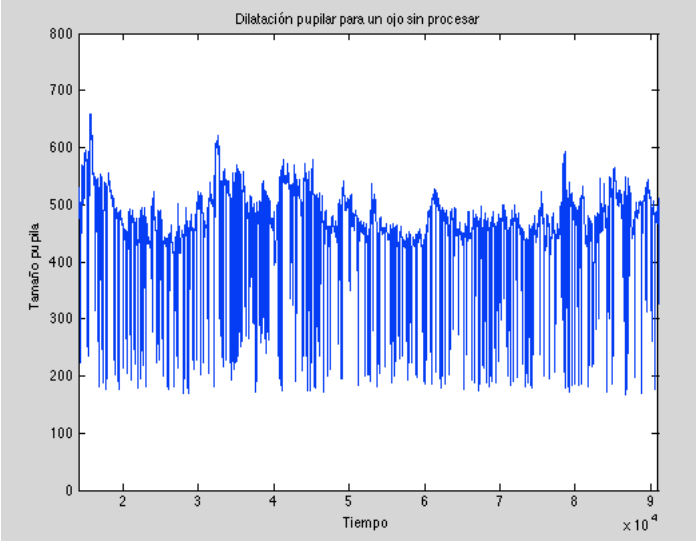

Figura 5.1: Gráfico de la dilatación pupilar sin procesar para una ventana de tiempo aleatoria.
Fuente: Elaboración propia.

Preprocesamiento

El tamaño promedio de los archivos ASCII es bajo (39 MB), por lo que es posible generar códigos de *Matlab* que tomen todos los archivos de una vez para realizar la limpieza y el preprocesamiento de manera rápida.

Para cada archivo se debe eliminar ciertos artefactos que producen ruido en la señal. Esto se hace de manera análoga a trabajos anteriores [3], [19]. Los elementos a eliminar son los siguientes:

1. Parpadeos: Cuando un sujeto parpadea, es decir, cada vez que cierra sus ojos, por más corto que parezca este movimiento, el dispositivo deja de captar la pupila que está grabando. De esa manera, al récord de datos, se le asigna un valor igual a cero en cada momento de pérdida de señal, que debe ser arreglado mediante una interpolación. Esto se puede ver claramente en la imagen 5.2.

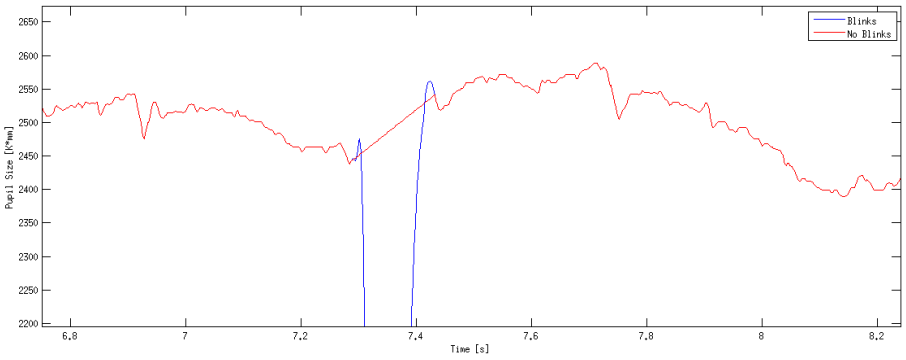

Figura 5.2: Gráfico de la interpolación de parpadeos en la señal de dilatación pupilar.
Fuente: Imagen realizada por Kristofher Muñoz para el proyect AKORI.

2. Sacadas: La eliminación de estos artefactos tiene su justificación en el hecho de que al pasar de una fijación a otra, el ángulo con el que el dispositivo capta la pupila cambia, es decir, reconoce una pupila de distinto tamaño y esto no necesariamente refleja la realidad. Por lo mismo, lo que se hace para corregir este fenómeno es llevar la señal desde el punto donde donde termina una sacada al punto donde comienza y generar una interpolación lineal entre los extremos de ese intervalo. En la figura 5.3 se puede apreciar una pupila con las sacadas en bruto y una con el efecto de las sacadas corregido.

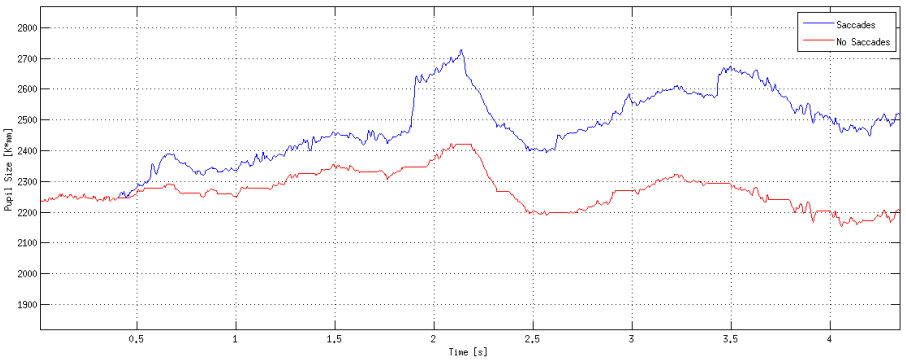

Figura 5.3: Gráfico de la corrección de sacadas en la señal de dilatación pupilar.
Fuente: Imagen realizada por Kristofher Muñoz para el proyect AKORI.

3. Altas frecuencias: Se aplica un filtro pasa-bajo para eliminar todas las frecuencias mayores a 2 Hz que para el análisis de dilatación pupilar se consideran como ruidos. El resultado de esta limpieza se pude apreciar en la imagen 5.4

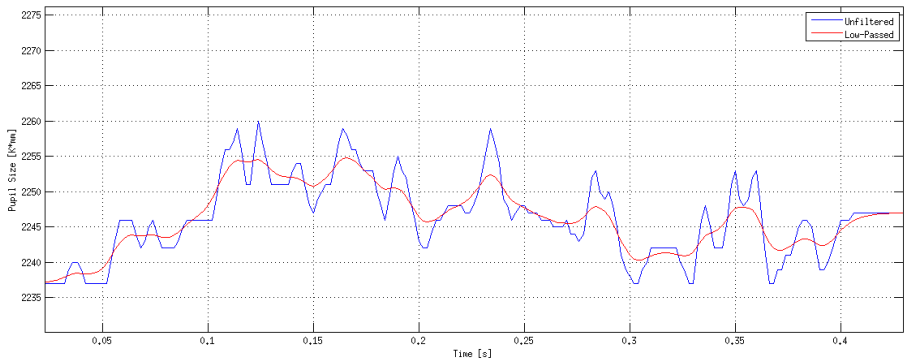

Figura 5.4: Gráfico de la eliminación de altas frecuencias en la señal de dilatación pupilar. Fuente: Imagen realizada por Kristofher Muñoz para el proyect AKORI.

Transformación

Una vez terminada la limpieza de los datos, es posible continuar con la siguiente etapa del proceso KDD: la transformación. En esta etapa lo que se busca es caracterizar los objetos presentes en el sitio web para su posterior clasificación en *Website Keyobjects* y objetos no relevantes.

Como la idea es replicar la metodología de [4], se deben obtiener las mismas características para describir cada objeto.

La primera transformación, entonces, es separar cada una de las matrices en intervalos. El resultado es una nueva matriz que indica, para cada sujeto, la página en la que se encuentra, el objeto que está mirando, el índice inicial y el índice final. Es decir, se separa la data en observaciones, y cada una de éstas presenta los siguientes requerimientos [3]:

- Una observación empieza cuando parte una fijación dentro de un objeto y termina cuando ocurre una sacada que lleva la posición ocular fuera de ese objeto.
- Para que sea considerada como observación, es necesario que el sujeto permanezca en el mismo objeto al menos 300ms.
- Una observación considera una señal de dilatación pupilar desde su inicio hasta los siguientes 1000ms.

Creados los intevalos, es sencillo juntar todas las observaciones correspondientes a los objetos vistos por todos los sujetos. Con esto, se puede realizar la obtención de características para cada objeto. Cabe destacar que no se obtienen características para la totalidad de los objetos presentes en el sitio, esto es esperable porque los usuarios no prestan atención a todos los objetos, incluso si se fijaran en todos, existe el límite inferior de los 300ms. Las

	Tiempo	Delta	Visitas
Tiempo	1		
Delta	0.0750	1	
Visitas	0.7451	0.0447	1

Tabla 5.1: Correlación entre variables.
Fuente: Elaboración propia.

características a extraer son las siguientes:

1. Tiempo promedio por cada objeto.
2. Indicador Delta: Definido por la diferencia entre la máxima dilatación y la mínima contracción de la pupila en el objeto.
3. Número de vistas al objeto.

Las variables físicas de cada objeto, a nombrar Alto, Ancho, Área, no se consideran ya que no presentan relación con la importancia del objeto para definir el contenido del sitio [4]. En la tabla 5.1 se encuentra la correlación entre las variables escogidas como características de clasificación.

Minería de Datos

En esta fase del proceso KDD se busca generar una clasificación de los objetos vistos en *Website Keyobjects* y objetos no relevantes. El método de Reglas de Asociación entrega el mismo resultado que el trabajo de Martínez por lo que se omite en este informe. Entonces, partiendo de esa base, se utiliza el modelo de agrupamiento *K-Means*, considerando como variables de clasificación el Tiempo promedio y el indicador Delta.

El resultado de este *clustering* consiste en tres grupos de objetos definidos como Tiempo Alto, Indicador Delta Alto y Baja Importancia. Como es de esperar en este trabajo, los objetos presentes en el grupo de baja importancia son mucho más en cantidad que los presentes en los otros dos grupos.

En la figura 5.5 se puede ver gráficamente los tres grupos formados. Los objetos pertenecientes al *cluster* Tiempo Alto son los elementos de color verde, los de Indicador Delta Alto son los de color rojo, y los de Baja Importancia son los de color azul.

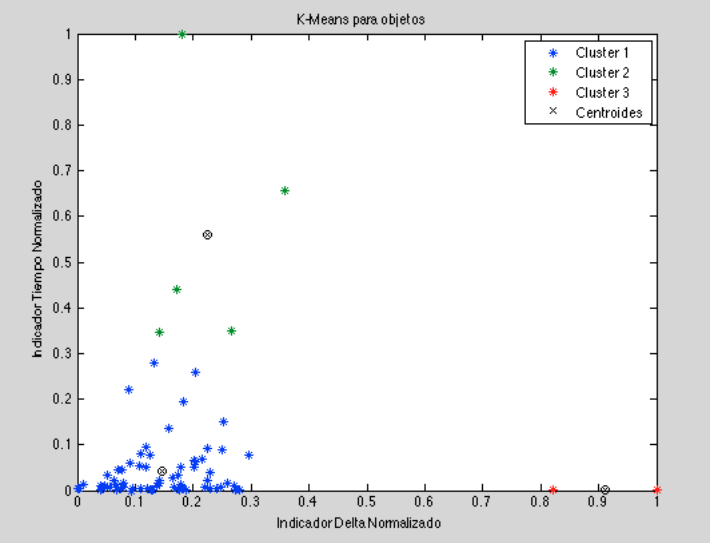

Figura 5.5: Grupos formados con los indicadores Delta y Tiempo.
Fuente: Elaboración propia.

En las tablas 5.2 y 5.3 se puede apreciar los objetos que componen los *clusters* de mayor importancia.

Tiempo Alto
Objeto 5
Objeto 94
Objeto 215
Objeto 244
Objeto 265

Tabla 5.2: *Cluster* Tiempo Alto
Fuente: Elaboración propia.

Indicador Delta Alto
Objeto 103
Objeto 127

Tabla 5.3: *Cluster* Indicador Delta Alto
Fuente: Elaboración propia.

De la misma manera que en el trabajo previo de Martínez, para obtener el grupo final de *Website Keyobjects* se debe considerar la cantidad de visitas para cada objeto. Luego de esto se crea un nuevo indicador definido como el producto entre el Indicador Delta y el Tiempo (usados en el *clustering anterior*), y se ordenan de mayor a menor, dejando fuera los que tienen este valor igual a cero.

Finalmente se seleccionan los veinte primeros objetos del ordenamiento realizado, dando como resultado la lista de *Website Keyobjects* presente en la tabla 5.4.

Objeto	Descripción
94	Párrafo 1 Contacto
244	Imagen Plataforma 1
5	Header picture
265	Párrafo Corporación 1
306	Párrafo Meteodología 1
215	Párrafo propuesta 1
245	Párrafo Perfil Egresados 1
267	Párrafo Quienes 1
243	Párrafo Plataforma 1
356	Picture Detalle Profe 1
307	Párrafo Metodología 2
266	Imagen Corporación 1
357	Párrafo Testimonio Detalle 1
84	Párrafo 1 Doble Grado
73	Plan de Estudios Chart
195	Párrafo Detalle noticia 1
269	Párrafo Quienes 2
35	Charla Informativa Picture
295	Párrafo Biblioteca 1
23	Main Post Text 1

Tabla 5.4: *Website Keyobjects* encontrados usando metodología de [4]
Fuente: Elaboración propia.

Evaluación

Los objetos reconocidos como relevantes son comparados con los que se obtuvieron en en [4]. Es importante destacar que existen diferencias entre las listas de objetos para ambos trabajos.

Del total de objetos, sólo 12 coinciden con la lista de Martínez. Esto se puede deber a variados factores, como por ejemplo:

- La cantidad de sujetos de de experimentación no es suficiente para lograr resultados significativos y extrapolables a la población.
- Existen diferencias entre los segmentos de personas a los que se les aplicó el experimento en cada trabajo. Es decir, puede ser que el sitio web en estudio no está dentro de las mismas preferencias de las personas utilizadas como sujetos en este experimento que de las personas en el anterior.
- Al momento de realizar el experimento la instrucción no fue comunicada de la misma manera o no fue entendida como debiese haber sido.

De todas maneras es importante destacar que dentro de los objetos catalogados como relevantes en este trabajo, no están incluidos cuatro objetos que sí fueron determinados por Martínez, los que finalmente fueron descartados por el *webmaster* (objetos 1, 3, 4, 130).

Por otro lado, existen otros objetos que no fueron encontrados por Martínez, pero fueron encontrados por otro memorista, Jorge Dupré, por medio del análisis de medidas de centralidad de Teoría de Grafos aplicado a los datos de [4]. Estos son los objetos 73, 84, 94 y 269, que fueron clasificados como relevantes y validados por el *webmaster*.

Además hay tres objetos que en [4] son validados como relevantes pero en la nueva investigación de Dupré, son rechazados por el *webmaster*. Se habla de los objetos 5, 244 y 356.

Finalmente, es posible notar que la lista posee cuatro objetos nuevos (35, 266, 295, 357), que debiesen ser validados por el *webmaster*.

Por las diferencias y ambigüedades entre los trabajos de Martínez y Dupré, para la siguiente parte de análisis con el EEG, se utilizará como base de *Website Keyobjects* los objetos encontrados en este trabajo, es decir, los presentes en la tabla 5.4.

5.1.2 Electroencefalograma (EEG)

Selección

Tal como se menciona al principio de esta sección, los datos de un sujeto debieron ser descartados del análisis debido a un desperfecto técnico en la grabación del EEG. En consecuencia, se seleccionan 19 archivos BDF para llevar a cabo el análisis.

Los archivos que entrega el EEG, en un principio son analizados con el *plug-in Eeglab* de *Matlab*, herramienta que permite realizar de forma sencilla y clara variados procedimientos con este tipo de datos, entre los que destacan gráficos, filtros de frecuencia, cambios de tasas de sampleo, crear épocas de estudio, entre otros. Además, se hace uso de otro *plug-in*, *Eye-Eeg*, que permite sincronizar una grabación de EEG con una de ET, y aplicar diversas funciones de manejo de datos e investigación.

Por último, se trabaja directamente con matrices formadas por los valores de las señales de los electrodos en función del tiempo.

En la figura 5.6, es posible ver un gráfico de la data obtenida para un sujeto en particular

sin procesar. En el eje x está el tiempo y en el eje y cada uno de los canales del EEG, se incluyen los de la cabeza y los externos. Se puede notar además uno de los marcadores, incluidos en el experimento para delimitar secciones importantes para el análisis.

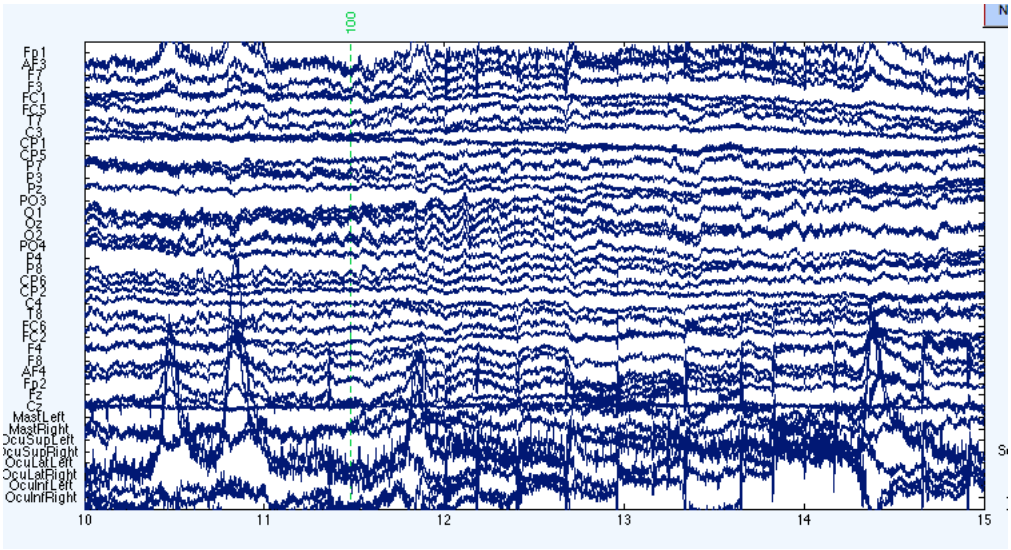

Figura 5.6: Ejemplo de data cruda del EEG.
Fuente: Elaboración propia.

De la imagen 5.6 se puede inferir que la data cruda posee una gran cantidad de artefactos que producen ruido, por lo que debe ser limpiada y preprocesada.

Preprocesamiento

Al igual que para el caso del ET, para el EEG algunos procesos fisiológicos generan ruidos en la data. Los más importantes son los siguientes [19]:

- Parpadeos.
- Movimientos oculares (sacadas).
- Electromiograma (cuando el sujeto aprieta los dientes o traga saliva).
- Altas y bajas frecuencias.
- Irregularidades, por ejemplo, cuando el sujeto tose, se rasca o mueve la cabeza, se suelta algún electrodo, etcétera.

Antes de eliminar estos ruidos, se realizan ciertos tratamientos detallados a continuación:

1. Bajar la tasa de sampleo de 2000Hz a 500Hz, esto sirve para ajustar la cantidad de datos con el fin de que calcen con las mediciones del ET.

2. Filtrado de frecuencias, dejando la data entre 1 y 60Hz.
3. Sincronización de la medición del EEG con la del ET, mediante el uso del *plug-in Eye-Eeg*.

Una vez listos estos pasos, se procede a realizar un Análisis de Componentes Principales (ACP), cuyo objetivo es reconocer los dintintos componentes en los que se puede desglosar la señal del EEG. Esto se realiza para la eliminación de los ruidos presentados anteriormente.

Cuando se lleva a cabo el ACP, se consigue distinguir, en un principio, 32 componentes que muestran las activaciones en las zonas de la corteza cerebral que son medidas por el EEG. La idea es reconocer cuáles corresponden a ruidos o irregularidades.

El procedimiento es primero efectuar el ACP, posteriormente, detectar las sacadas y fijaciones dentro de la data del EEG, para luego eliminar las componentes que tengan relación con los artefactos de ruido.

Es posible reconocer visualmente las componentes que presentan ruidos, ya que muestran gran activación en zonas cercanas a los ojos (ruidos de parpadeos y movimientos oculares), a las orejas y parte de atrás de la cabeza (electromiograma), activación fuerte en toda la cabeza (ruidos musculares y electromiograma), etcétera.

En la figura 5.7 es posible apreciar algunas componentes con ruidos que fueron eliminadas. Es importante destacar que esto se hace para todos los sujetos por separado, implicando una alta cantidad de tiempo y recursos computacionales.

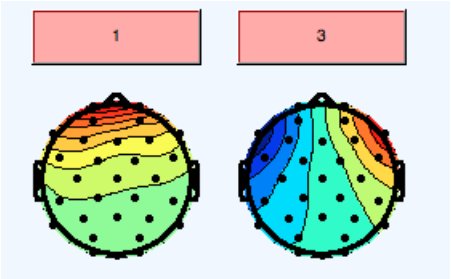

Figura 5.7: Ejemplo de componentes a eliminar por estar relacionadas con artefactos de ruido.
Fuente: Elaboración propia.

Al mismo tiempo que se eliminan las componentes de ruido, la data de los canales del EEG sufre cambios: donde existían los ruidos, los datos son interpolados para su corrección. Por lo tanto, una vez finalizado este procesamiento, se puede trabajar tanto con los canales como con las componentes.

Transformación

Una idea general de los fenómenos que se estudian utilizando el EEG, se puede obtener gráficamente a través del ERP [25], [24], [26]. De esta forma, se grafican los ERP para comparar la actividad cerebral de los *Website Keyobjects* con los objetos no relevantes. Se utilizan las señales promediadas para los objetos de la lista 5.4, y el promedio de las señales de otros objetos vistos por los sujetos. Para poder obtener estas señales, previamente se separan por intervalos, usando en conjunto la data del ET, obteniendo matrices que indican el tiempo, el objeto, los intervalos de tiempo y las señales de todos los canales del EEG.

Se grafican los ERP para todos los canales y para electrodos por separado; en la literatura se puede ver que los investigadores utilizan distintas combinaciones para sus análisis. En particular, en este trabajo se ocupan todos los canales [43], [44], el electrodo PZ [26], y el electrodo FC6 [23], [35], para la medición de bioactividad cerebral.

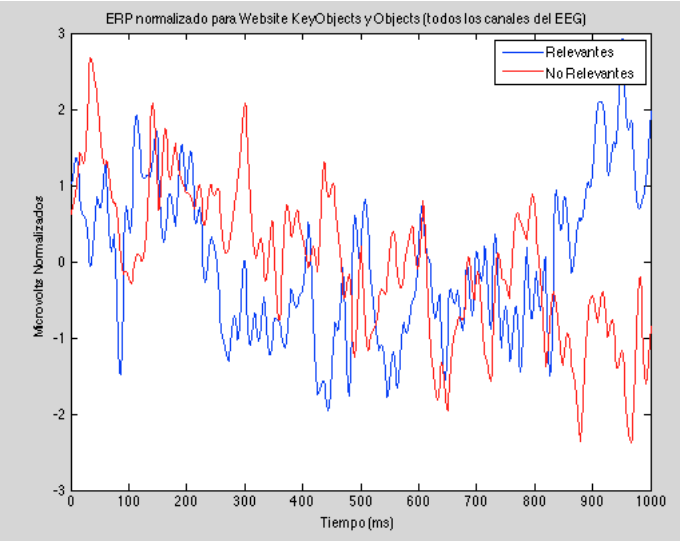

Figura 5.8: ERP normalizado para objetos relevantes y no relevantes (todos los canales).
Fuente: Elaboración propia.

Como se puede apreciar en la figura 5.8, no existen diferencias notorias en el ERP para objetos relevantes y no relevantes a simple vista. Esto sucede también para los electrodos FC6 y PZ, por lo que se procede a analizar la data de otra manera: se intenta caracterizar las señales por medio de diversas transformaciones.

En los estudios de electroencelografía, la data es caracterizada utilizando diversas metodologías. En [43], [44], [45], entre otros estudios, se propone descomponer las señales en bandas de frecuencia tales que coincidan con las ondas típicas del EEG (Alpha, Beta, Gamma, Delta, Theta descritas en 2.4.3), a través de las transformadas Wavelet. En [23] y [35] caracterizan

la excitación y valencia emocional definiendo dimensiones fractales para las ondas. En otros estudios, como [29], se usan estadísticos típicos como media o varianza de las ondas.

Por lo anterior, en esta etapa se crean tres bases de datos, una que contenga las señales de cada objeto caracterizadas por 14 variables para todos los canales; y otras dos que contengan las mismas variables para el electrodo PZ y el FC6, respectivamente. A continuación se detallan las 14 variables a utilizar:

En primer lugar se utilizarán las variables definidas con la metodología típica de *Website Keyobjects*, es decir: Tiempo, Indicador Delta y Vistas, además se agrega el indicador Delta*Tiempo. Adicionalmente, se calculan la media, desviación estándar y varianza de cada señal.

Por otro lado se descompone la señal por medio de la transformada Wavelet. En este caso, como la señal está medida con una tasa de sampleo de 500Hz, se utiliza la descomposición en 6 niveles de detalle (D1-D6) y uno final de aproximación (A6), junto a la función Wavelet Daubechies de orden 5 [45]. El resultado de este proceso se puede apreciar en la tabla 5.5.

Nivel de descomposición	**Banda de frecuencia**	**Rango de frecuencia (Hz)**
A6	Delta	0 - 4
D6	Theta	4 - 8
D5	Alpha	8- 16
D4	Beta	16 - 32
D3	Gamma	32 - 64
D2	Ruido	64 - 128
D1	Ruido	128 - superior

Tabla 5.5: Descomposición por bandas de frecuencia para tasa de sampleo de 500Hz.
Fuente: Elaboración propia.

Para cada banda se obtienen las características que sirven para clasificar. En [43], [44], utilizan las siguientes:

1. Energía: Definida como la cantidad de energía que posee cada banda, calculada con la ecuación 5.1. C^i representa los coeficientes de la banda i y n_i es la cantidad de coeficientes en la banda i.
$$Energia_i = \sum_{k=1}^{n_i} (C_k^i)^2 \tag{5.1}$$
2. Procentaje de energía: Calculado con la ecuación 5.2, se refiere al porcentaje de energía

que tiene cada banda con respecto del total.

$$\%\ Energia_i = \frac{Energia_i}{Energia\ Total} \cdot 100 \tag{5.2}$$

3. RMS: Representa la raíz cuadrada promedio, calculada con la ecuación 5.3.

$$RMS_j = \sqrt{\frac{\sum_{i=1}^{j} \sum_{k=1}^{n_i} (C_k^i)^2}{\sum_{i=1}^{j} n_i}} \tag{5.3}$$

Se busca representar cada señal con las bandas que aporten con más información. En este caso, las bandas elegidas son la Delta y la Alpha [45].

Finalmente se calcula la Dimensión Fractal a través del algoritmo Higuchi tal como se define en [23].

Agrupando todo lo anterior se tiene un archivo que contiene las 14 variables para todos los canales, para el canal PZ y para el canal FC6.

Minería de Datos y Evaluación

El objetivo principal de esta parte es poder clasificar utilizando las variables anteriormente definidas los objetos en dos grandes grupos, relevantes y no relevantes. El procedimiento a seguir es análogo a la metodología de [4], es decir, mediante el algoritmo de *clustering K-Means*. Se intentará encontrar las variables que más aporten en la clasificación.

Tiempo y Energía Banda Delta: Se realiza el agrupamiento utilizando estas dos variables, entregando malos resultados para tres segmentos, como se ve en la figura 5.9.

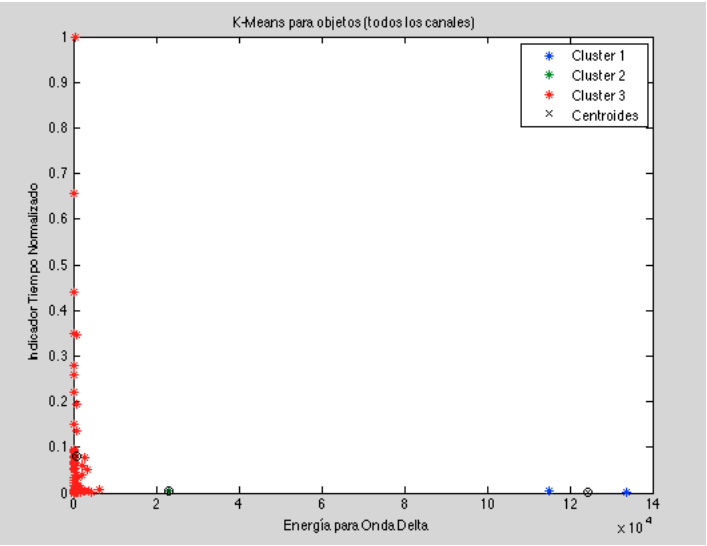

Figura 5.9: K-Means para objetos utilizando Energía Banda Delta y Tiempo (todos los canales).
Fuente: Elaboración propia.

Tiempo y Porcentaje de Energía Banda Delta: En este caso es posible distinguir tres grupos diferenciados por sus características, como muestra la imagen 5.10. El primero, en verde, presenta una gran importancia relacionada con el tiempo con bajo porcentaje de energía, el segundo (azul) tiene valores medios de porcentaje de energía, y el tercero, correspondiente al color rojo, presenta el mayor porcentaje de energía.

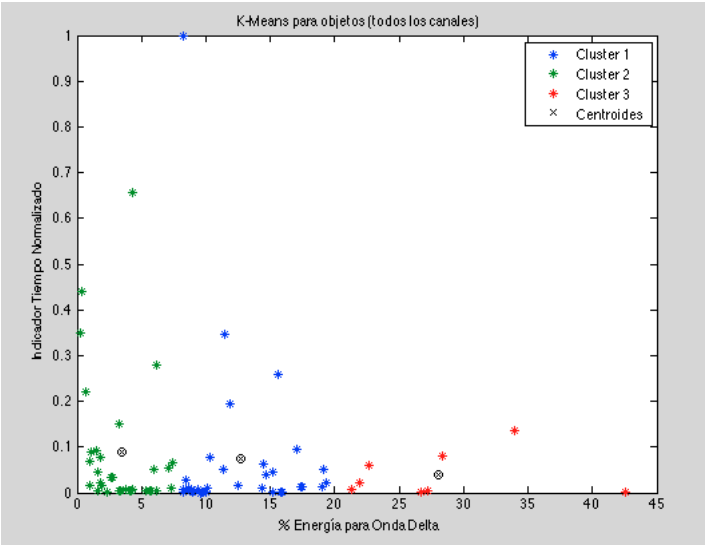

Figura 5.10: K-Means para objetos usando Porcentaje de energía Banda Delta y Tiempo (todos los canales).
Fuente: Elaboración propia.

Para determinar los objetos relevantes luego de este agrupamiento, se procede a calcular el indicador Tiempo*Porcentaje de Energía, agregar la dimensión de Vistas y ordenar. Los resultados se encuentran en la tabla 5.6.

Objeto
244
307
306
215
94
243
295
267
271
74
214
359
87
343
195
35
75
245
73
218

Tabla 5.6: Objetos relevantes usando Porcentaje de Energía Banda Delta y Tiempo (32 canales). Fuente: Elaboración propia.

Este *clustering*, arroja un 60% de precisión, comparándolo con la lista de *Website Keyobjects* base de la tabla 5.4. De los 8 objetos que están mal clasificados, 5 son párrafos, 1 es un formulario, y 2 son imágenes (mapa y un ex alumno).

Análogamente para el electrodo PZ se obtienen los objetos de la tabla 5.7, nuevamente con un 60% de precisión.

De la misma manera se analiza para el electrodo FC6, pero esta vez se considera la Banda Theta. Este nuevo agrupamiento entrega un mejor resultado, que equivale a un 80% de precisión. Los objetos encontrados en esta oportunidad están en la tabla 5.8.

Tiempo y Estadísticos Típicos: Si se toma el tiempo junto con la varianza de las señales por objeto, para el canal FC6, el resultado de la *clusterización* no es bueno, es decir, no es posible distinguir claramente tres grupos distintos de objetos como se espera. Esto se puede ver en la figura 5.11, mientras que, usando la media de las señales para todos los canales, se obtiene un resultado similar, mostrado en la figura 5.12.

Objeto
35
94
195
215
218
243
244
245
265
267
269
271
295
296
306
307
341
343
357
359

Tabla 5.7: Objetos relevantes usando Porcentaje de Energía Banda Delta y Tiempo (PZ).
Fuente: Elaboración propia.

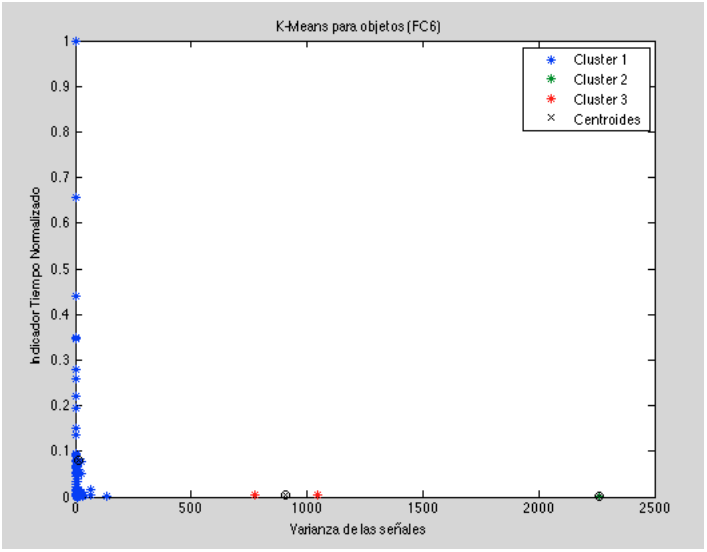

Figura 5.11: K-Means para objetos utilizando Varianza y Tiempo (FC6).
Fuente: Elaboración propia.

Objeto
267
94
215
244
271
245
306
356
307
87
295
269
266
357
359
73
23
265
214
195

Tabla 5.8: Objetos relevantes usando Porcentaje de Energía Banda Theta y Tiempo (FC6).
Fuente: Elaboración propia.

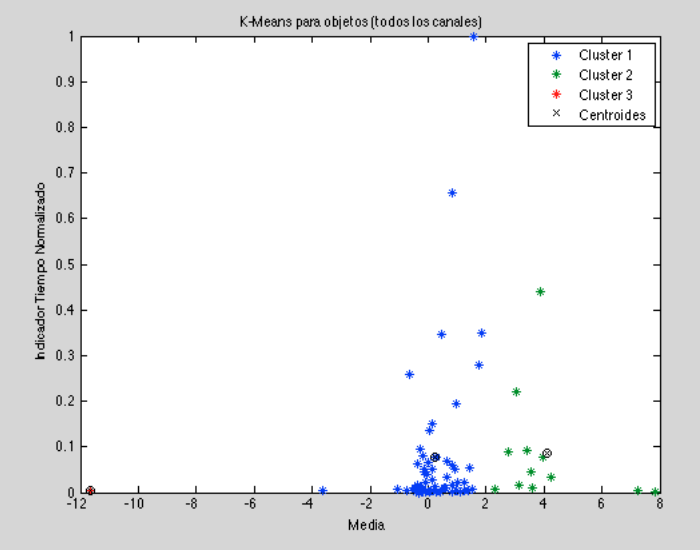

Figura 5.12: K-Means para objetos utilizando Media y Tiempo (Todos los canales).
Fuente: Elaboración propia.

Con estos resultados se descarta seguir utilizando los estadísticos típicos para la clasificación.

Tiempo y RMS: Haciendo uso de estas variables como dimensiones de clasificación, los resultados no son acorde a lo esperado para la banda Delta en el electrodo FC6 5.13. Ocurre algo similar con la banda Theta en el electrodo PZ.

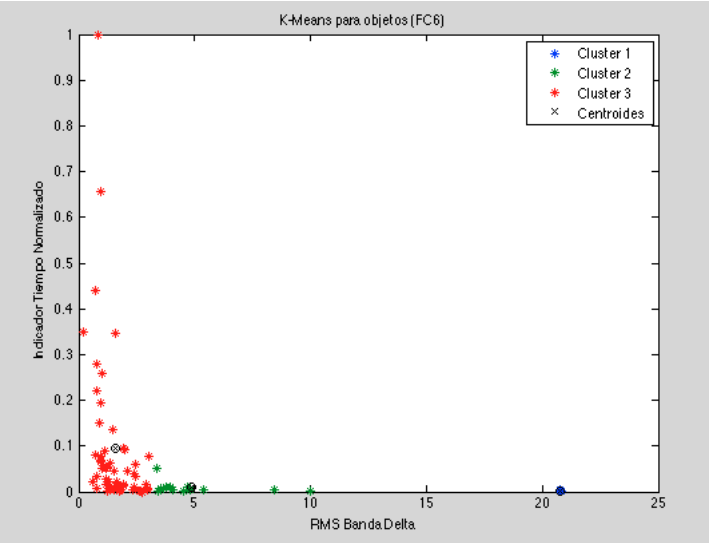

Figura 5.13: K-Means para objetos utilizando RMS de la Banda Delta y Tiempo (FC6).
Fuente: Elaboración propia.

Tiempo y Dimensión Fractal: Cuando las variables de clasificación son el tiempo y la dimensión fractal obtenida con el algoritmo Higuchi, para el canal PZ, la separación de objetos en tres grupos, propone que la variable predominante es el tiempo, pero no es posible distinguir diferencias claras entre los valores de la dimensión fractal (imagen 5.14). Por lo tanto, se descarta esta variable para terminar el proceso de clasificación.

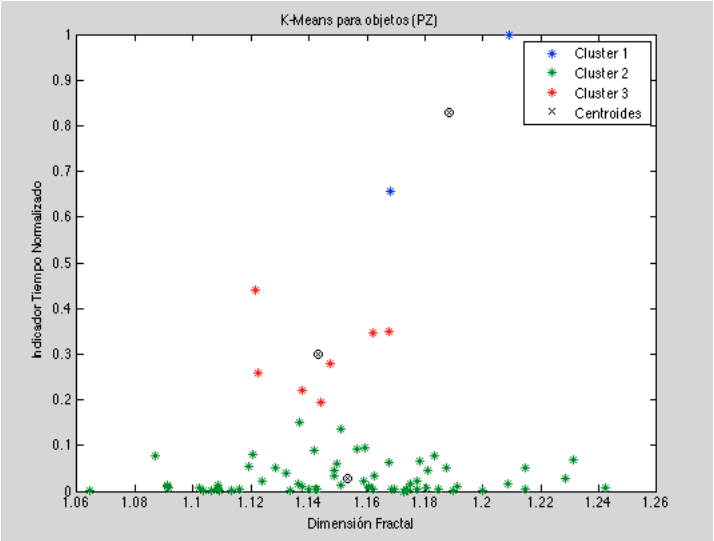

Figura 5.14: K-Means para objetos utilizando Dimensión fractal y Tiempo (PZ).
Fuente: Elaboración propia.

Con todos estos resultados se finaliza el estudio utilizando como primera variable el tiempo. A continuación se desarrolla un análisis similar, incluyendo esta vez la dimensión que incluye el efecto en el comportamiento de la pupila del sujeto, es decir, el indicador Delta. Esta variable es usada en conjunto con el tiempo, de manera que al igual que anteriormente, se crea un indicador definido como el producto de ambas variable.

Tiempo*Delta y Energía: Como se vio anteriormente, la energía y el tiempo otorgaban un 60% de precisión. Se analizará si agregando el efecto en la dilatación pupilar produce un mejoramiento. Como se ve en la figura 5.15, no es posible sacar conclusiones, usando estas variables para la clasificación.

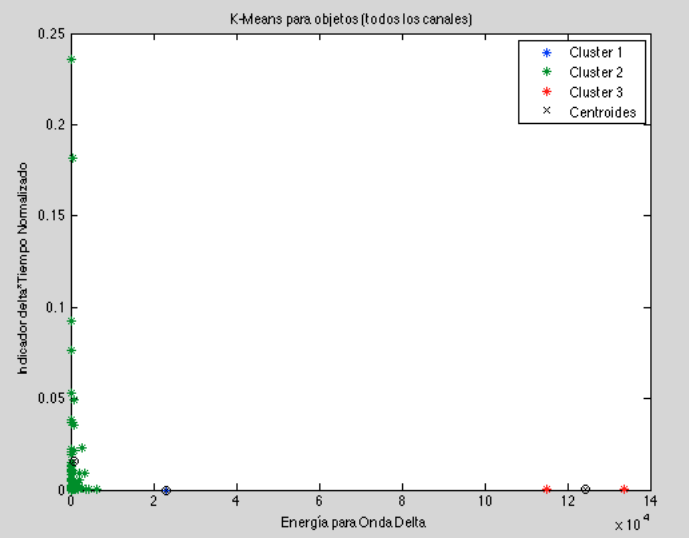

Figura 5.15: K-Means para objetos utilizando Energía Banda Delta y Tiempo*Indicador Delta (todos los canales).
Fuente: Elaboración propia.

Tiempo*Delta y Porcentaje de Energía: Se analiza para el canal FC6 en la banda Delta, donde nuevamente se ve una clara separación clara en tres grupos de objetos (imagen 5.16). Entonces, se cuentan las vistas, se crea el nuevo indicador, definido como el producto de las dos variables de clasificación iniciales y se ordena de mayor a menor. Los resultados están en la tabla 5.9. Esta clasificación posee una precisión del 70%.

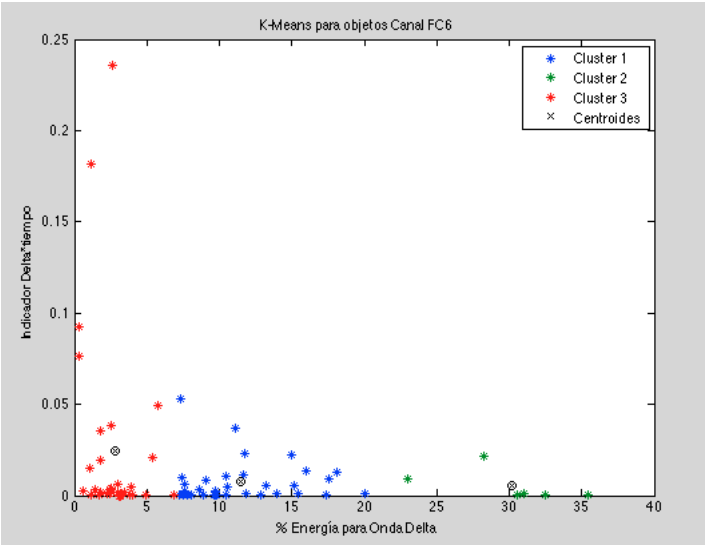

Figura 5.16: K-Means para objetos utilizando Porcentaje de Energía Banda Delta y Tiempo*Indicador Delta (FC6).
Fuente: Elaboración propia.

Objeto
94
307
267
306
356
215
87
359
244
73
214
35
271
266
195
245
74
295
88
269

Tabla 5.9: Objetos relevantes usando Porcentaje de Energía Banda Delta y Tiempo*Indicador Delta (FC6).
Fuente: Elaboración propia.

Finalmente se analiza usando la Banda Theta. Anteriormente se descubrió que la combinación Tiempo y porcentaje de energía en esta banda, entregaba el mejor resultado de todas

las combinaciones realizadas, con una precisión del 80%. Con esta nueva combinación se espera, entonces, al menos igualar esa precisión y en lo posible superarla.

Se puede ver que para el primer *clustering* los datos siguen comportándose de la manera esperada; esto puede ser notado en la imagen 5.17. Por lo tanto, se procede a terminar la clasificación de manera análoga a casos anteriores. El resultado de este procedimiento arroja un 90% de precisión, siendo el más alto que se obtuvo en este trabajo. El listado de los objetos encontrados está en la tabla 5.10, donde del total de los objetos, sólo dos no coinciden con la lista de los *Website Keyobjects* originales tomados como base de análisis.

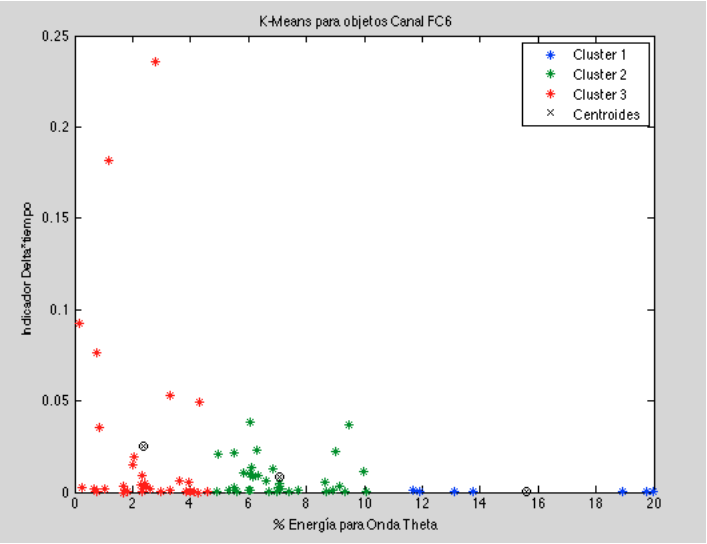

Figura 5.17: K-Means para objetos utilizando Porcentaje de Energía Banda Theta y Tiempo*Indicador Delta (FC6).
Fuente: Elaboración propia.

5.1.3 Discusión

Primero, es importante destacar que todos los resultados obtenidos en la parte de minería de datos de los datos del EEG son en comparación con la lista base de objetos relevantes derivada en este mismo trabajo. No se trabaja con los objetos encontrados por Martínez (A), ya que durante el último tiempo surgieron ciertas ambigüedades con el veredicto final del *webmaster*, en particular en la clasificación de los objetos encontrados por Jorge Dupré en su trabajo de memoria.

Los objetos encontrados en 5.1.1, difieren de los de Martínez, pero lo importante es que de todas maneras, en esta ocasión no se clasificaron como relevantes los objetos que el *webmaster* rechazó de ese trabajo. Es decir, el presente estudio corrobora la opinión del experto del negocio.

Objeto
94
267
245
244
215
356
306
307
266
73
195
269
265
295
88
357
23
84
243
309

Tabla 5.10: Objetos relevantes usando Porcentaje de Energía Banda Theta y Tiempo*Indicador Delta (FC6).
Fuente: Elaboración propia.

Acerca de los indicadores utilizados en la parte 5.1.2 cumplen de buena manera su función de clasificar objetos relevantes dentro del sitio web en estudio. Se ve que utilizando algunas características de las señales cerebrales derivadas por medio de diversos métodos, se pueden obtener clasificaciones con distintos resultados, como se vio en la parte anterior.

6 Conclusiones

En este capítulo se describen las conclusiones finales de lo que fue este trabajo de memoria. En la primera sección se explicitan las conclusiones generales del proyecto, los resultados, el cumplimiento de objetivos y validación de hipótesis. En la segunda sección se entregan algunas recomendaciones y líneas de trabajo futuro.

6.1 Conclusiones Generales

El presente trabajo forma parte de las investigaciones del proyecto AKORI, en el que se une la expertiz en el área biológica con el conocimiento de la ingeniería. Este proyecto está en marcha bajo la tutela del Laboratorio de Neurosistemas de la Facultad de Medicina y el Departamento de Ingenería Industrial de la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Chile.

La hipótesis de investigación de este trabajo planteaba el desafío de responder si se podía clasificar o distinguir entre dos tipos de objetos dentro de un sitio web, llámese relevantes y no relevantes, utilizando variables para caracterizar la actividad bioeléctrica cerebral. Los resultados obtenidos validan la hipótesis, de manera que efectivamente la medición de la bioactividad cerebral es una buena fuente de información para este tipo de análisis.

Para validar la hipótesis se ha propuesto un análisis de datos generados por un dispositivo de *Eye Tracking* y electroencefalografía, mediante la metodología KDD. Como primer paso, se realizó una exhaustiva investigación de las técnicas existentes y del estado del arte de la investigación de emocionalidad y bioactividad cerebral con impulsos relacionados con los intereses de usuarios de sitios web.

Se diseñó e implementó un experimento que permitió generar una base de datos con la que se desarrolló la investigación. Despues de aplicar el proceso KDD y en base a lo resultados obtenidos se pueden declarar las siguientes conclusiones:

1. Utilizando la misma metodología que en el trabajo de Martínez en [4], se obtuvo un resultado diferente. Sin embago, esto no es considerado como un fracaso, pues existen varias razones por lo que esto puede haber sucedido. De estas, la más importante es la

diferencia entre el grupo de personas utilizadas en ambos experimentos y la ambigüedad con que se termina por analizar en el caso anterior. Es de cierta manera ilógico que se desarrolle una metodología compleja para objetivizar la identificación de los elementos más importantes en un sitio web, para que al final del camino la opinión subjetiva de una persona sea la que termine por definir el resultado.

2. La utilización de diversas formas de caracterizar las señales bioeléctricas cerebrales, en particular, los potenciales eléctricos de distintas zonas de la cabeza, es un buen apronte para la identificación de los objetos relevantes en un sitio web. Se demostró que existen descomposiciones utilizadas en la literatura, que permiten parametrizar las señales para posteriormente realizar metodologías de clasificación.
   En este trabajo, la mejor combinación de variables fue la del indicador Delta*Tiempo y el porcentaje de energía para la banda Theta para el electrodo FC6, obteniendo un 90% de precisión en comparación con la lista base derivada en este trabajo.
3. Utilizar sólo el tiempo y alguna variable de actividad cerebral puede dar buenos resultados, en este caso fue un 60%, que puede ser mejorado incluyendo la componente de efecto sobre la dilatación pupilar. Esto es análogo al resultado de Martínez que indica que la dilatación por sí sola agrega valor, pero puede ser mejorado con la utilización de otras variables, por ejemplo de actividad cerebral.
4. En cuanto a la obtención de rasgos de emocionalidad de cada uno de los objetos del sitio web, cabe destacar que no se obtuvo una clasificación a base de emociones. Para llevar a cabo una clasificación de ese tipo, se requiere contar con etiquetas que representen la valencia emocional de los estímulos (en este caso, los objetos del sitio), y posteriormente aplicar técnicas de clasificación. El sitio ocupado en esta ocasión no presenta valencias positivas o negativas de una manera clara, más bien, al ser un sitio informativo, sus objetos son en su mayoría, textos y fotos explicativas, de carácter neutro.
5. El uso del electroencefalograma genera un impacto positivo en la identifiación de los *Website Keyobjects*, aportando con un mayor grado de objetividad a la metodología. Este impacto puede ser aún mayor si se puede extender y llegar a lograr clasificaciones de estados emocionales reales a partir de los estímulos que entregan los sitios web.

6.2 Recomendaciones y Trabajo Futuro

Para cerrar este trabajo de memoria, se entregan algunas recomendaciones para tener en cuenta en eventuales continuaciones del proyecto AKORI o líneas de investiación similares.

- **Sitio Web:** El sitio web es una parte fundamental en trabajos de este tipo. Si se quieren

obtener resultados que definan de mejor manera el comportamiento de los usuarios, es necesario contar con un sitio web real en los experimentos. En el caso de este estudio, el sitio web utilizado era una adaptación mediante una secuencia de imágenes del sitio web real. Si bien se les daba la instrucción de navegar libremente a los usuarios, de todas maneras era un acción forzada y no reflejaba en su totalidad a la navegación que tendría en la realidad el usuario.

- **Determinación de Objetos:** Otro ámbito fundamental para lograr este trabajo es definir los objetos y sus posiciones para saber dónde está poniendo su atención el usuario. Este trabajo fue realizado completamente a mano por Martínez, por lo que tomó una gran cantidad de tiempo y nuevamente, un cierto grado de subjetividad. Es necesario contar con un sistema autónomo que pueda descomponer un sitio en objetos de forma sencilla. Además, así se tendrá la posibilidad de analizar varios sitios a la vez o de manera más rápida.
- **Facilidad de Experimentar y Segmentación:** Uno de los cuellos de botella de este trabajo fue poder contar con la base de datos necesaria para el análisis. Los experimentos duraban cerca de dos horas, considerando preparación, toma de muestra y limpieza de instrumentos. Por lo tanto, si en el futuro se quiere detectar patrones de comportamiento por segmentos, es necesario contar con una forma más fácil y rápida de tomar las muestras. Se propone el uso del EEG Epoc para disminuir tiempos de preparación y experimentación.
- **Ambigüedades:** Como se menciona antes, en la opinión del autor de este trabajo, resulta en cierto modo contraproducente el término de la investigación utilizando la opinión subjetiva de una persona, por más que se trate de un experto. Si se han desarrollado técnicas y algoritmos que analizan respuestas fisiológicas de las personas, y que tienen una base científica robusta, no debiese validarse el resultado por la opinión de una persona, que además se ha demostrado puedo ir cambiando y generando ambigüedades como las exhibidas en este trabajo.

6.3 Reflexión Final

Una vez finalizado el trabajo de memoria, se han mencionado diferentes conclusiones y entregado recomendaciones referentes al proyecto en sí, a la identificación de los *Website Keyobjects*. Pero más allá de ésto, es posible dar origen a otro tipo de reflexiones o pensamientos que permiten ligar este estudio con otros ámbitos.

Los volúmenes de datos analizados en este trabajo correspondientes a señales eléctricas cerebrales, señales de dilatación pupilar y posicionamiento ocular, son inmensos. Muchas

veces tardaron horas en diversos procesamientos y requirieron el uso demás de un computador en estas tareas. Los algoritmos usados son similares a los que se usan en otros campos de investigación, que adaptados a esta realidad particular mostraron ser útiles y eficientes.

Lo que se quiere decir es que el conocimiento desarrollado en este trabajo, el análisis de grandes cantidades de datos y la derivación de técnicas y metodologías puede ser usado en variados problemas o campos para cambiar o mejorar las situaciones que existen hoy en día. Por ejemplo, en la salud, donde los registros electrónicos guardan información importante de los pacientes, sus enfermedades y síntomas, se tiene una cantidad considerable de datos en forma de texto que puede ser analizada de forma similar para obtener patrones de comportamiento, predecir enfermedades e incluso generar atenciones a distancia y en el momento preciso, es decir, es posible proponer mejoras en el sistema de salud pública.

Otro ejemplo, es la astronomía; Chile está posicionado como una potencia mundial en este ámbito, donde nuevos proyectos en el país significan una muchísima mayor cantidad de datos a analizar para seguir creando conocimiento del universo, por lo tanto, se necesitará del uso de este tipo de algoritmos y tecnología para avanzar.

Ejemplos como estos hay muchos más, lo importante es saber que a medida que pasa el tiempo los volúmenes de datos creados por los sistemas han ido creciendo exponencialmente y el haber trabajado en este proyecto de forma similar abre el camino para aplicar lo aprendido en otras áreas que no necesariamente están relacionadas con la ingeniería industrial, y de esa manera crear valor en distintos campos, en el país y en la sociedad.

Bibliografía

[1] L. E. Dujovne and J. D. Velásquez, "Design and implementation of a methodology for identifying website keyobjects," in *Knowledge-Based and Intelligent Information and Engineering Systems*, pp. 301–308, Springer Berlin Heidelberg, 2009.

[2] O. Maimon and L. R. (Eds.), *The Data Mining and Knowledge Discovery Handbook*. Springer, 2005.

[3] J. Jadue, "Incidencia de la dilatación pupilar como variable predictiva del comportamiento de los usuarios en una página web antes de tomar una decisión," *Universidad de Chile*, 2014.

[4] G. Martínez, "Mejoramiento de una metodología para la identificacion de website keyobjects mediante la aplicación de tecnologias eye-tracking, análisis de dilatación pupilar y algoritmos de web mining," *Universidad de Chile*, 2013.

[5] Miniwatts Marketing Group, "Estadísticas de uso de internet y sitios web." `http://www.internetworldstats.com/stats.htm`, 2013. Visto en 10/11/2013.

[6] J. D. Velásquez, R. Weber, H. Yasuda, and T. Aoki, "A methodology to find web site keywords," in *Procs. IEEE Int. Conf. on e-Technology, e-Commerce and e-Service*, (Taipei, Taiwan), pp. 285–292, 2004.

[7] L. González, "Mejoramiento de una metodología para la identificacion de website keyobjects mediante la aplicación de tecnologias eye-tracking y algoritmos de web mining," *Universidad de Chile*, 2011.

[8] J. D. Velásquez and L. Donoso, "Aplicación de técnicas de web mining sobre los datos originados por usuarios de páginas web. visión crítica desde las garantías fundamentales, especialmente la libertad, la privacidad y el honor de las personas," *Revista de Ingeniería de Sistemas*, vol. 24, pp. 47–68, June 2010.

[9] W. W. W. Consortium, "Basic HTTP as defined in 1992." `http://www.w3.org/Protocols/HTTP/HTTP2.html/`, 1992. Visto en 08/12/2013.

[10] W. W. W. Consortium, "Introduction to HTML 4." `http://www.w3.org/TR/1999/REC-html401-19991224/intro/intro.html`, 1999. Visto en 08/12/2013.

[11] W. W. W. Consortium, "Uniform Resource Locators." `http://www.w3.org/Addressing/URL/url-spec.html`. Visto en 08/12/2013.

[12] J. D. Velásquez, "Combining eye-tracking technologies with web usage mining for identifying website keyobjects," *Engineering Applications of Artificial Intelligence*, vol. 26, pp. 1469 – 1478, 2013.

[13] J. Velásquez and L. Jain, *Advanced Techniques in Web Intelligence - Part 1*. No. v. 1 in Studies in Computational Intelligence, Springer, 2010.

[14] L. González and J. Velásquez, "Una aplicación de herramientas de eye-tracking para analizar las preferencias de contenido de los usuarios de sitios web," *Revista de Ingeniería de Sistemas*, vol. 26, pp. 95–118, September 2012.

[15] F. J. A. Fernández, J. F. Pérez, and S. F. López, *Oftalmología en Atención Primaria*. Alcalá, 2002.

[16] H. Kolb, E. Fernandez, and R. Nelson, *Webvision: The Organization of the Retina and Visual System*. Salt Lake City (UT): University of Utah Health Sciences Center, 1995.

[17] F. E. Léeon-Sarmiento, D. G. Prada, and C. Gutiérrez, "Pupila, pupilometría y pupilografía," in *Acta Neurol Colomb, Volumen 24*, pp. 188–197, Neurol Colomb, 2008.

[18] E. Kandela, J. Schwartz, and T. Jessell, *Principles of Neural Science*. McGraw-Hill, 2000.

[19] C. Aracena, "Estudio de la relación entre neurodatos, dilatación pupilar y emocionalidad basado en técnicas de minería de datos," *Universidad de Chile*, 2014.

[20] S. J. Luck, *An Introduction to Event-Related Potential Technique*. The MIT Press, 2005.

[21] E. Niedermeyer and F. L. da Silva, *Electroencephalography , 5th Edition*. Lippincott Williams and Wilkins, 2005.

[22] J. T. Cacioppo, L. G. Tassinary, and G. G. Bertson, *Handbook of Psychophysiology, 3th edition*. Cambridge University Press, 2007.

[23] O. Sourina and Y. Liu, "A fractal-based algorithm of emotion recognition from eeg using arousal-valence model," in *Biosignals*, pp. 209–214, 2011.

[24] A. Keil, M. Bradley, T. Elbert, and P. Lang, "Large-scale neural correlates of affective picture-processing," *Psychophysiology*, p. 39:641–649, 2002.

[25] M. C. Pastor, M. M. Bradley, and P. J. Lang, "Affective picture perception: Emotion, context, and the late positive potential," *Brain Research*, pp. 1189:145–151, 2008.

[26] B. Herbert, O. Pollatos, and R. Schandry, "Interoceptive sensitivity and emotion processing: An eeg study," *International Journal of Psychophysiology*, p. 65:214–227, 2007.

[27] K. Ishino and M. Hagiwara, "A feeling estimation system using a simple electroencephalograph," *Systems, Man and Cybernetics*, pp. 4204–4209, 2003.

[28] K. Takahashi, "Remarks on emotion recognition from multi-modal bio-potential signals," *Industrial Technology*, pp. 1138–1143, 2004.

[29] G. Chanel, J. Kronegg, D. Grandjean, and T. Pun, "Emotion assessment: Arousal evaluation using eeg's and peripheral physiological signals," in *Multimedia content representation, classification and security*, pp. 530–537, Springer, 2006.

[30] G. Chanel, J. Kierkels, M. Soleymani, and T. Pun, "Short-term emotion assessment in a recall paradigm," *International Journal of Human Computer Studies*, pp. 67:607–627, 2009.

[31] Y. P. Lin, C. H. Wang, T. L. Wu, and S. K. Jeng, "Eeg-based emotion recognition in music listening: A comparison of schemes for multiclass support vector machine," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 489–492, ICASSP, 2009.

[32] Z. Khalili and M. H. Moradi, "Eeg-based emotion recognition in music listening: A comparison of schemes for multiclass support vector machine," in *Proceedings of the International Joint Conference on Neural Networks*, pp. 1571–1575, 2009.

[33] Q. Zhang and M. Lee, "Analysis of positive and negative emotions in natural scene using brain activity and gist," *Neurocomputing*, pp. 72:1302–1306, 2009.

[34] P. C. Petrantonakis and L. J. Hadjileontiadis, "Emotion recognition from eeg using higher order crossings.," *IEEE Transactions on Information Technology in Biomedicine*, pp. 14:186–197, 2010.

[35] O. Sourina, Q. Wang, and Y. Liu, "A real-time fractal-based brain state recognition from eeg and its applications," in *Biosignals*, pp. 82–91, 2011.

[36] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI magazine*, vol. 17, pp. 37–54, 1996.

[37] B. Liu, *Web Data Mining*. Springer, 2007.

[38] G. Myatt, *Making sense of data: A practical guide to exploratory data analysis and data mining*. John Wiley, 2007.

[39] R. Christensen, *Log-Linear Models and Logistic Regression, 2th Edition*. Springer, 1997.

[40] R. Agrawal, R. Srikant, *et al.*, "Fast algorithms for mining association rules," in *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, vol. 1215, pp. 487–499, 1994.

[41] A. I. Mora, "Guía para elaborar una propuesta de investigación," *Educación*, vol. 2, pp. 67–97, 2005.

[42] Emotiv, "Sitio oficial de Emotiv." `http://www.emotiv.com/`, 2013. visto en 08/04/2014.

[43] M. Murugappan, M. Rizon, R. Nagarajan, S. Yaacob, I. Zunaidi, and D. Hazry, "Eeg feature extraction for classifying emotions using fcm and fkm," *International Journal of Computers and Communications*, vol. 1, no. 2, pp. 21–25, 2007.

[44] M. Murugappan, M. Rizon, and R. Nagarajan, "Time-frequency analysis of eeg signals for human emotion detection," in *Biomed*, p. 262–265, 2008.

[45] L. Zou, R. Zhou, S. Hu, J. Zhang, and Y. Li, "Single trial evoked potentials study during an emotional processing based on wavelet transform," in *Advances in Neural Networks - ISNN 2008* (F. Sun, J. Zhang, Y. Tan, J. Cao, and W. Yu, eds.), vol. 5263 of *Lecture Notes in Computer Science*, pp. 1–10, Springer Berlin Heidelberg, 2008.

Apéndices

A Lista de *Website Keyobjects* obtenidos por Martínez

Object Number	Website object	Name	Is keyobject?
1	Object 5	header picture	Yes
2	Object 356	picture detalle profe 1	Yes
3	Object 6	navigation menu	Yes
4	Object 195	paragraph detalle noticia 1	Yes
5	Object 4	logo mba	No
6	Object 267	paragraph quienes 1	Yes
7	Object 306	paragraph metodologia 1	Yes
8	Object 244	image plataforma 1	Yes
9	Object 1	ingenieria industrial logo	No
10	Object 265	paragraph corporacion 1	Yes
11	Object 307	paragraph metodologia 2	Yes
12	Object 215	paragraph propuesta 1	Yes
13	Object 243	paragraph plataforma 1	Yes
14	Object 277	title contenido 1	Yes
15	Object 245	paragraph perfil egresados 1	Yes
16	Object 130	title elegirnos 10	No
17	Object 3	search form	No
18	Object 59	titulo cuerpo	Yes
19	Object 23	main post text 1	Yes
20	Object 22	main post picture	Yes

Tabla 6.1: Lista de *Website Keyobjects* de Martínez

B Consentimiento Informado

Ver la siguiente página.

COMITÉ DE ÉTICA
INVESTIGACIÓN EN SERES HUMANOS
FACULTAD DE MEDICINA
UNIVERSIDAD DE CHILE

CONSENTIMIENTO INFORMADO

TITULO

Nombre del Investigador principal: Dr. Pedro Maldonado A.
Institución: Programa de Fisiología y Biofísica, ICBM, Facultad de Medicina, Universidad de Chile.
Teléfonos: 9786035

Se le entregará una copia del consentimiento informado completo.

Introducción

Mi nombre es Enzo Brunetti. Mi profesión es médico-cirujano. Poseo un doctorado en ciencias biomédicas y actualmente llevo a cabo el proyecto de investigación al cual usted ha sido invitado, en la Facultad de Medicina de la Universidad de Chile. A lo largo de la lectura de este documento usted es libre de manifestar cualquiera de sus inquietudes respecto al procedimiento que se llevará a cabo, tanto hacia mi como con alguien con quien usted se sienta cómodo. Asimismo, puede tomarse el tiempo que requiera para reflexionar respecto a si desea participar del proyecto. Si no entiende alguna(s) de las informaciones contenidas en el presente documento, siéntase libre de expresármelo en cualquier momento para explicarle la información en detalle.

Invitación a participar: Le estamos invitando a participar en el proyecto de investigación "Rol de la respuesta autonómica durante las emociones como determinante de la integración sensorial rápida y la conducta motora".

Objetivos: Esta investigación tiene por objetivo estudiar los cambios en el diámetro de la pupila que están asociados a la actividad que genera el cerebro durante la visión de imágenes con contenido emocional. El total de sujetos propuesto para realizar este estudio es de 40 personas.

Procedimientos: Si usted acepta participar de la investigación propuesta usted estará aceptando ser sometido, por una sola vez, al siguiente procedimiento: la medición de la variación del diámetro de su pupila mediante un sistema de cámaras que van adosados a su cabeza, junto con la medición de la actividad que genera el cerebro a través de electrodos que se ubicarán sobre la superficie de su cuero cabelludo

(electroencefalograma). Ambos registros son superficiales, esto es, no invasivos, y no producen daño ni efectos adversos. Durante todo el experimento sólo se *medirán* parámetros provenientes de usted. En ningún momento se le administrará ningún tipo de energía así como ningún tipo de fármaco. Durante la tarea se le presentarán imágenes que presentan distintos niveles de contenido emocional, de carácter positivo, negativo o neutro. Algunas de esas imágenes pueden ser de alto contenido emocional. Las imágenes serán presentadas secuencialmente y de manera aleatoria, intercaladas con imágenes que no contienen información emocional ni visual conocida. Las imágenes deben vistas libremente por usted. Después de la presentación de cada imágen, la única tarea que requiere realizar el la categorización del tipo de imagen presentada según las categorías de *positivo*, *negativo* o *neutro*. Usted es libre de retirarse de la tarea en cualquier momento a lo largo de esta, aunque no haya llegado a su fin. El investigador se encontrará en la misma sala que usted a lo largo de toda la tarea, y usted puede solicitar de él en todo momento cualquier información o expresar cualquier necesidad que estime pertinente.

Riesgos: Bajo los sistemas de registro que utilizaremos no existen riesgos ni efectos adversos conocidos. Ambos registros mencionados son ampliamente utilizados en el mundo entero para fines de investigación, como en el caso de la tarea a la cual usted será sometido.

Costos: Las técnicas utilizadas en este proyecto no tienen costo alguno para Ud.

Beneficios: Los beneficios del presente estudio no irán en beneficio directamente de usted. El presente proyecto tiene por objetivo contribuir al conocimiento científico de cómo las emociones son procesadas por el cerebro humano.

Alternativas: La decisión de no participar del presente estudio no significara ningún perjuicio para su persona.

Compensación: No se considera la entrega de una compensación económica para usted en el presente estudio.

Confidencialidad: Toda la información derivada de su participación en este estudio será conservada en forma de estricta confidencialidad, lo que incluye el acceso de los investigadores o agencias supervisoras de la investigación. Cualquier publicación o comunicación científica de los resultados de la investigación será completamente anónima.

Voluntariedad: Su participación en esta investigación es totalmente voluntaria y se puede retirar en cualquier momento comunicándolo al investigador.

Complicaciones: Aunque no existe riesgo asociado descrito para el procedimiento a realizar, en el caso de que usted. presente complicaciones directamente dependientes de la aplicación de las técnicas utilizadas

en este estudio, usted recibirá el tratamiento médico completo de dicha complicación, financiado por el proyecto al cual se asocia este estudio, y sin costo alguno para Ud. o su sistema previsional.

Derechos del participante: Si Ud. requiere cualquier otra información sobre su participación en este estudio puede llamar a:

Investigador: Enzo Paolo Brunetti; teléfono 6 236 64 07
Autoridad de la Institución: Dr. Pedro Maldonado A; teléfono: (56 2) 978 60 35

Conclusión:
Después de haber recibido íntegramente y comprendido la totalidad de la información contenida en este documento, no teniendo actualmente ninguna duda respecto a la tarea a realizar, las técnicas de medición, así como los riesgos asociados, otorgo mi consentimiento para participar en el proyecto "Rol de la respuesta autonómica durante las emociones como determinante de la integración sensorial rápida y la conducta motora".

Nombre del sujeto	Firma	Fecha
Nombre de informante	Firma	Fecha
Nombre del investigador	Firma	Fecha