Universidades de Burgos, León y Valladolid

Máster universitario

Inteligencia de Negocio y Big Data en Entornos Seguros

TFM del Máster Inteligencia de Negocio y Big Data en Entornos Seguros

Investigación y Desarrollo de un Sistema de Reconocimiento Biométrico mediante Dispositivos Ponibles (Wearables)

Presentado por Irene Salvador Ortega
en Universidad de Valladolid — 29 de noviembre de 2020
Tutores: D. Carlos Vivaracho Pascual
Dña. María Aránzazu Simón Hurtado

Agradecimientos

Hace tiempo, en Pinterest [59], encontré una frase sin autor que decía, "Jamás pares de aprender, la vida nunca para de enseñar". De mi aprendizaje, hoy doy gracias a la Universidad de Valladolid, por acompañarme en mis 5 años de Grado y por hacerlo, de nuevo, un año más, con el Máster.

Este Trabajo Fin de Máster es fruto de la ayuda de muchas personas que han estado presentes en mi vida, desde profesores que me han permitido aprender lo que hoy sé, hasta familiares y amigos que no han permitido que me sintiera sola en ningún momento, a pesar de las malas circunstancias. Me gustaría dedicaros unas palabras de agradecimiento.

A mis tutores, Carlos Vivaracho y Mª Aránzazu Simón, sin vuestra ayuda y consejos no podría haber llegado tan lejos. Muchas gracias por haberme guiado y apoyado, en todo momento. Hemos formado un gran equipo juntos, sin duda, os volvería a elegir. Espero que este proyecto que empezó hace 2 años, continúe hasta conseguir el lugar que se merece.

A mi familia, especialmente a mis padres, por acompañarme en cada locura y aceptar mis decisiones a pesar de no estar de acuerdo. Gracias por confiar siempre en mí y ayudarme en todo lo que podéis. Me hacéis la vida mucho más fácil y ojalá siga siendo así muchos años más. A mi hermano, gracias por a pesar de la distancia y las discusiones, estar siempre a mi lado y ser como esa sombra que me acompaña para en los momentos más difíciles, ayudarme.

A Raúl, por ser uno más de la familia. Tú, mejor que nadie, conoces el esfuerzo que ha habido detrás de este trabajo. Gracias por haberme hecho la vida mucho más fácil y haber aportado ese toque de humor, incluso cuando ni tú lo tenías para ti.

A mis compañeros de trabajo, a vosotros también os doy las gracias de mi aprendizaje. Ha sido un año muy intenso. Gracias por haber confiado en mí y haberme enseñado todo lo que sabíais. A Gustavo por ser mi complemento en las 9 horas diarias de trabajo y ayudarme con cada duda.

A cada persona que ha participado en este proyecto como sujeto, andando durante una hora: familiares, amigos y compañeros de

trabajo. También formáis parte de este proyecto y sin vosotros, no habría sido posible.

A mis abuelos, porque sé que me guiais y ayudáis en cada decisión, intentando que salga de la mejor manera posible. Nunca os olvidaré. Sois las estrellas que más brillan en el cielo y sin duda, todo esto va por vosotros.

Al COVID por enseñarme la fragilidad de la vida y la necesidad de aprovechar cada momento, antes de que se acabe.

A todos, muchas gracias.

Resumen

El continuo aumento en el mercado de los dispositivos portátiles ha ido acompañado de una mejora en las capacidades y sensores incorporados a ellos. Por esta razón, su uso y el trabajo relacionado van más allá del seguimiento de la actividad o para complementar un teléfono inteligente.

En el presente proyecto se va a trabajar con una biometría emergente basada en las características del comportamiento del ser humano que permite verificación no intrusiva, continua, fácil de conseguir y difícil de robar o falsificar. El objetivo principal es determinar si el uso de los sensores presentes en los dispositivos ponibles puede permitir o no la verificación biométrica de personas mediante su forma de caminar y proponer un sistema de reconocimiento para su propósito.

Ya existen trabajos en el tema que serán usados como referencia, pero la gran mayoría utilizan IMUs (Unidades de Medida Inercial) construidas de manera específica o dispositivos comerciales no portátiles para simularlos como, por ejemplo, el mando de Wii. Aquí se utiliza un reloj y una pulsera inteligente, ambos dispositivos comerciales.

Este trabajo es una continuación de trabajos previos ya realizados por el grupo de investigación, en los que se desarrolló el sistema móvil de captura de datos, se obtuvo un corpus con el que trabajar y se realizó un análisis profundo de dichos datos, demostrando la existencia de periodicidad y proponiendo un sistema de reconocimiento que se utilizará como partida. Aquí, se van a adquirir nuevos datos y el estudio se va a centrar en la construcción de un sistema automático de limpieza de la señal, un análisis de distintas técnicas de aprendizaje automático ajustando las técnicas de preprocesamiento y diversos parámetros de interés que ayuden a la construcción de un sistema final. Se han logrado resultados muy buenos, no publicados anteriormente, sobre 32 usuarios, coincidiendo en los dos dispositivos comerciales distintos probados. También se demostrará que es irrelevante la mano utilizada para llevar el dispositivo. Todo ello muestra líneas interesantes para futuras investigaciones.

Descriptores

Autenticación biométrica, dispositivos ponibles, dispositivos comerciales, reconocimiento de la forma de andar, análisis de Fourier, base de datos, biometría, dominio de la frecuencia, dominio del tiempo, máquinas de vectores soporte, arquitectura, inteligencia de negocio, ciberseguridad.

Abstract

The continuous increase in the wearable device market has been accompanied by an improvement in the capabilities and sensors incorporated to them. For this reason, their use and related work go beyond activity tracking or for complementing a smartphone.

In this project we will work with an emerging biometrics based on the characteristics of human behavior that allows an unobtrusive, continuous, easy to obtain and difficult to steal or falsify verification. The main objective is to determine if the use of the sensors present in the wearable devices can allow or not the biometric verification of people by their gait and to propose a recognition system for this purpose.

There are already works on the subject that will be used as a reference, but the great majority use specifically built IMUs (Inertial Measurement Units) or commercial non-wearable devices to simulate them, e.g., Wii Remote. A smart watch and bracelet are used here, both commercial devices.

This work is a continuation of previous works already carried out by the research group, in which the mobile data capture system was developed, a corpus was obtained to work with and a deep analysis of data was carried out, demonstrating the existence of periodicity and proposing a recognition system that will be used as a starting point. Here, new data will be acquired and the study will focus on the construction of an automatic signal cleaning system, an analysis of different machine learning techniques adjusting the pre-processing techniques and various parameters of interest that help the building a final system. Interesting outcomes not previously reported have been achieved on 32 users, coinciding on the two different commercial devices tested. The hand used to carry the device will also be shown to be irrelevant. All this shows interesting lines for future research.

Keywords

Biometric user authentication, wearable devices, commercial devices, gait recognition, Fourier Analysis, database, biometrics, frequency domain, time domain, support vector machines, architecture, business intelligence, cybersecurity.

Índice general

Índice de figuras

Índice de tablas

Memoria

Capítulo 1

Introducción

La interacción humano-dispositivo está en continuo crecimiento. En el mercado, cada vez son más los dispositivos portátiles, acompañados de una mejora de sus capacidades y de los sensores incorporados en ellos. El término "Internet de las Cosas" (IoT, Internet of Things) hace referencia a la conexión e integración de dispositivos a internet junto con el análisis de la información obtenida a través de ellos. El año 2020 fue calificado como clave para los 4 componentes del modelo del IoT: Sensores, Redes (Comunicaciones), Analítica y Aplicaciones [8], palabras clave que aparecerán a lo largo de este trabajo.

Respecto a los wearables[1] (pulseras de actividad, relojes inteligentes, etc.), hay quienes afirman que pasarán a fusionarse con los smartphones. Es más cómodo y eficiente tener un dispositivo en la muñeca que en el bolsillo, aunque lograr este hito no será tan fácil [58]. Actualmente, permiten realizar infinidad de funciones: enviar mensajes, transferir dinero, usar el asistente virtual, reproducir música, hacer un seguimiento de la salud y un largo etcétera que lleva a que, con un pequeño dispositivo colocado en la muñeca, el usuario sea capaz de tomar decisiones fundamentales y más acertadas respecto a su salud.

Cuando un usuario desea llevar a cabo cualquier tipo de comunicación o transacción a través de Internet, necesita algún tipo de autenticación. Los sistemas más utilizados actualmente requieren que el usuario se registre y recuerde contraseñas, provocando que este termine realizando

[1]Durante el presente trabajo se va a usar "wearable" en inglés o "ponible" (una posible traducción al castellano) indistintamente para referirnos a dispositivos como pulseras o relojes inteligentes. No existe traducción aceptada para el término "wearable", por lo que en este trabajo nos vamos a permitir la licencia de usar "ponible", porque los relojes o las pulseras en castellano no se "visten", se ponen.

prácticas poco seguras, como emplear contraseñas simples, repetir una misma contraseña en varios sitios web, guardarlas en algún archivo, no cambiarlas con el paso del tiempo, etc. También hay que añadir problemas administrativos derivados de posibles pérdidas de claves, o en el caso de que el usuario necesite llevar consigo alguna tarjeta o dispositivo, esta también se puede perder, ser robada o transferirse.

Una alternativa es el uso de sistemas basados en biometría, donde el usuario no necesita recordar ni llevar consigo nada, empleando únicamente sus características intrínsecas (físicas o de la forma de actuar). Porque todos los seres humanos tenemos características morfológicas únicas que nos diferencian: la huella dactilar, la geometría de partes de nuestro cuerpo como las manos, nuestros ojos, la forma de la cara... Aunque la biometría se lleva aplicando desde finales del siglo XIX para la identificación de las personas con métricas como la huella dactilar, la cual sigue usándose hoy en día. Las mejoras tecnológicas y la necesidad de incrementar y simplificar la identificación de los usuarios han provocado que a lo largo de los últimos años los estudios basados en sistemas de reconocimiento biométrico hayan cobrado una mayor relevancia.

Tradicionalmente, los métodos empleados en reconocimiento biométrico estaban relacionados con las características físicas del individuo como su cara o su huella dactilar. Estas características dan buenos resultados, pero requieren que el usuario ponga su cara o su dedo en algún dispositivo, proceso que termina siendo incómodo para el individuo. Por eso, en la actualidad, y en este proyecto, se busca emplear el comportamiento de la forma de actuar del individuo para el reconocimiento biométrico, ya que es un método menos intrusivo.

Se propone aprovechar el desarrollo y gran éxito en ventas en los últimos años de los dispositivos ponibles comerciales (figura 1.1) para ver si es posible verificar a una persona a partir de su *forma de andar*.

La diferencia entre la forma de abordar el problema en este proyecto y el resto de los trabajos de la bibliografía, exceptuando los tres previos a este ya mencionados, es que se van a utilizar dispositivos comerciales, cuando en la bibliografía se utilizan dispositivos creados ad hoc para el propio propósito del proyecto o smartphones.

La biometría puede ser aplicada para la identificación o la autenticación/verificación de las personas. Los sistemas de verificación son configurados para cada usuario registrado con el objetivo de autenticar la identidad de dicho usuario en una etapa posterior mientras que, en

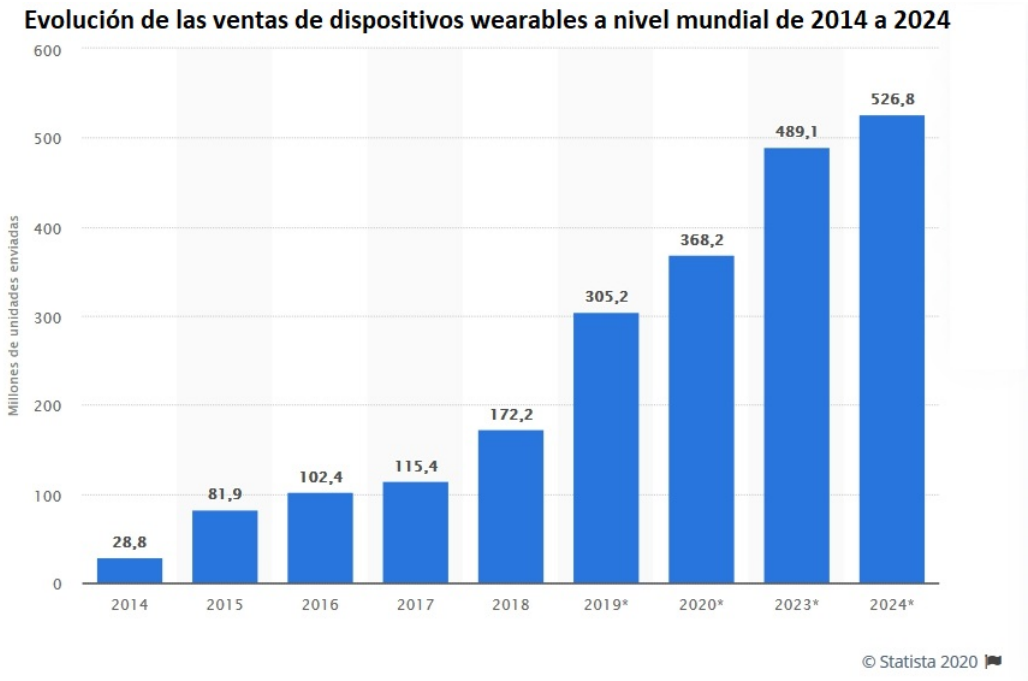


Figura 1.1: Evolución (en millones de unidades) de las ventas de dispositivos wearables a nivel mundial [28].

la identificación, el sistema presenta una señal biométrica y se debe decidir quién es el propietario de esa señal de entre un grupo de usuarios registrados. En otras palabras, en verificación se busca respuesta a la pregunta *Am I who I claim I am? (¿Soy yo quien digo que soy?)*, mientras que la identificación busca respuesta a la pregunta *Who am I? (¿Quién soy yo?)*. En este proyecto, nos vamos a centrar únicamente en la verificación, cuyas fases se muestran, de manera resumida, en el esquema de la figura 1.2.

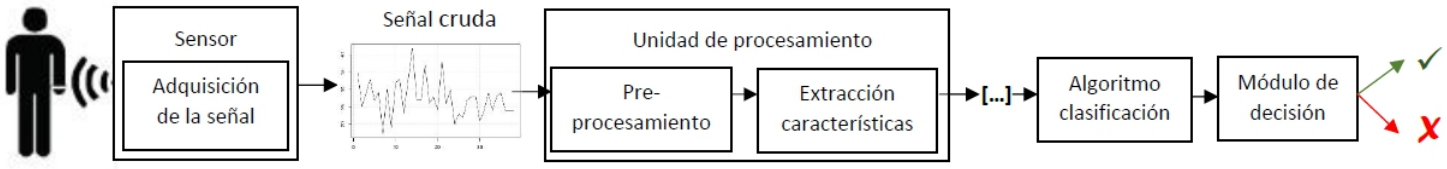


Figura 1.2: Fases involucradas en el sistema de verificación biométrica.

El sensor captura la señal biométrica cruda, se realiza una fase de preprocesamiento y extracción de características de las señales, después [...] indica la generación de los ficheros de salida necesarios para aplicar

el clasificador, el cual generará una métrica, que servirá para evaluar el rendimiento del sistema construido, y posteriormente podrá ser utilizada para tomar una decisión: soy yo o no soy yo el usuario.

Este trabajo es una continuación de tres trabajos previos realizados en la Universidad de Valladolid (UVa), dos llevados a cabo por otros autores y un tercero que fue el Trabajo de Fin de Grado que desarrollé para los Grados en Estadística y en Ingeniería Informática:

- El primero [31] hizo un estudio detallado de todos los sensores de distintos dispositivos ponibles, llegando a la conclusión de que los únicos que proporcionaban una información susceptible de ser usada en biometría para la forma de andar son los que aquí se van a usar: el acelerómetro y el giroscopio. Además, se construyó una aplicación Android (APK) que permitía la recogida de los datos de dos dispositivos seleccionados, que también se utilizarán en el presente trabajo.
- El segundo [4] recogió datos de diversos usuarios y realizó un estudio preliminar con ellos que demostraba la existencia de periodicidad en la señal de los datos y resultados positivos que indicaban su posible uso en biometría.
- El tercero [57] realizó un análisis más profundo de los datos recogidos en [4], comparando diferentes configuraciones del sistema y el rendimiento entre dispositivos, sensores y sesiones, logrando resultados interesantes, que afirmaban la posibilidad de utilizar estos datos como reconocimiento biométrico.

En este trabajo se recogen nuevos datos y nos centramos en la "Unidad de procesamiento", limpiando y analizando la señal, y en el "Algoritmo de clasificación", tratando de construir un sistema final eficiente que demuestre su posible uso en reconocimiento biométrico, con independencia de la base de datos y el dispositivo.

1.1. Motivación

Mis intereses siempre han sido aprender y explorar cosas nuevas, que no hubiera hecho antes. Participar en un grupo de investigación era una cosa interesante que no quería dejar de lado en mi vida académica y profesional. Eso me llevó a buscar, dentro de la Universidad de Valladolid,

un tema original que me llamará la atención como fue este, con el que empecé a trabajar en el curso 2018/2019 para la realización de una Beca de Investigación y de mis Trabajos de Fin de Grado (TFG). La biometría desde el principio me pareció un tema novedoso, emergente y con muchas posibilidades en las que trabajar. Tantas que, tras la finalización de mis TFGs, decidí seguir formándome dentro de la misma Universidad y continuar tanto con el Grupo de Investigación como con el proyecto.

Entre las asignaturas que más me habían gustado a lo largo de mis cinco años de Grado y con las que había conseguido mi afición por los datos, se encontraban:

- **Análisis de Datos:** consiguiendo ese primer contacto con las técnicas de análisis de datos, tanto para la selección de características como para la aplicación de modelos de clasificación.
- **Análisis Multivariante:** como continuación de la asignatura anterior, más centrada en profundizar los problemas de inferencia multivariante. Empezando un primer contacto con técnicas de aprendizaje automático (Machine Learning) como redes neuronales, máquinas de vectores soporte, árboles de clasificación o Random Forests.
- **Técnicas de Aprendizaje Automático:** donde conseguí un conocimiento más profundo de mayor cantidad de técnicas de Machine Learning, trabajando con diferentes metodologías experimentales y aplicando los conocimiento teóricos en la resolución de prácticas.
- **Minería de Datos:** donde asenté y afiance mis intereses al reforzar los conocimientos de todas las etapas del proceso de minería de datos, desde el preprocesamiento de los datos hasta la evaluación de los resultados, extrayendo conocimiento con diversas técnicas de aprendizaje más profundo, así como el conocimiento de los métodos de ensamblado.

Tras la finalización de las asignaturas del Máster, he conseguido ampliar mis conocimientos en ellas a través de las asignaturas de Ciencia de Datos y abrir nuevos intereses que me parecen muy importantes a lo largo del desarrollo de un proyecto.

- **Arquitectura de Datos:** me ha permitido conocer tecnologías informáticas enfocadas al Big Data. Incluyendo todo lo relacionado con el almacenamiento y el acceso a grandes volúmenes de datos.

En muchas ocasiones trabajaba con datos, pero no me preocupaba por mejorar su almacenamiento y he podido ver cómo es algo muy necesario que puede mejorar la eficiencia en las fases posteriores de ciencia de datos.

- **Business Intelligence:** pudiendo unir el análisis de datos con el mundo empresarial con el objetivo de seguir buenas prácticas que permitan el acceso y el análisis a la información mejorando y optimizando las decisiones y el rendimiento.
- **Seguridad de datos y Ciberseguridad:** cuando trabajamos con datos, muchas veces no nos damos cuenta de las repercusiones éticas y legales que puede tener en el mundo real y que siempre hay que tener en cuenta al empezar a trabajar con nuevos datos y antes de tomar una decisión.

Por ello, he querido tratar todos estos temas a lo largo de la memoria, habiendo podido tener la experiencia de trabajar un mismo proyecto en todas sus fases, desde la adquisición supervisada de los datos hasta la construcción de un modelo de Machine Learning capaz de tomar una decisión.

1.2. Objetivos del proyecto

A continuación, se indica el objetivo general y se enumeran los objetivos específicos en que se divide el presente proyecto.

Objetivo general

El objetivo principal de este trabajo es profundizar y ampliar en el estudio del uso de dispositivos ponibles comerciales para el reconocimiento biométrico de personas mediante la forma de andar para intentar lograr un sistema de reconocimiento totalmente automático en todas sus partes, escalable, seguro y eficiente.

Objetivos específicos

Para poder cumplir el objetivo general, se han creado una serie de objetivos específicos que se desarrollarán de forma progresiva. Estos objetivos servirán para determinar si merece la pena continuar futuros estudios en este tema o, por el contrario, si es mejor abandonar esta línea

de investigación. Los objetivos específicos que se han planteado llevar a cabo han sido los siguientes:

1. Adquisición de nuevos datos, de manera supervisada.
2. Realizar un análisis de los nuevos datos para ver qué características tienen y demostrar su compatibilidad o no con los tomados anteriormente.
3. Probar técnicas que permitan la elección de la zona de interés (eliminación del ruido) de manera automática.
4. Probar y comparar distintos clasificadores.
5. Probar y comparar el rendimiento de características en el dominio del tiempo y el dominio de la frecuencia.
6. Probar y comparar el rendimiento de distintos sensores inerciales (acelerómetro y giróscopo), así como la influencia en el rendimiento de usar distintos tipos de ponibles.
7. Estudiar la influencia de la muñeca portadora del dispositivo y si es indiferente para la toma de decisiones.
8. Teniendo en cuenta los datos disponibles, enfocarlo hacía una adquisición continua (Big Data) y analizar teóricamente arquitecturas escalables siguiendo un escenario tanto experimental como real.
9. Analizar procesos de negocio de interés en estos datos, que pudieran ayudar a la toma de decisiones de manera más rápida, enfocando el problema a Business Intelligence.
10. Analizar teóricamente los riesgos, las consecuencias legales y éticas de los datos, reforzando la seguridad del sistema.

1.3. Estructura de la obra

Siguiendo el actual capítulo introductorio donde se muestra el problema, la motivación por el trabajo y los objetivos, se encuentran los siguientes capítulos.

Capítulo 2. Conceptos teóricos: En este capítulo se hablará sobre una serie de conceptos comunes en biometría, se analizarán las ventajas e

inconvenientes de su uso para el reconocimiento biométrico y de las alternativas históricas existentes.

Capítulo 3. Trabajos relacionados: En este capítulo se analizarán los trabajos previos existentes en el campo de la biometría, centrándose en los aspectos de interés, que se van a aplicar en el presente proyecto.

Capítulo 4. Corpus biométrico: se explicarán los datos con los que se va a trabajar y se realizará un análisis inicial que incluirá la limpieza de los mismos con los problemas que han ido surgiendo, construyendo un sistema automático de limpieza de la señal.

Capítulo 5. Configuración experimental: En este capítulo se explicarán los parámetros estáticos que se van a fijar en el sistema de reconocimiento final. Las decisiones serán sobre el preprocesamiento y los parámetros de interés, qué características se van a extraer de los datos y en qué dominios, cómo se van a medir los resultados, qué procedimiento experimental se va a seguir y qué algoritmos de clasificación se van a utilizar, justificando cada una de las decisiones.

Capítulo 6. Pruebas experimentales: aprovechando las decisiones del capítulo anterior, aquí se van a realizar pruebas experimentales sobre los datos, tomando decisiones para llegar a la construcción de un sistema de reconocimiento biométrico. Se explicarán de manera teórica cada una de las pruebas.

Capítulo 7. Resultados: como continuación del capítulo anterior, se mostrarán los resultados de cada una de las pruebas realizadas.

Capítulo 8. Arquitectura: Este capítulo está orientado a Big Data. Se hablará sobre el almacenamiento, el acceso a grandes volúmenes de datos y las tecnologías informáticas enfocadas a este problema.

Capítulo 9. Business Intelligence: Se realizará el análisis del problema orientado al seguimiento y monitorización de los datos, construyendo cuadros de mandos orientados a mejorar el análisis de la información, optimizando la toma de decisiones y el rendimiento.

Capítulo 10. Ciberseguridad: En este capítulo se hablará de la seguridad de los sistemas biométricos, sus beneficios e inconvenientes orientados a la ciberseguridad, las amenazas y vulnerabilidades existentes y la serie de controles mitigantes y buenas prácticas que se podrían aplicar.

Capítulo 11. Conclusiones y líneas de trabajo futuras: Este es el capítulo final donde se expondrán las conclusiones obtenidas y las posibles alternativas a probar en un futuro.

Para finalizar se encuentra una sección donde se explican los acrónimos y abreviaturas utilizadas a lo largo de la memoria, un índice alfabético, los anexos del trabajo que incluyen el manual de uso de una aplicación interactiva de visualización de los datos de los usuarios y la bibliografía.

Capítulo 2

Conceptos teóricos

En este capítulo, se van a explicar una serie de conceptos comunes en biometría y los sistemas biométricos [13] que aparecerán de manera recurrente a lo largo del trabajo. También se analizan las ventajas e inconvenientes del problema que se quiere resolver y las alternativas históricas que existen.

2.1. Biometría

El concepto biometría proviene de las palabras bio (o del griego *bios*, vida) y metría (o del griego *metron*, medida), que permite inferir que todo equipo biométrico mide e identifica alguna característica propia de la persona [83].

La **biometría** es el estudio estadístico de los fenómenos o procesos biológicos. Tiene muchas aplicaciones posibles, pero dentro de las tecnologías de la información, la más destacada es el estudio del reconocimiento de los seres humanos a partir de sus características, que se suelen clasificar en dos tipos:

- **Características fisiológicas:** Son características físicas de los individuos. Dentro de este grupo cabe destacar la huella dactilar, el iris, etc. Se caracterizan por ser estáticas, es decir, no cambian con el tiempo.
- **Características del comportamiento:** Son propiedades de la forma de actuar de los individuos. Dentro de este grupo se encuentra el modo con el que interactúan con los dispositivos, su voz, firma, forma de andar, etc. Se caracterizan por ser dinámicas, es decir, pueden cambiar con el paso del tiempo.

Un **sistema de reconocimiento biométrico** es una aplicación informática con la capacidad de identificar o verificar a una persona a partir de sus características, bien sean fisiológicas o de comportamiento.

Los sistemas de reconocimiento necesitan algún tipo de patrón para poder identificar o verificar a los individuos. Un **patrón** es un modelo creado mediante capturas o datos del usuario para representarle.

Evidentemente, no todas las características de un individuo pueden ser empleadas para el reconocimiento biométrico. Según [61, 38, 93], para que una característica biométrica pueda ser considerada como tal, ésta ha de cumplir las siguientes propiedades.

- **Universalidad:** todas las personas han de tener dicha característica biométrica.
- **Unicidad:** no ha de haber dos personas que sean idénticas atendiendo únicamente a esa característica.
- **Permanencia:** o biológicamente constante, es decir, la característica no tiene que variar con el tiempo.
- **Recolectable:** la característica ha de poder ser medible cuantitativamente.

Buscando conseguir un sistema de reconocimiento biométrico que tenga las siguientes características.

- **Rendimiento:** precisión que tiene el sistema biométrico empleado a la hora de identificar o verificar a un individuo.
- **Aceptabilidad:** el grado en que el público se muestra positivo a utilizar el sistema biométrico.
- **Invulnerabilidad:** el grado de facilidad del sistema a ser engañado mediante el uso de técnicas fraudulentas.

En la biometría basada en comportamiento como la forma de andar, en ocasiones, la propiedad de *permanencia* no se cumple, denominando a este tipo de biometrías como suaves o débiles *(Soft Biometrics)*.

Por otro lado, la posibilidad de verificar o identificar a un individuo a través de su forma de andar está sujeto a cuestiones éticas, teniendo una serie de ventajas e inconvenientes.

- **VENTAJAS**
  - No requiere interacción durante el proceso de verificación o identificación, el usuario simplemente tiene que andar.
  - Reconocimiento continuo, el propietario se mantiene automáticamente autorizado para el acceso al dispositivo.
  - Proceso discreto, sin molestar al usuario. No requiere cooperación explícita del sujeto.
  - Se puede capturar la información a distancia.
  - Podría utilizarse como ventaja en el campo de la asistencia sanitaria, detectando los cambios de la forma de andar para ayudar a identificar los primeros indicadores de la aparición de la enfermedad de Parkinson y la esclerosis múltiple, así como otras enfermedades.
  - Un impostor puede observar cómo camina un usuario, pero aun así tendrá dificultades para replicar su patrón de marcha, es decir, es difícil de robar o falsificar.
- **INCONVENIENTES**
  - Existen factores externos que influyen en la forma de andar de las personas: condiciones de la superficie, meteorológicas, la ropa o los zapatos que lleve el usuario, etc.
  - Existen factores internos que influyen en la forma de andar de las personas: estado físico, mental, una enfermedad, etc.
- **CUESTIONES ÉTICAS**
  - Los conjuntos de datos contienen información muy sensible, por poderse utilizar para identificar de forma única a las personas y dependiendo del tipo de sensor usado, se podría incluir información que pudiera revelar las condiciones médicas de los usuarios.
  - No requiere el consentimiento del individuo que se está observando, por lo que se podría estar extrayendo su información sin que el usuario lo sepa.

La privacidad de los datos es un problema cada vez más presente en nuestra sociedad. Hay que tener mucho cuidado porque, aunque aparentemente sólo estemos observando su forma de andar, puede existir gente que

de manera maliciosa aproveche esa información y consiga conocer la identidad física, fisiológica o psíquica de los usuarios. Como se ha podido ver en las ventajas e inconvenientes, está información se puede utilizar en el campo de la *medicina* de manera positiva, en la prevención de enfermedades o la posible actuación temprana de las mismas, o de manera negativa, revelando las condiciones médicas de los usuarios y utilizándolo para perjudicarlos.

2.2. Alternativas históricas

Si se entiende el concepto *biometría* en términos muy amplios, se puede decir que se lleva practicando desde el principio de los tiempos y, de hecho, cualquier persona lo practica muchas veces a lo largo del día sin casi darse cuenta: cuando descolgamos el teléfono y escuchamos la voz de nuestro interlocutor, nuestro cerebro trata de comprobar si esa voz se parece a cualquiera de las muestras que tiene almacenadas en su memoria y que ha ido recopilando a lo largo de su vida. Si nuestro cerebro encuentra similitudes suficientes entre alguno de sus recuerdos y lo que está escuchando en ese momento, entonces reconoce a la persona que nos está llamando. Si no, asumimos que estamos ante alguien a quien no conocemos.

El reconocimiento biométrico consiste en aplicar técnicas estadísticas y matemáticas sobre las características fisiológicas o del comportamiento de un individuo para su reconocimiento, ya sea identificación o verificación. Estos sistemas, como se ha dicho en la Introducción y en la sección 2.1, presentan una serie de ventajas, tales como que los usuarios no necesitan recordar claves complejas para su autenticación ni llevar consigo llaves, tarjetas u otros objetos físicos, que pueden perderse o transferirse.

El uso de la forma de andar para la autenticación de los usuarios presenta ventajas sobre otras características, por ser discreto, difícil de robar o falsificar. Sin embargo, su desempeño en el reconocimiento biométrico suele ser menor que el de otras características de comportamiento más utilizadas como la voz o la firma, por ejemplo.

Existen sistemas biométricos tradicionales y portátiles. Los sistemas portátiles, por su naturaleza, están siempre con el usuario pudiendo almacenar los datos dentro del dispositivo, siendo capaces de leer la señal del sujeto en cualquier momento y por tanto, permitiendo la autenticación continua, mientras que los sistemas biométricos tradicionales son generalmente colocados en un lugar fijo, menos susceptibles de deteriorarse

así como más fácilmente reemplazables, haciendo uso de procesos más costosos computacionalmente, ya que pueden utilizar fuentes externas de energía [13]. Ejemplo de sistema biométrico tradicional es una característica de Windows 10 llamada *Windows Hello* [11] que permite al usuario autenticarse usando la cara, el iris o la huella digital.

La creciente popularidad de los dispositivos portátiles está llevando a nuevas formas de interactuar con otros dispositivos inteligentes y con otras personas. Los *wearables* equipados con una serie de sensores son capaces de capturar los rasgos fisiológicos y de comportamiento del propietario, resultando apropiados para biometría, siendo éstos los que se van a utilizar en el presente proyecto. Los sensores predominantes en los dispositivos portátiles actuales son [13]:

- **Sensores de luz:** dependiendo de la resolución del sensor, pueden ser utilizados para medir la intensidad de la luz, como por ejemplo los sensores fotopletismográficos (PPG) [77] que miden el volumen de cambio sanguíneo dentro del tejido microvascular, o proporcionar imágenes completas, como es el caso de los lectores de huellas dactilares [46] o cámaras digitales [37] que pueden capturar las características fisiológicas como la cara u otras características corporales como la forma de andar de los individuos a través de lo que se llaman *técnicas de visión*.

- **Sensores de fuerza:** mide la fuerza que afecta al dispositivo de medición, ya sea originada por el movimiento, ejemplo de ello es el acelerómetro tridimensional [44, 49] o por la fuerza de Coriolis como hace el giroscopio o el campo magnético de la Tierra con el magnetómetro o la presión del aire con el barómetro.

- **Sensores eléctricos:** mide la actividad eléctrica de algunas partes del cuerpo, como por ejemplo, un electrocardiograma para el corazón [23] o cómo cambia una corriente cuando se aplica al cuerpo, como por ejemplo, la conductividad de la piel con un sensor de respuesta galvánica de la piel [41].

- **Sensores de temperatura:** funcionan como una cámara infrarroja. Se captura la energía infrarroja y se transforma en una señal digital que representa la temperatura. Los sensores de temperatura de la piel generalmente se colocan a una distancia muy corta o en contacto directo con la piel. La miniaturización de la tecnología ha permitido el desarrollo de pequeños sensores de temperatura de la piel que

pueden incorporarse en casi cualquier dispositivo electrónico, como los dispositivos ponibles [80].

- **Sensores de sonido:** un micrófono traduce las ondas de sonido que viajan por el aire en una señal eléctrica. Hay micrófonos comerciales que están preparados para capturar la voz humana a una distancia razonable (60dB a 1 metro), ya que la voz de una persona se define por las características fisiológicas del sistema respiratorio de la persona [67].

- **Sensores de localización:** El Sistema de Posicionamiento Global (GPS) consta de 32 satélites y cualquier número de receptores GPS ubicados en la superficie de la Tierra. Un receptor GPS utiliza la señal de cuatro satélites de línea de visión diferentes para triangular la ubicación del dispositivo, ofreciendo sus coordenadas (longitud y latitud), proporcionando información de comportamiento solo con respecto a la ubicación del sujeto.

Hasta la aparición de los smartphones había que utilizar dispositivos tradicionales para captar el movimiento del usuario, como cámaras o sensores del suelo [82]. Primero, con la popularización de los teléfonos inteligentes y, más recientemente, la de los dispositivos portátiles (wearables), ha hecho que la investigación de la forma de andar se centre en el uso de dispositivos portátiles [82, 48], principalmente utilizando sensores de fuerza como el acelerómetro y el giroscopio. Con estos dispositivos, la marcha es más fácil de obtener y se puede adquirir de forma continua. Pero la biometría es un problema difícil en continuo estudio, con cada vez más tipos de sensores diferentes, que algún día podrán ser usados de forma complementaria para conseguir mejores resultados.

Capítulo 3

Trabajos relacionados

A lo largo de este capítulo se va a exponer la situación actual de la biometría, centrándonos en el estudio de aquellas cosas que nos afectan, como la manera de actuar con los datos o las técnicas de aprendizaje automático que se emplean, haciendo un resumen de los resultados obtenidos en investigaciones similares utilizando dispositivos ponibles y el comportamiento de las personas como patrón.

Una vez adquirida una muestra de datos del usuario mediante el sensor existen dos formas de abordar el trabajo: considerando toda la muestra adquirida o dividiendo esa muestra en marcos temporales. Los artículos encontrados trabajan de la segunda forma, ya que justifican que de esta manera se captura la variabilidad del individuo con el tiempo, pero dependiendo del trabajo, se hace de diferente forma. Todos ellos consideran ciclos de marcha y que la forma de caminar humana es un movimiento periódico, compuesto por un paso de la pierna derecha y un paso de la pierna izquierda. Es decir, un ciclo de marcha empieza cuando un pie toca el suelo y termina cuando el mismo pie toca el suelo nuevamente como se muestra en la figura 3.1. Dentro de los trabajos leídos cabe destacar [78] por utilizar solamente la dimensión Z del acelerómetro para hacer la partición del ciclo de la forma de andar, ya que afirma existir una asociación entre la fuerza de reacción del suelo y la fuerza de la señal de este eje, que forma picos de gran magnitud y busca esos cambios del eje Z para dividir la señal en ventanas. Otros como [24] utilizan el periodo de la señal para detectar los ciclos y hacer la división. Por último, en [9] se hace una revisión extensa del enfoque de ventanas, mostrando el tamaño utilizado, en segundos, de distintas publicaciones en las que se realizan diversas actividades, no solo la de caminar, y se sitúan distinto número de acelerómetros en distintas posiciones, que también se indican. Considera la

creación de ventanas, para cada actividad, en función del flujo de datos del sensor y los cambios que se producen, pudiendo identificar dichos cambios a través de un análisis de variaciones en las características de la frecuencia de la señal; o bien detectando el contacto inicial y final del pie con el suelo a través de la aceleración lineal del pie. Introduce la superposición entre ventanas adyacentes, lo que llamaremos "solapamiento" y demuestra que su efecto es beneficioso para el reconocimiento de actividades periódicas como caminar o correr, y estáticas como estar de pie o sentado, pero de utilidad cuestionable para la detección de actividades esporádicas, en las que su naturaleza es más compleja e intercalada. La publicación [87] considera ventanas con 20 % de solapamiento y [10, 95] consideran un 50 %.

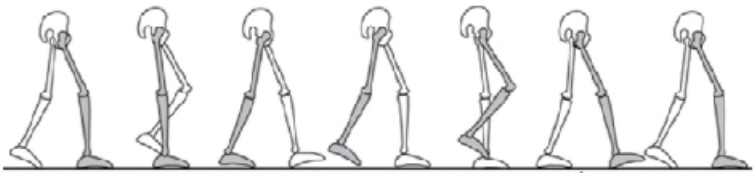


Figura 3.1: Esquema de un *ciclo de marcha*.

Por otro lado, casi nunca se utiliza la señal cruda de los datos, ya que un buen preprocesamiento puede ayudar a mejorar los resultados. Lo que todos los artículos hacen es eliminar el ruido, destacando [40, 92, 71, 24, 30] por hacerlo asignando pesos a los datos a través del filtro Weighted Moving Average (WMA), en [40, 71] también se eliminan los falsos mínimos a través del ciclo medio, calculando aquellos puntos fuera del rango (*media* ± *desviacion_estandar*) o bien con filtros de la mediana como en [94, 12] o filtros *lowpass* o *highpass* para eliminar las interferencias fuera de la banda como hacen [92, 70, 78]. La segunda técnica más aplicada es la de la interpolación por tener los datos disponibles en intervalos de tiempo irregulares, [40, 29, 71] aplican una interpolación de spline, mientras que [92, 24, 78] justifican que utilizar una interpolación lineal es suficiente y más sencillo. Por último, la mayoría de los estudios analizados normalizan los datos, tanto si trabajan en el dominio del tiempo como si lo hacen con las amplitudes de Fourier en el dominio de la frecuencia, destacando [40, 29, 70, 71, 78], pero ninguno de ellos compara el efecto de lo que ocurre si no se normalizan los datos.

Una vez se ha decidido si trabajar con toda la muestra o con una división de ella en ventanas y el preprocesado que aplicar a los datos, hay que decidir si trabajar con la señal preprocesada cruda o si realizar una ex-

tracción de características que represente al usuario. Cuando se trabaja con la señal preprocesada cruda se suele aplicar el método de Dynamic Time Warping (DTW) para obtener la distancia entre las ventanas o las señales, como ocurre en [13]. Pero la mayoría de las investigaciones se centran en la extracción de características, utilizando tanto el dominio del tiempo como el dominio de la frecuencia. En el dominio de la frecuencia, lo más utilizado es la transformada de Fourier, donde destacan las publicaciones [92, 70, 24, 12, 75]. No obstante también se aplica la Transformada Discreta del Coseno (DCT) en [13, 44, 24] donde se utilizan sus coeficientes para intentar representar al usuario. En el dominio del tiempo lo más utilizado son medidas estadísticas como la media y la mediana en [40, 34, 71, 27], la desviación estándar en [33, 34, 78], el mínimo y el máximo en [55, 34, 92, 44, 2]. También se extraen otras características como la curtosis en [44, 27], el coeficiente de asimetría en [75] o los ratios medios de las componentes XY, XZ o YZ como representantes de la gravedad, en [92]. Otro tipo de características incluidas son las correlaciones en [29, 70, 24] o la información sobre los ángulos de los ejes en [55]. Respecto a las características, cabe destacar [55], donde se demuestra cómo la precisión mejora cuando se agregan valores estadísticos adicionales al vector de características y se indica que el vector óptimo es aquel que contiene los menos datos posibles sin perder ningún criterio de información discriminativo. Entre las técnicas para reducir la dimensión del vector de características destacan Principal Component Analysis (PCA), Independent Component Analysis (ICA), Linear Discriminant Analysis (LDA) y Support Vector Machines (SVM), por aplicarse en [12], mientras que PCA también se aplica en [78].

Se ha hecho un resumen de varios artículos en los que se han utilizado dispositivos wearables (tabla 3.1), para los que se va a indicar la siguiente información[1]:

- **Técnica ML:** Técnica de aprendizaje automático utilizada.
- Referencia al **artículo**, junto con el nombre del autor y el año.
- Tipo de **sensor** utilizado entre los explicados anteriormente y que nos interesan, que son los de fuerza.
- **Modo** de reconocimiento biométrico: identificación (I) o verificación (V).

[1]En todos los casos se está utilizando el sensor de un teléfono móvil (smartphone).

- Los resultados, indicando la tasa de equierror **(EER)**, el área bajo la curva ROC (**AUC**), la precisión **(H)** y/o la tasa de falsos positivos **(FPR)**, según qué información esté disponible.
- Número de **sujetos** en la Base de Datos.
- El número de **características** utilizadas.

Los algoritmos de Machine Learning usados en biometría intentan resolver dos problemas diferentes.

- **Identificación biométrica:** modo I, es un problema de clasificación multiclase.
- **Verificación biométrica:** modo V, es un problema de clasificación de una clase. Más frecuente en los sistemas portátiles.

La salida de los algoritmos ejecutados es un valor numérico que mide el grado de similitud entre la señal consultada y un sujeto registrado. Después de obtener este resultado, generalmente se aplica un umbral t para determinar la decisión final. La variación de t ajusta las tasas de falsos positivos y falsos negativos (FPR y FNR, respectivamente), generando lo que se llama las curvas Receiver Operating Characteristic (ROC). Las métricas más usadas en biometría y que se van a utilizar en el estudio de los diversos artículos son además de la tasa FPR, la tasa de equierror (EER) que es el punto de la curva ROC en el que FPR es igual a FNR, y una medida más general de la precisión (H) que corresponde con el número de veces que el sistema produce la decisión correcta.

Se han aplicado muchas técnicas distintas de clasificación, pero las más frecuentes para este tipo de problemas, y cuyos resultados aparecen en la tabla 3.1 son:

- ***K*-Nearest Neighbour (KNN):** o *k*-vecinos más próximos, es un método de aprendizaje perezoso por almacenar vectores de características en el conjunto de entrenamiento y retrasar todo el procesamiento hasta la clasificación. Muy utilizado y popular por su simplicidad y efectividad.
- **Support Vector Machines (SVM):** o máquinas de vectores soporte, es un método estadístico que construye un hiperplano que separa

de manera óptima las diferentes clases de las muestras de entrenamiento. La efectividad de SVM depende del kernel seleccionado y de un parámetro de margen que describe la influencia de una sola muestra en el hiperplano. En [17, 33] los investigadores utilizaron el acelerómetro para verificar la identidad de los sujetos mientras realizaban gestos, pudiendo verificar la identidad del sujeto solo mientras caminaban en un ambiente muy restringido. En cambio, en [21] propusieron un sistema multimodal que consistía en acelerómetro, giroscopio y posicionamiento GPS para verificar la identidad de un sujeto, consiguiendo resultados prometedores, pero utilizando únicamente 3 personas. Por otro lado, destaca [35] por obtener el 100 % de precisión utilizando los datos del acelerómetro, pero su estrategia de clasificación no fue exactamente un sistema biométrico, sino una mezcla de identificación y verificación, con una población también muy limitada de 5 individuos.

- **Gaussian Mixture Model (GMM):** o modelo de Mixtura Gaussiana, es un modelo probabilístico que asume que todas las muestras del mismo sujeto pueden generarse por una suma ponderada de un número finito de distribuciones gaussianas. Los pesos de cada distribución y sus parámetros se obtienen a través de diferentes métodos de ajuste, por ejemplo, el más común en la literatura es el de maximización de las expectativas (EM). En un sistema de verificación, se debe establecer un umbral de probabilidad para seleccionar muestras como válidas para ese GMM. En un sistema de identificación biométrica, la muestra de consulta se pasa a través de todos los GMM de los sujetos, y se selecciona el que tiene más probabilidad. Pero es un método más utilizado para los sonidos emitidos por el cuerpo, como por ejemplo, del corazón, donde los experimentos [65, 96] han logrado precisiones entre 0.86 y 1 con poblaciones de sujetos de tamaño medio (entre 10 y 80 individuos). También se ha empleado en verificación utilizando el acelerómetro y la respuesta galvánica de la piel [49, 44] con peores resultados de EER y FPR por encima de 0.14 en todos los casos.

- **Hidden Markov Model (HMM):** o modelo oculto de Markov es un tipo particular de red Bayesiana, donde el sistema realiza la transición de un estado a otro según las observaciones y un conjunto de probabilidades de transición que se desconocen previamente. Los HMM se han utilizado ampliamente en varios problemas de aprendizaje automático, pero son especialmente conocidos por sus

aplicaciones en reconocimiento de voz [32], donde se ha logrado EERs promedio de 0.10 con 48 individuos diferentes.

- **Decision Trees (DTree):** o árboles de decisión, donde cada nodo evalúa una característica y las hojas del árbol especifican la decisión a tomar. El algoritmo más utilizado y conocido se llama C4.5 [64], en su funcionamiento va (i) calculando la característica que proporciona la mayor ganancia de información en las muestras, (ii) crea un nodo de decisión utilizando el atributo que mejor divide el conjunto de datos de entrenamiento, (iii) crea listas secundarias de muestras utilizando los criterios de decisión creados y (iv) crea un Decision Trees (DTree) para todas las listas secundarias a partir del nodo de decisión. El algoritmo se detiene cuando todas las muestras en una lista secundaria pertenecen a una clase específica, que es cuando el algoritmo crea un nodo de decisión para esa clase. En [74] se encontró que los árboles de decisión podrían identificar a los sujetos con alta precisión usando datos del acelerómetro, pero utilizó una población muy pequeña de sólo 5 individuos.

Técnica ML	**Referencia**	**Sensor**	**Modo**	**EER**	**H**	**FPR**	**Sujetos**	**Caract.**
SVM	[Casale et al. 2012] [17]	ACC	V	-	-	0.01	20	18
	[Hestbek et al. 2012] [33]	ACC	I	0.1	-	-	36	12
	[Ho et al. 2012] [35]	ACC	V	-	1	-	36	-
	[Sugimori et al. 2011] [74]	ACC	I	-	0.98	-	5	2
GMM	[Lu et al. 2014] [44]	ACC	V	0.14	-	-	12	87
	[Meharia and Agrawal 2015] [49]	ACC	V	-	0.8	0.14	10	-
KNN	[Nickel et al. 2012] [54]	ACC	I	-	0.82	-	36	52
HMM	[Nickel et al. 2011] [53]	ACC	I	0.1	-	-	48	26
DTree	[Sugimori et al. 2011] [74]	ACC	I	-	0.98	-	5	2

Tabla 3.1: Resultados de artículos de la bibliografía que estudian el problema utilizando un smartphone.

En la tabla 3.1, las máquinas de vectores soporte consiguen buenos resultados, tanto en el caso de identificación como de verificación, pero utilizando Bases de Datos muy pequeñas, con un máximo de 36 individuos. Destaca [74], ya que consigue una precisión alta de 0.98; y cuando se vuelve a utilizar con árboles de decisión consigue los mismos resultados, pero sigue ocurriendo lo mismo, con 5 individuos y 2 características no se

pueden extraer muchas conclusiones. El modelo de Mixtura Gaussiana (GMM) parece obtener peores resultados que SVM, además llama la atención [44], ya que con 12 individuos se están usando 87 características, pudiendo existir un problema de sobreajuste que este perjudicando a los resultados. *K*-vecinos más próximos consigue una precisión de 0.82, peores resultados que SVM, pero de nuevo se están utilizando muchas más características, por lo que podría existir también aquí un problema de sobreajuste. Por último, en el modelo oculto de Markov (HMM) se obtiene una tasa de equierror razonable, de 0.1, y es la Base de Datos que contiene más sujetos, de un total de 48, con un número de características más pequeño que antes, de 26. Sin duda y como se puede ver en la tabla, el sensor más utilizado en la literatura es el acelerómetro, aunque hablan sobre el giroscopio, no lo utilizan tanto para obtener resultados. Pero existen artículos como [18], que comparan ambos sensores utilizando señales PPG durante la realización de ejercicio físico. Estas señales recogen el estado del corazón y otros órganos. Considerando que los acelerómetros por sí solos no pueden diferenciar entre aceleración debida al movimiento o a la gravedad y que las correlaciones presentes en las diferentes señales de movimiento (3-ejes del acelerómetro y los 3-ejes del giroscopio) recogen diferente información.

El trabajo [7] se centra en explorar la implementación de algoritmos de Machine Learning para el reconocimiento de la forma de andar. Utiliza una base de datos open-source, HugaDB, que recopila los datos de 18 individuos realizando una serie de actividades: caminar, correr, sentarse y estar de pie, aplicando los algoritmos de manera independiente para cada actividad. Utiliza una red de sensores corporales que consta de seis sensores inerciales portátiles (acelerómetro y giroscopio) ubicados a la derecha e izquierda de los muslos, las espinillas y los pies junto con dos sensores Electromyography (EMG) en los cuádriceps para medir la actividad muscular. Se prueban los siguientes algoritmos de aprendizaje automático.

- **RIPPER**: es uno de los algoritmos de aprendizaje basados en reglas más populares. Las clases se examinan en tamaño creciente y se establece un conjunto inicial de reglas para la clase utilizando el error acumulado reducido. El algoritmo procede tratando todas las muestras de los datos de entrenamiento como una clase y encuentra un conjunto de reglas que cubren a todos los miembros de esa clase. En consecuencia, pasa a la siguiente clase y hace lo mismo, repitiéndolo hasta que se hayan cubierto todas las clases.

- **Perceptrón multicapa (MLP)**: se organiza en capas. Las capas están formadas por varios nodos interconectados (neuronas) que contienen una función de activación. Los patrones se presentan a la Artificial Neural Networks (ANN) a través de la capa de entrada, que se comunica con una o más capas ocultas donde el procesamiento real se realiza a través de un sistema de conexiones ponderadas. Las capas ocultas luego se vinculan a una capa de salida. Un MLP consta de una capa de entrada, una o más capas de unidad de umbral lineal (LTU) denominadas capa oculta y una capa de salida. Otras capas, excepto la capa de salida, contienen la neurona de polarización y están completamente conectadas a otras capas.

- **Árboles de decisión (DTree)**: ofrecen una estrategia de arriba a abajo. Un árbol de decisión es una estructura que se utiliza para dividir un conjunto de datos que contiene una gran cantidad de registros en conjuntos más pequeños aplicando una serie de reglas de decisión. En otras palabras, es una estructura que se utiliza para dividir grandes cantidades de registros en registros muy pequeños aplicando pasos simples de toma de decisiones.

- **Random Forest (RF)**: su objetivo es aumentar el valor de la clasificación y ser más preciso mediante el uso de más de un árbol de decisión durante el proceso de clasificación.

- **Bootstrap aggregating (bagging)**: pertenece al grupo de metaalgoritmos, diseñados para mejorar la estabilidad y la precisión de los algoritmos de aprendizaje automático. También reduce la variación y ayuda a prevenir la inserción excesiva. Se aplica con el método de árboles de decisión, pero se puede utilizar con otros algoritmos.

- **Naive Bayes (NB)**: tiene como objetivo determinar la clase de datos presentados al sistema mediante una serie de cálculos definidos según los principios de probabilidad. La clasificación de Naïve Bayes proporciona datos que se enseñan al sistema a un cierto ritmo. Los datos presentados para el entrenamiento (docencia) deben tener una clase. Con las operaciones probabilísticas realizadas sobre los datos entrenados, los nuevos datos de prueba presentados al sistema se operan de acuerdo con los valores de probabilidad obtenidos previamente y se intenta determinar qué categoría de datos de prueba se da. Cuanto mayor sea el número de datos entrenados, más preciso será determinar la categoría real en los datos de prueba.

La tabla 3.2 muestra los resultados para la actividad de caminar, donde el peor funcionamiento corresponde con el perceptrón multicapa, un algoritmo potente pero que puede llegar a funcionar mal si se utilizan malos criterios a la hora de decidir sus parámetros (parámetros de inicialización, número de neuronas, capas ocultas...) o por tener clases poco balanceadas. Random Forest es el que mejor funciona, superando los valores de AUC de los árboles de decisión y bagging, que por sí solos ya consiguen buenos resultados, consiguiendo, por tanto, RF su objetivo, aumentar la precisión tras multiplicar el número de árboles de decisión. El algoritmo basado en reglas (RIPPER), cuyo funcionamiento es similar al de los árboles de decisión, también funciona bien. Al igual que el algoritmo basado en probabilidades (NB). En definitiva, a excepción de MLP, todos los algoritmos consiguen buenos resultados, lo cual puede estar relacionado con la adquisición de los datos. Al utilizar múltiples sensores en diferentes partes del cuerpo (muslos, espinillas, pies y cuádriceps) y en ambos lados (derecha e izquierda) la monitorización del usuario es alta.

Técnica ML	**Referencia**	**Modo**	**AUC**	**Sujetos**	**Caract.**
RIPPER	[Aybuke Kececi, 2020] [7]	I	0.9910	18	36
MLP	[Aybuke Kececi, 2020] [7]	I	0.8298	18	36
DTree	[Aybuke Kececi, 2020] [7]	I	0.9705	18	36
RF	[Aybuke Kececi, 2020] [7]	I	1	18	36
Bagging	[Aybuke Kececi, 2020] [7]	I	0.9993	18	36
NB	[Aybuke Kececi, 2020] [7]	I	0.9994	18	36

Tabla 3.2: Resultados de artículos de la bibliografía que estudian el problema y tratan de mejorar los resultados variando la técnica de Machine Learning con la base de datos HugaDB.

Un problema del uso de *smartphones* se encuentra en la ubicación del dispositivo, ya que no todos los usuarios lo llevan siempre en el mimo sitio y posición. Destaca el trabajo [30], donde se emplearon dispositivos ponibles diseñados específicamente para capturar los datos del movimiento mediante un acelerómetro, probando diferentes posiciones del dispositivo: en el pie, la cadera, el bolsillo del pantalón y la muñeca de los distintos usuarios, y se consiguieron los resultados de la tabla 3.3. Como en cada localización se utiliza distinto número de usuarios, se va a realizar una especie de EER por usuario, utilizando el cociente del EER y el número de individuos para poder compararlos. Parece que los mejores

resultados se consiguen con el dispositivo en la cadera y en el bolsillo del pantalón, zonas muy similares, donde el acelerómetro podría capturar mejor el movimiento, mientras que en el tobillo los resultados son peores, probablemente porque sea una zona con más ruido, al ir demasiado en contacto con el pie y la muñeca parece capturar peor la información consiguiendo peores resultados. Pero hace falta recoger el EER medio utilizando las 4 localizaciones y el mismo número de individuos para que las conclusiones sean más realistas, ya que los 30 individuos usados para obtener el EER con el dispositivo en la muñeca pueden ser los que peores resultados estén dando por tener movimientos muy similares entre ellos.

Localización dispositivo	EER	Nº individuos	EER por individuo
Tobillo	5 %	21	0.24
Cadera	13 %	100	0.13
Bolsillo del pantalón	7.3 %	50	0.15
Muñeca	10 %	30	0.33

Tabla 3.3: Resultados de artículos de la bibliografía que estudian el problema y alternan la posición del dispositivo.

No obstante, los sistemas biométricos rara vez alcanzan una precisión perfecta en la práctica debido a muchos factores, tales como el ruido, entrenamiento incompleto o un algoritmo de aprendizaje automática no ideal, lo cual afecta a la tasa obtenida de falsos positivos y falsos negativos. Todos estos resultados y la investigación realizada ha sido utilizando, la mayoría de las veces, smartphones o dispositivos creados para el propio propósito del estudio cuando en el presente proyecto se utilizan *wearables* comerciales, donde un algoritmo muy preciso podría drenar la batería rápidamente o tomar demasiado tiempo para tomar una decisión, lo cual no es factible.

Se pueden encontrar varios trabajos de investigación centrados en el mismo enfoque que el que aquí se propone en los últimos años, pero hasta donde llega nuestro conocimiento, solo dos utilizan un dispositivo comercial (un reloj inteligente en ambos casos) [39, 91]. La razón principal es que no es fácil acceder a la señal sin procesar generada por los sensores portátiles, ya que la mayoría de los wearables comerciales no lo permiten. Aunque el uso de wearables simulados es interesante, para aplicaciones prácticas, el uso de reales es fundamental. Tienen características diferenciales e importantes, entre las que cabe destacar.

- Primero y más importante, es necesario probar la viabilidad de acceder a las señales sin procesar de los sensores.
- No podemos controlar la adquisición. Por ejemplo, no podemos controlar la frecuencia de muestreo.
- No son dispositivos dedicados. Esto significa, primero, que el sistema operativo puede tomar el control en cualquier momento de cualquiera de sus tareas prioritarias, lo que interfiere con la adquisición; y, en segundo lugar, que los sensores no se crean para la autenticación de usuarios, entonces es necesario investigar si sus señales son adecuadas para esta tarea.

Se muestran las diferencias de los dos proyectos más similares para que quede clara la contribución diferenciadora de este proyecto. Respecto a [39]:

- Solo adquiere una muestra de datos por usuario, usando la mitad para entrenar al clasificador y la otra para probar. Esto no es un caso realista. En este proyecto se han adquirido datos en dos días diferentes (dos sesiones) y en cada sesión cada dispositivo fue capturado tres veces, no consecutivamente, es decir, se adquiere, primero, con un dispositivo, luego con el otro, repitiendo este proceso tres veces por sesión. Nunca se entrena y prueba con la misma muestra.
- Utiliza un único dispositivo portátil (LG smartwatch). Aquí se utilizan dos: Microsoft Band 2 y Motorola Moto 360, construyendo, por tanto, un trabajo experimental más completo con comparaciones y pruebas cruzadas.

Por otro lado, [91] utiliza un escenario de prueba más realista, similar al de este proyecto. Las condiciones de adquisición del corpus biométrico son muy similares, incluido el número de individuos. Sin embargo, también existen diferencias importantes:

- Se centra en la detección de actividad para mejorar el reconocimiento del usuario. Se estudian varios parámetros del sistema propuesto para realizar esa detección, pero con respecto al sistema de reconocimiento de la forma de andar, solo se analiza el número de ciclos de la marcha. Aquí nos centramos en el sistema de reconocimiento

de la marcha, utilizando para los experimentos este parámetro y muchos más. El análisis de los mismos lo llevamos a cabo en el trabajo anterior [57], del cual el proyecto actual es continuación.

- Utiliza ataques de impostores, pero solo para probar el sistema final. Aquí, se consideran en todos los experimentos y pruebas.
- Solo se utilizan un dispositivo portátil (Samsung Smartwatch) y un sensor (acelerómetro). Aquí se utilizan dos dispositivos: Microsoft Band 2 y Motorola Moto 360 y 2 sensores (acelerómetro y giroscopio) construyendo, por tanto, un trabajo experimental más completo con comparaciones y pruebas cruzadas.

Considerando, por ende, el presente proyecto como complementario al mostrado en [91], ya que el abordaje del problema es similar, pero con perspectivas diferentes y complementarias.

Capítulo 4

Corpus biométrico

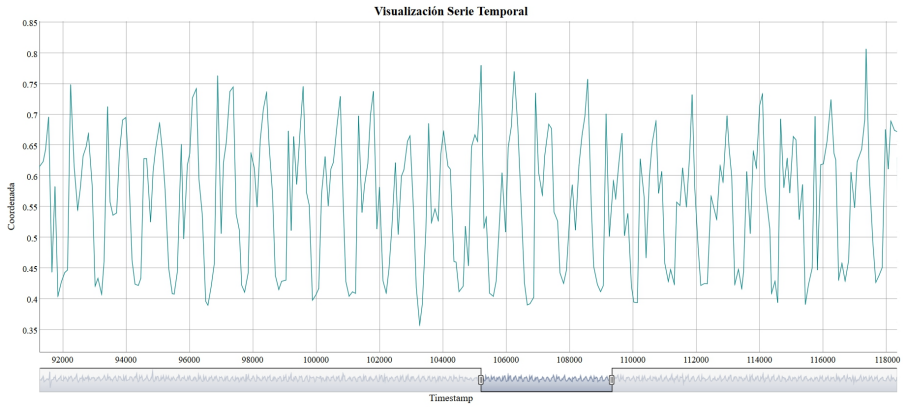
Los apartados anteriores han permitido conocer el tipo de datos con los que se va a trabajar, su naturaleza y el problema que se quiere resolver. En este apartado se va a invertir tiempo en limpiar los datos y garantizar una cierta calidad de los mismos, que es también una fase inicial y muy importante, que puede llevar a finalizar con éxito o fracaso cualquier proyecto.

Dos de los muchos objetivos que subyacen bajo la palabra *Análisis de Datos* son los de encontrar relaciones insospechadas y resumir los datos de maneras novedosas que los hagan comprensibles y útiles. Esto, aunque se puede hacer de muchas maneras, en el presente capítulo se va a realizar a través de un conocimiento previo grande, un análisis visual de los datos y la prueba de técnicas como la autocorrelación, el cálculo de la energía, el número de cruces por el valor medio de la señal o el tiempo de adquisición entre dos muestras consecutivas.

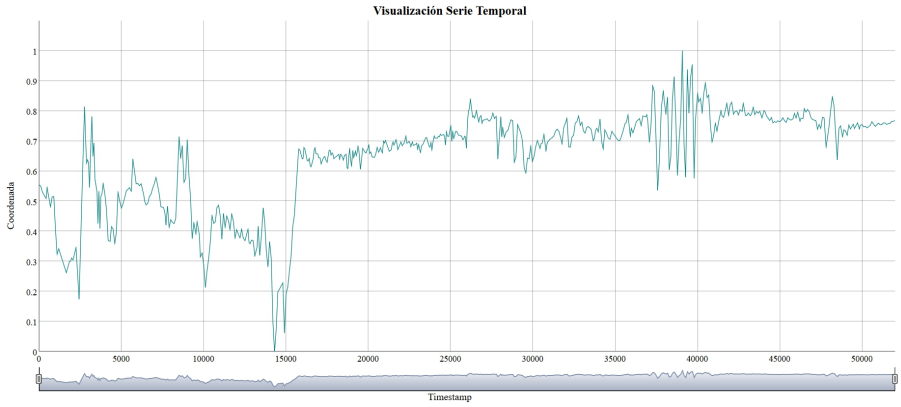
Es importante que exista periodicidad en la señal, es decir, el mismo patrón repetido en todas las muestras correspondientes al mismo usuario, mientras que aquellas que corresponden a otro usuario distinto, tuvieran otro patrón periódico pero diferente. Un ejemplo de serie periódica puede verse en la figura 4.1 (a). El caso opuesto sería una serie no periódica, cuyo ejemplo se puede ver en la figura 4.1 (b).

4.1. Base de datos

En trabajos anteriores [4, 57], pertenecientes a la Universidad de Valladolid y al grupo de investigación del Departamento de Informática, se ha abordado este problema utilizando una única Base de Datos recogida



(a) Periódica



(b) No Periódica

Figura 4.1: Tipos de series de datos.

durante la realización del primer trabajo [4]. En ella, se disponía de 20 usuarios: 13 hombres y 7 mujeres en edades comprendidas entre 16 y 57 años, utilizando el reloj y la pulsera situada en la muñeca de uso habitual del reloj, a la que se llamará mano portadora del dispositivo[1], o la opuesta en función del usuario. Cada uno realizó un recorrido andando de minuto y medio, aproximadamente, durante 2 sesiones en diferentes días. Y dependiendo del usuario, cada día realizó el recorrido una única vez o dos.

[1]Con mano dominante se hace referencia a la mano en la que el usuario tiene más habilidad (en personas diestras, la derecha, normalmente) y con mano portadora se hace referencia a la mano en la que se pondría habitualmente el ponible, normalmente la no dominante.

Característica	BD Inicial	BD Nueva
Duración del recorrido	1min 30seg	5min
Número de muestras	Depende del usuario	6 por sesión
Número de sesiones	Depende del usuario	2 por usuario
Mano de uso del dispositivo	Aleatorio	Supervisado

Tabla 4.1: Diferencias entre las Bases de Datos adquiridas en [4] (BD Inicial) y en el presente trabajo (BD Nueva).

Tras haber analizado y trabajado con esta primera Base de Datos y llegar a la conclusión de que se trata de señales periódicas, susceptibles de poder ser usadas en biometría, se ha decidido capturar nuevos datos con el conocimiento adquirido de los anteriores. Se trata de datos supervisados con las siguientes características:

- Recorrido andando de una mayor duración de tiempo, 5 minutos, aproximadamente. El mismo recorrido para todos los usuarios.
- Controlar la mano de uso del dispositivo y la orientación de los mismos.
- Capturar datos en 2 sesiones en días diferentes. Para poder estudiar la dependencia a lo largo del tiempo se adquirieron dos sesiones distintas, con una separación mínima entre ambas de dos semanas.
- Cada día realizar 6 recorridos: 3 con cada dispositivo, los dos primeros utilizando la muñeca de la mano de uso habitual del reloj o portadora[1] y el último utilizando la opuesta, para comprobar si la mano en que se lleva el dispositivo modifica los resultados.

En la tabla 4.1 se recogen las diferencias entre las dos bases de datos, en la 4.2 se muestra la información disponible de cada uno de los usuarios de la primera base de datos (BD Inicial), mientras que en la tabla 4.3 se muestra la información de los nuevos usuarios (BD Nueva).

La nueva Base de Datos va a disponer de la misma cantidad de datos en 2 dispositivos comerciales y con 2 sensores en cada uno de ellos. Los dispositivos son un reloj *Motorola Moto 360* (Moto) y una pulsera *Microsoft Band 2* (Micro). Los mismos utilizados en los trabajos anteriores a los que se hace referencia [4, 57]. Los dispositivos se pueden ver en la figura 4.2.

BD	Usuario	Sexo	Edad	Mano Dominante	Mano Portadora	Nº de datos
BD Inicial	usuario1	Hombre	21	Derecha	Izquierda	4
BD Inicial	usuario2	Hombre	57	Derecha	Izquierda	4
BD Inicial	usuario3	Hombre	50	Derecha	Izquierda	4
BD Inicial	usuario4	Hombre	50	Derecha	Izquierda	2
BD Inicial	usuario5	Mujer	53	Derecha	Izquierda	2
BD Inicial	usuario6	Hombre	21	Derecha	Derecha	2
BD Inicial	usuario7	Mujer	16	Derecha	Izquierda	2
BD Inicial	usuario8	Mujer	56	Derecha	Derecha	4
BD Inicial	usuario9	Mujer	46	Derecha	Izquierda	4
BD Inicial	usuario10	Mujer	19	Derecha	Izquierda	4
BD Inicial	usuario11	Mujer	46	Derecha	Derecha	4
BD Inicial	usuario12	Hombre	16	Derecha	Derecha	4
BD Inicial	usuario13	Hombre	49	Derecha	Derecha	4
BD Inicial	usuario14	Hombre	20	Derecha	Izquierda	4
BD Inicial	usuario15	Hombre	22	Derecha	Derecha	4
BD Inicial	usuario16	Mujer	48	Derecha	Izquierda	2
BD Inicial	usuario17	Hombre	53	Derecha	Derecha	2
BD Inicial	usuario18	Hombre	22	Derecha	Izquierda	4
BD Inicial	usuario19	Hombre	23	Derecha	Derecha	2
BD Inicial	usuario20	Hombre	21	Derecha	Izquierda	2
TOTAL MUESTRAS DE DATOS DISPONIBLES						66

Tabla 4.2: Metadatos de los usuarios en la Base de Datos Inicial. Una persona con mano dominante derecha es lo que se conoce como diestro y una persona con mano dominante izquierda sería zurdo.

(a) Microsoft Band 2 (b) Motorola Moto 360

Figura 4.2: Dispositivos disponibles.

BD	Usuario	Sexo	Edad	Mano Dominante	Mano Portadora	Nº de datos
BD Nueva	usuario1	Hombre	27-45	Derecha	Izquierda	6
BD Nueva	usuario2	Hombre	19-26	Derecha	Izquierda	6
BD Nueva	usuario3	Hombre	27-45	Derecha	Izquierda	6
BD Nueva	usuario4	Hombre	19-26	Derecha	Izquierda	6
BD Nueva	usuario5	Hombre	19-26	Izquierda	Derecha	6
BD Nueva	usuario6	Hombre	27-45	Derecha	Izquierda	6
BD Nueva	usuario7	Hombre	19-26	Derecha	Izquierda	6
BD Nueva	usuario8	Mujer	46-65	Derecha	Izquierda	6
BD Nueva	usuario9	Mujer	19-26	Derecha	Izquierda	6
BD Nueva	usuario10	Mujer	46-65	Derecha	Izquierda	6
BD Nueva	usuario11	Hombre	27-45	Derecha	Izquierda	6
BD Nueva	usuario12	Hombre	46-65	Derecha	Izquierda	6
BD Nueva	usuario13	Mujer	46-65	Derecha	Izquierda	6
BD Nueva	usuario14	Hombre	46-65	Derecha	Izquierda	6
BD Nueva	usuario15	Hombre	46-65	Izquierda	Izquierda	6
BD Nueva	usuario16	Hombre	27-45	Derecha	Izquierda	6
BD Nueva	usuario17	Mujer	46-65	Derecha	Izquierda	6
BD Nueva	usuario18	Hombre	46-65	Derecha	Izquierda	6
BD Nueva	usuario19	Hombre	27-45	Derecha	Izquierda	6
BD Nueva	usuario20	Hombre	27-45	Derecha	Izquierda	6
BD Nueva	usuario21	Hombre	27-45	Derecha	Izquierda	6
BD Nueva	usuario22	Mujer	19-26	Derecha	Izquierda	6
BD Nueva	usuario23	Mujer	19-26	Derecha	Izquierda	6
BD Nueva	usuario24	Hombre	19-26	Derecha	Izquierda	6
TOTAL MUESTRAS DE DATOS DISPONIBLES						144

Tabla 4.3: Metadatos de los usuarios en la Base de Datos Nueva. Una persona con mano dominante derecha es lo que se conoce como diestro y una persona con mano dominante izquierda sería zurdo.

Los sensores utilizados son el acelerómetro (ACC) y el giroscopio (GYR) tridimensional que poseen los dispositivos usados en la captura.

- **Acelerómetro:** mide la orientación de una plataforma fija respecto a la superficie terrestre. En esta situación podría verse como la rapidez con que algo se acelera.
- **Giroscopio:** mide la velocidad de rotación sobre un eje determinado.

Las 3 componentes son X, movimiento hacia la izquierda o derecha; Y, movimiento hacia adelante o hacia atrás; Z, movimiento hacia arriba o hacia abajo.

De manera resumida, al realizar cada recorrido, se va guardando en la Base de Datos la siguiente información.

- Identificador del usuario.
- International Mobile Equipment Identity (IMEI) del teléfono móvil o la herramienta utilizado para la adquisición de los datos. El IMEI es un código que identifica al aparato de forma exclusiva a nivel mundial.
- Dispositivo que se está utilizando (Micro o Moto).
- Tipo de sensor al que pertenece el dato (ACC o GYR).
- Timestamp: contiene tanto la fecha, como la hora con una precisión en milisegundos.
- Las coordenadas X, Y y Z del sensor indicado.
- Nombre del usuario.
- Número de la tarea, la sesión y la muestra para distinguir entre las diferentes tomas de datos del mismo usuario.

Con ello, se construye un fichero en formato CSV para cada toma de datos de cada usuario. El fichero contiene únicamente la información necesaria, que se va a utilizar a lo largo de este trabajo.

- Una primera columna con el tiempo relativo, que es la diferencia de tiempo entre una captura de las coordenadas X, Y, Z y la anterior. Los datos se almacenan con este valor temporal porque es más compacto que almacenar el timestamp.

- Tres columnas para las coordenadas X, Y, Z correspondientes a la captura de datos que marque el tiempo relativo. En la Base de datos tienen el nombre de dato1, dato2 y dato3 para hacer referencia a las coordenadas X, Y, Z respectivamente.

El recorrido dura, aproximadamente, 5 minutos, por lo que se tienen bastantes capturas para cada toma de datos de cada usuario. En la figura 4.3 se muestra un ejemplo de toma de datos, con el formato final con el que se va a trabajar.

	A	B	C	D	E
1	tiempoRelativo	dato1	dato2	dato3	
2	0	0,800781	-0,551514	0,602539	
3	68	1,062012	-0,692627	0,31665	
4	111	1,089844	-0,79248	0,296143	
5	102	0,954346	-0,553955	0,223877	
6	105	0,973145	-0,55835	0,19873	
7	94	1,097656	-0,663818	0,200439	
8	83	0,681885	-0,542725	0,268799	
9	148	1,078857	-0,369629	0,190674	
10	63	0,808838	-0,330078	0,204102	
11	71	0,764648	-0,465332	0,207764	
12	118	0,772461	-0,49585	0,30127	
13	99	0,766357	-0,543213	0,398438	
14	79	0,880615	-0,638672	0,471191	
15	136	0,775879	-0,593506	0,387939	
16	51	0,700195	-0,594727	0,356201	
17	101	0,719727	-0,477051	0,350586	
18	121	0,575928	-0,40625	0,341309	
19	79	0,584229	-0,411133	0,350098	

Figura 4.3: Formato de los datos que se van a utilizar.

La notación relativa al corpus y a los datos adquiridos que se va a utilizar va a ser la siguiente:

- **Dispositivo:** para referirse a cada dispositivo portátil (wearable) utilizado.
- **Muestra:** los datos biométricos adquiridos en el minuto y medio, en la primera base de datos, o en los cinco minutos, en la segunda base de datos, andando con cada dispositivo. Luego "una sesión tiene dos/tres muestras adquiridas con cada dispositivo", identificadas como muestra 1 (M1), muestra 2 (M2) y muestra 3 (M3).

- **Usuario:** cada uno de los individuos del corpus.

4.2. Análisis visual de los datos crudos

Como punto de partida, se tienen los datos crudos. Tras un análisis visual manual a través de la aplicación web interactiva construida en R que se explica en el apéndice A, se ha visto que a pesar de haber sido adquiridos de manera supervisada, existen varios problemas a señalar.

- **Perdidas de conexión al inicio y final de la muestra.** Se puede ver un ejemplo en las figuras 4.4 y 4.5. El comportamiento más frecuente es la perdida final, causado por el tiempo transcurrido entre que se pulsa el botón "parar" en la aplicación de adquisición de los datos y se guarda la muestra. La pérdida inicial es causada por sobrecarga de la aplicación que obliga a pararla y volverla a iniciar, comenzando de nuevo la muestra. Solo hay una muestra en la que la perdida de conexión afecta al tiempo total de los datos adquiridos (figura 4.6). En ese caso, se descarta al usuario[2]. En el resto, no se da importancia, considerándolo como ruido.

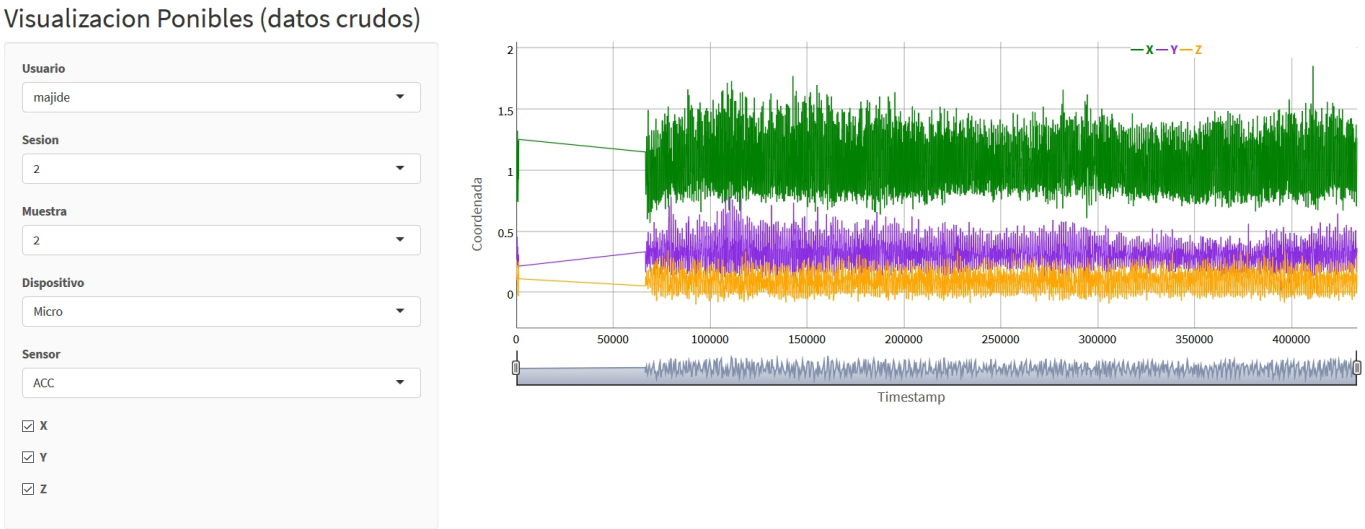


Figura 4.4: Estado de los datos crudos. Perdida de conexión al inicio de la muestra

[2]Se descartan usuarios tras un análisis visual pero posteriormente se buscará un sistema automático que descarte estos usuarios.

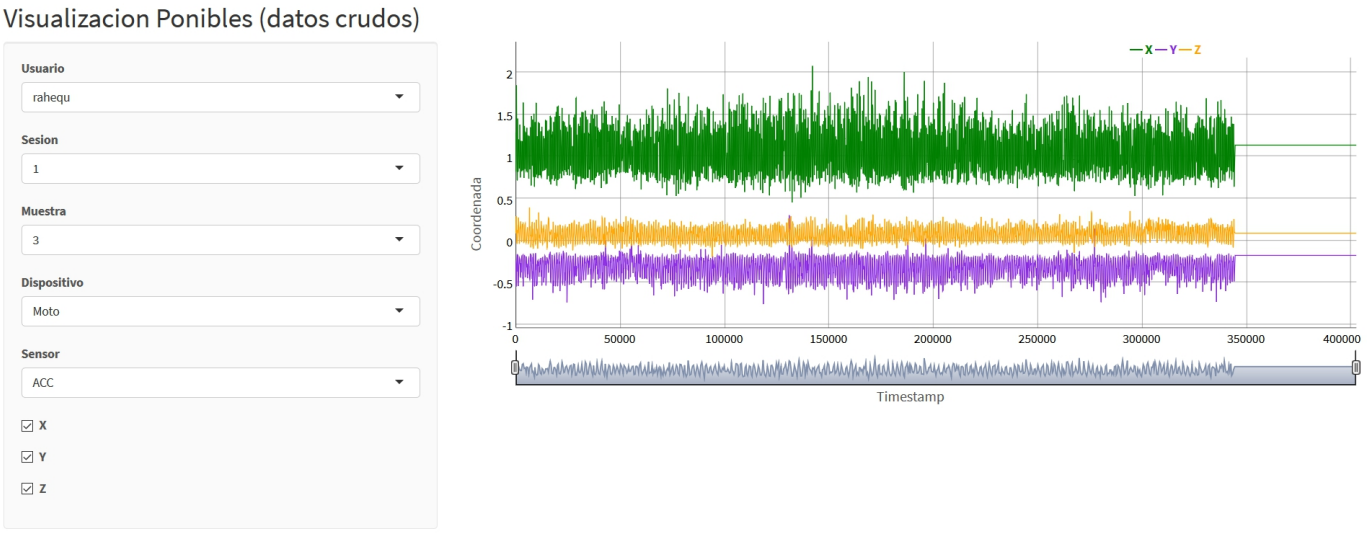


Figura 4.5: Estado de los datos crudos. Perdida de conexión al final de la muestra

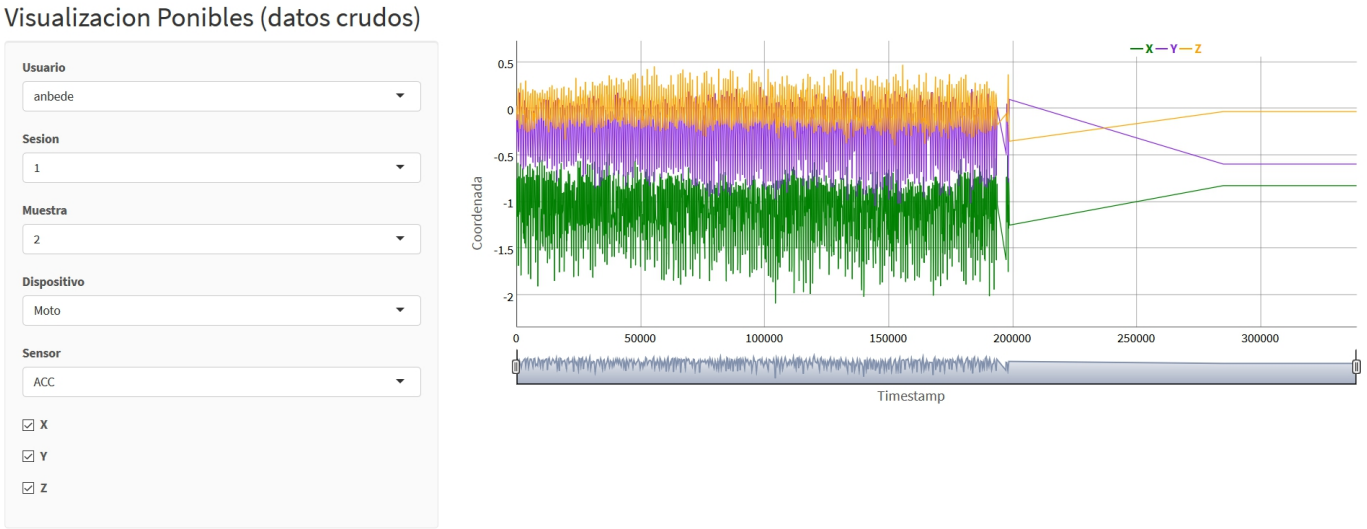


Figura 4.6: Estado de los datos crudos. Perdida de conexión grave al final de la muestra

- **Perdidas de conexión en un punto intermedio de la muestra.** Es un comportamiento poco frecuente. Se desconoce la causa que lo provoca. En dos muestras se considera lo suficientemente grave como para descartar al usuario (figuras 4.7 y 4.8).

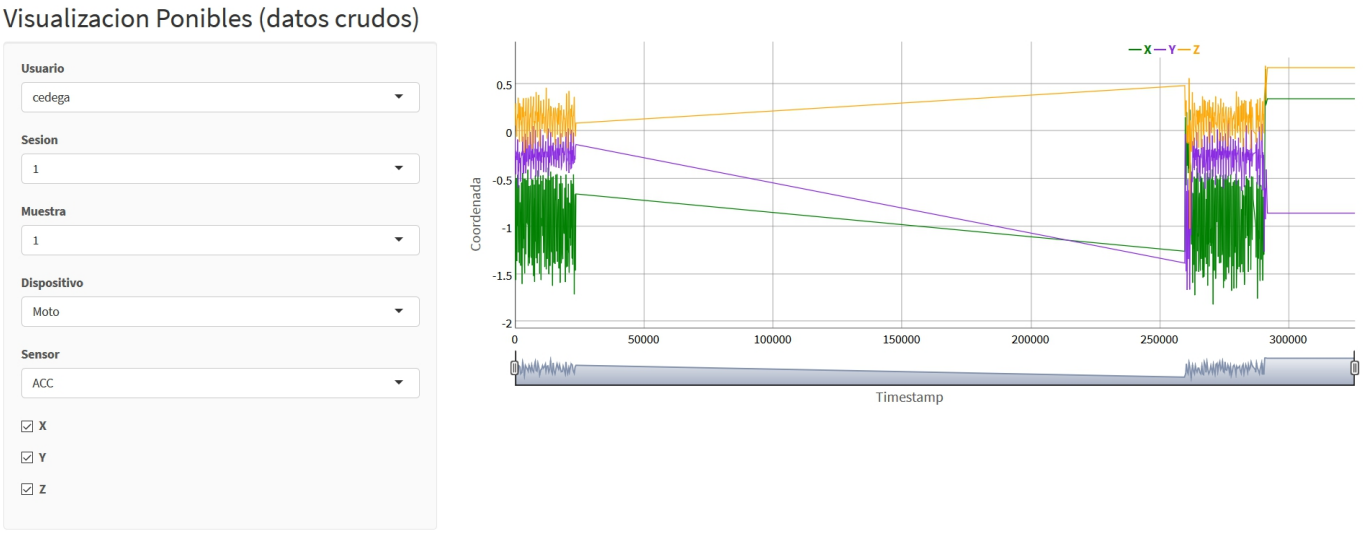


Figura 4.7: Estado de los datos crudos. Perdida de conexión intermedia-grave 1

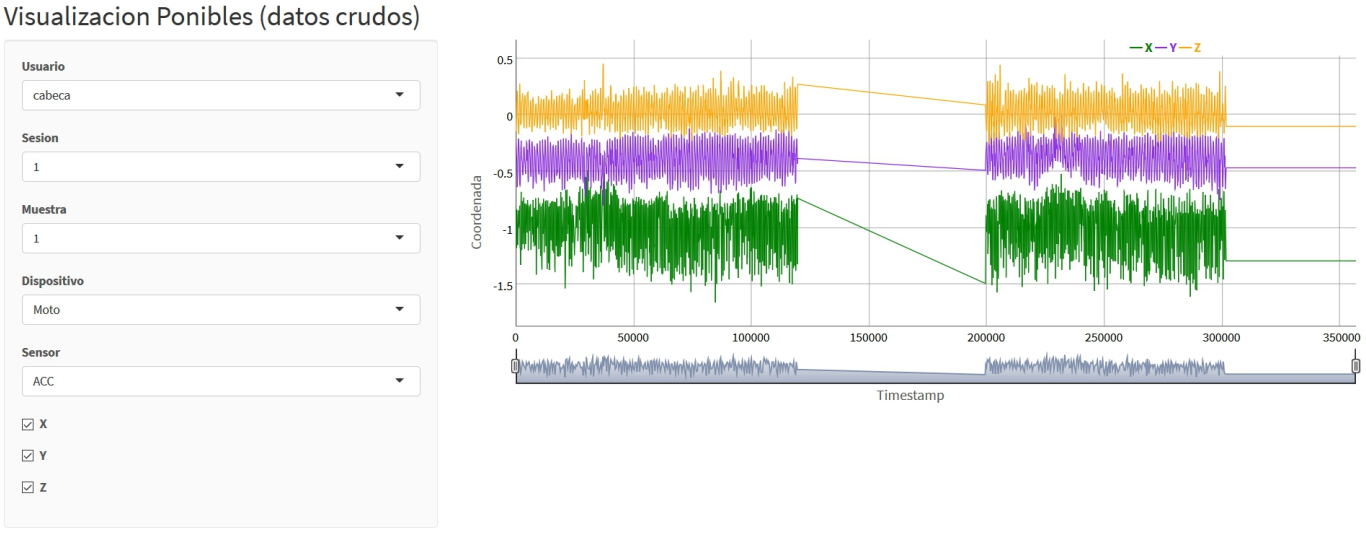


Figura 4.8: Estado de los datos crudos. Perdida de conexión intermedia-grave 2

- **Señal corta.** Se debe a un bloqueo de la aplicación, la cual aparentemente está funcionando bien, pero al guardar la muestra, la aplicación se bloquea y no hay más remedio que forzar su cierre. Ocurre en la muestra de la figura 4.9 que tiene un 26.8 % de los datos (75000ms, cuando lo normal en este usuario son 280000ms) y en la muestra de la figura 4.10 que contiene un 50 % de los datos. Ambos usuarios se descartan.

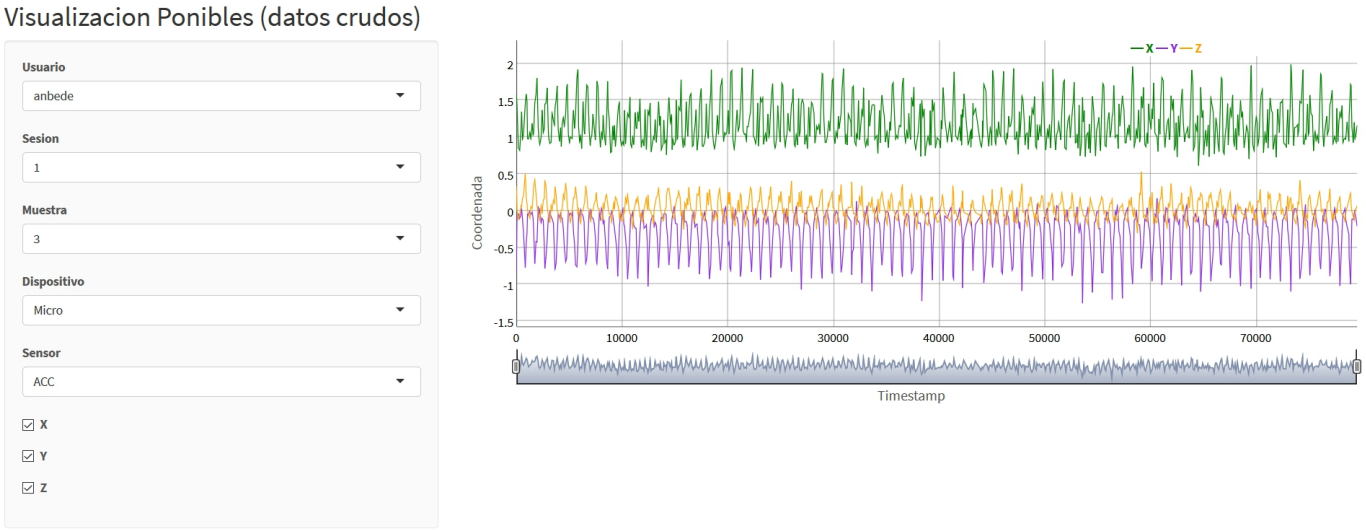


Figura 4.9: Estado de los datos crudos. Bloqueo grave de la aplicación 1.

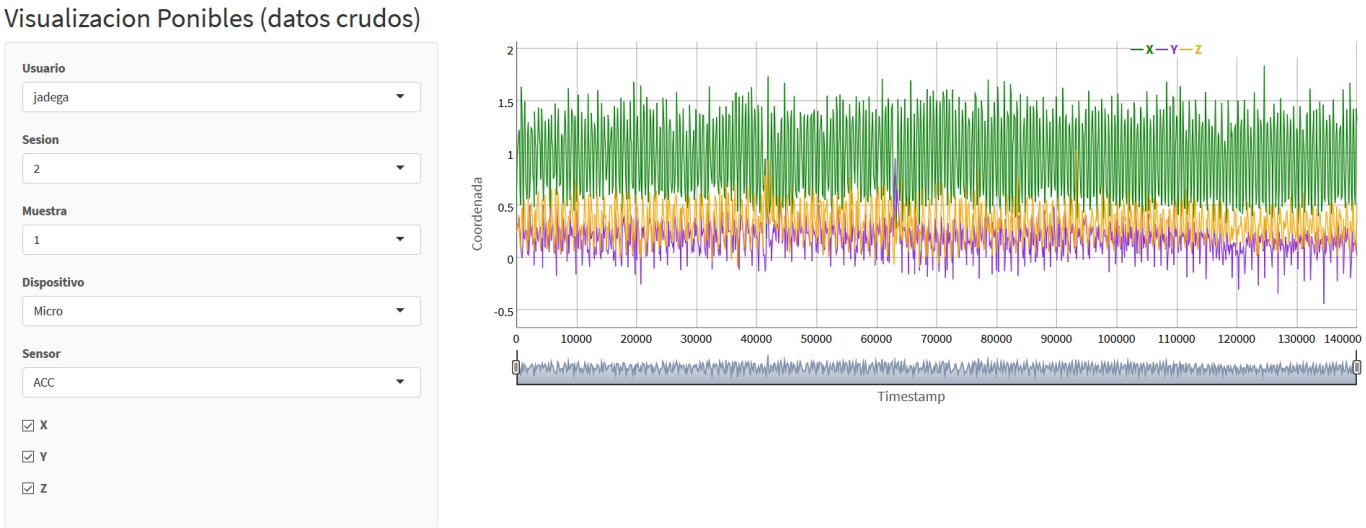


Figura 4.10: Estado de los datos crudos. Bloqueo grave de la aplicación 2.

- **Presencia de valores negativos.** Es un comportamiento poco frecuente pero que existe y se desconocen las razones que lo provocan. Se puede ver en la figura 4.11. Esta causado por la presencia de un valor negativo de tiempo entre una adquisición de datos y otra consecutiva, habiendo, en todos los casos, un único valor negativo por muestra (figura 4.12).

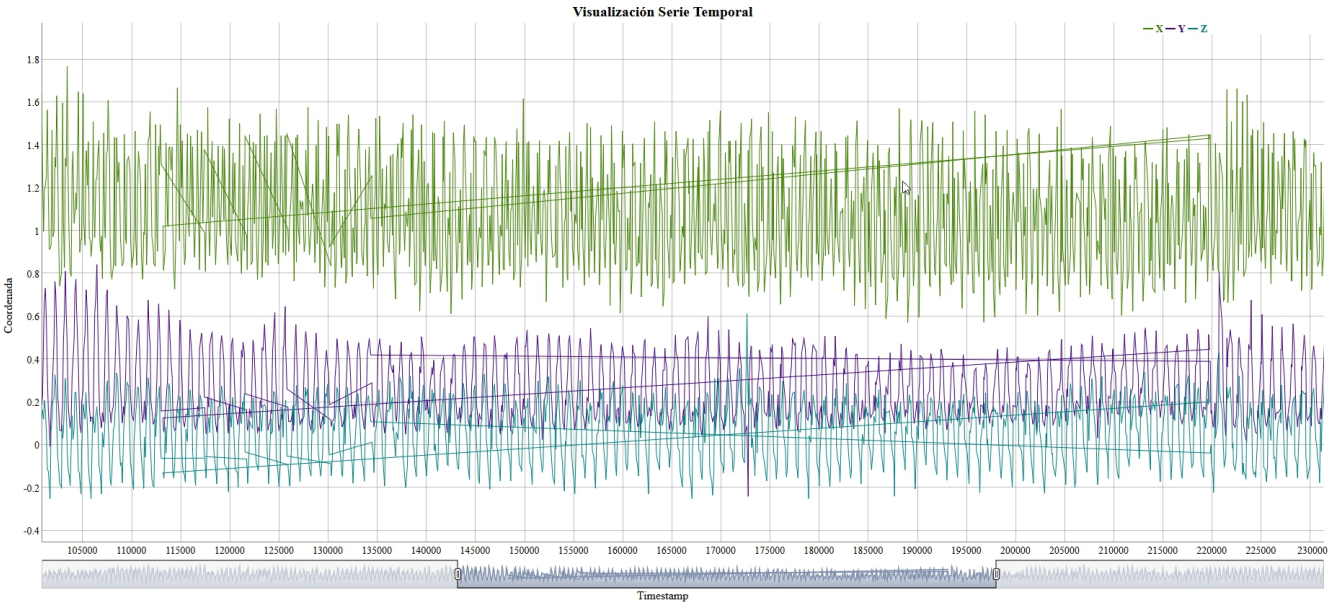


Figura 4.11: Estado de los datos crudos. Presencia de valores negativos en la muestra (visión del gráfico).

El análisis visual de los datos ha llevado a la eliminación de 4 usuarios por insuficiente información en alguna de sus muestras. Este análisis ha sido posible porque se tienen pocos usuarios, pero para corroborar lo visto y en espera de tener más usuarios y muestras, se ha construido un software automático de análisis y limpieza de los datos, que se explicará en el siguiente apartado.

4.3. Limpieza de los datos

La presencia de ruido en la señal es un problema que ya se ha tratado y resuelto en los trabajos anteriores a través de la eliminación de ventanas con baja autocorrelación. Zonas de la señal sin periodicidad, en las que el usuario anda mal o realiza un movimiento inesperado. La fórmula

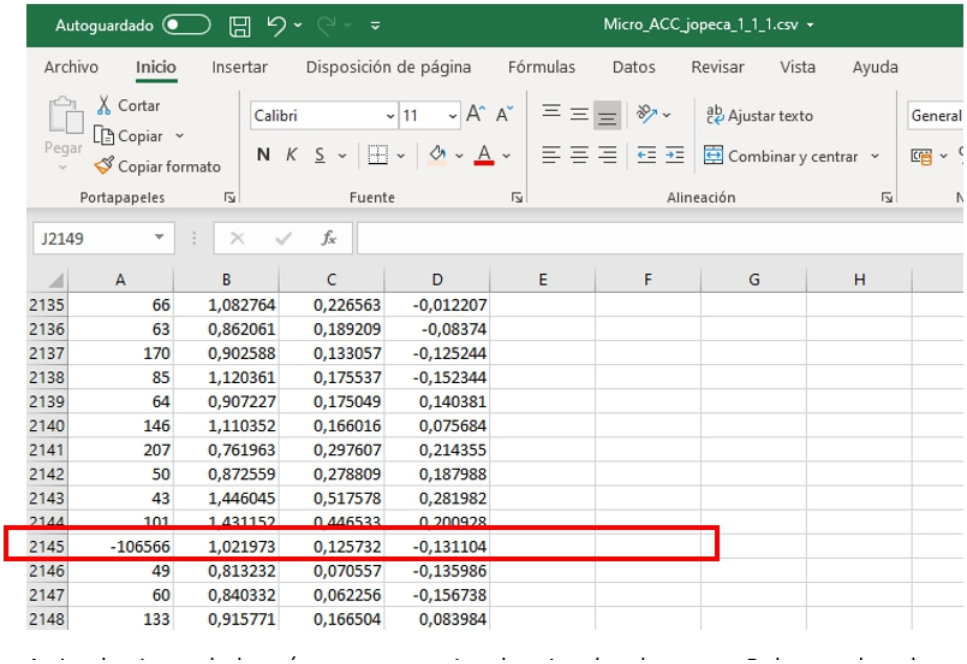


Figura 4.12: Estado de los datos crudos. Presencia de valores negativos en la muestra (visión de los datos).

empleada para calcular la autocorrelación de un proceso discreto X con n observaciones $X_1, X_2, ..., X_n$ es (4.1).

$$R(k) = \frac{1}{n \cdot \sigma^2} \cdot \sum_{t=1}^{n-k} (X_t - \mu) \cdot (X_{t+k} - \mu) \tag{4.1}$$

Donde n es el número de muestras de la ventana, X_t es el valor t-ésimo de X, μ y σ^2 representan respectivamente la media y la varianza de los valores de X, y el entero positivo $k<n$ es el desfase o desplazamiento en número de muestras para el cual queremos calcular la autocorrelación.

Los valores de *R(k)* están acotados entre -1 y 1. Un valor de autocorrelación de 1 indica que existe una correlación perfecta, mientras que un valor de -1 indica que hay una anticorrelación perfecta. Por otro lado, si el valor de autocorrelación es 0, indicará ausencia de correlación. Se buscan valores bajos en valor absoluto para encontrar ruido, es decir, ventanas en las que la señal no sigue un patrón que se repite a lo largo de la misma.

El cálculo de la autocorrelación se puede ver como si dividiéramos la ventana en subventanas de tamaño k muestras y obtuviéramos cuánto

se parecen esas subventanas entre sí. La relación entre el valor de k y la duración de la ventana en segundos, τ, la podemos ver en la fórmula (4.2), donde T es el periodo de muestreo de la señal. Como la señal se muestrea a una frecuencia de $Fm = 12s.$, el valor de T ($\frac{1}{Fm}$) es de, aproximadamente, 83 ms.

La duración de un paso andando es de, aproximadamente, un segundo, valor alrededor del cual se deben obtener los mayores valores de autocorrelación. Por lo que ese coeficiente se ha calculado para valores de k entre 11 y 16, que despejando en (4.2) nos da tamaños de ventana, τ, en milisegundos de entre 830 y 1245.

$$k = \frac{\tau}{T} + 1 \tag{4.2}$$

Pero existe un nuevo problema, la presencia de zonas planas donde no hay señal. Las ventanas de datos se construyen a través de un número fijo de adquisiciones de datos y hay que tomar la siguiente decisión sobre cada una de ellas:

- Es ruido y, por tanto, se descarta.
- No es ruido y se puede usar para reconocer al usuario.

Para detectar este problema de manera automática, se han utilizado las siguientes métricas:

1. Tiempo de adquisición entre dos datos consecutivos alto. Significa que ha pasado mucho tiempo y ha habido una posible pérdida de conexión. Al estar construyendo ventanas de datos con un número fijo de adquisiciones de datos, se pretende evitar la inclusión de estas pérdidas de conexión innecesarias.
2. Cruces por el valor medio de la señal. El análisis se ha hecho con los datos normalizados a media 0 y, por tanto, se ha buscado el número de cruces por el valor 0 en cada ventana de datos. Si la señal es plana se esperan valores bajos.
3. La energía de la señal (ecuación 4.3) proporciona una idea de la disposición tridimensional de los datos. En zonas planas se esperan valores de energía bajos.

$$Energia = \frac{1}{N} \cdot \sum_{n=1}^{N} (\sqrt{X[n]^2 + Y[n]^2 + Z[n]^2})^2 \tag{4.3}$$

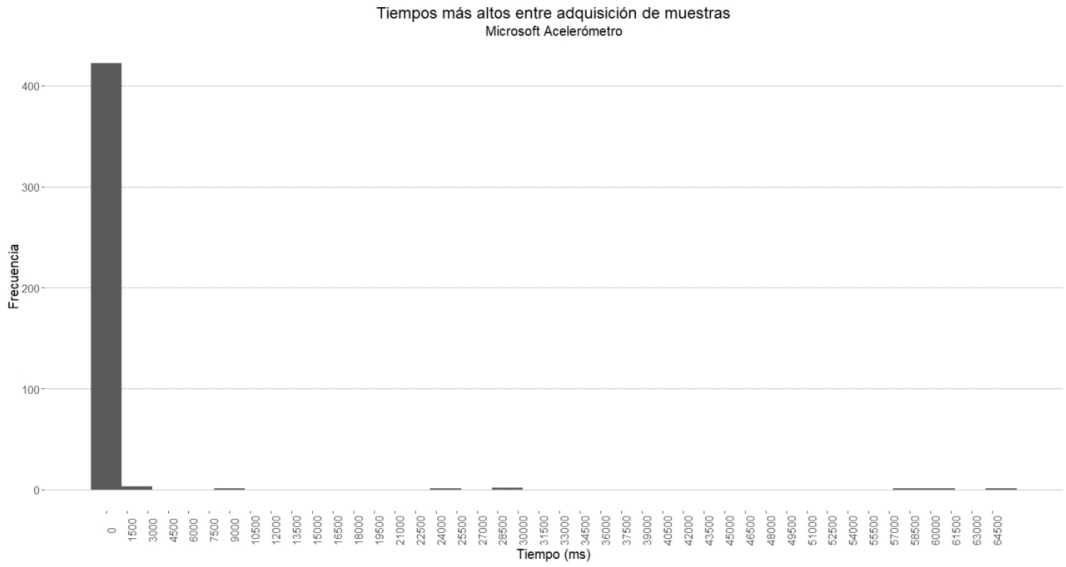
Micro ACC	Micro GYR	Moto ACC	Moto GYR
96.59	96.64	108.44	108.41

Tabla 4.4: Tiempos medios de adquisición entre muestras consecutivas, en milisegundos (ms).

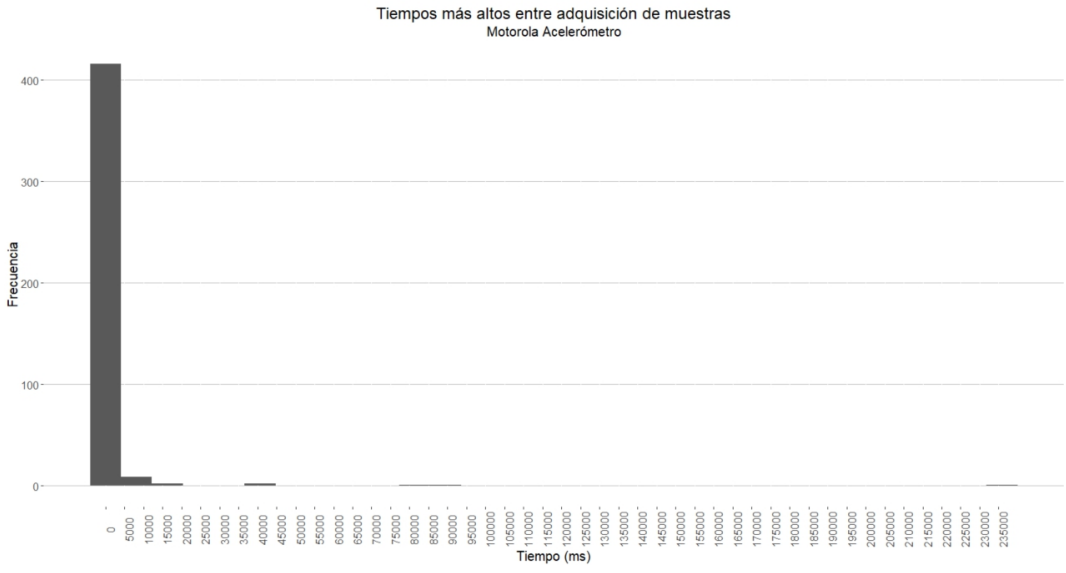
Analizando los tres tiempos más altos de adquisición entre datos consecutivos de todas las muestras adquiridas de todos los usuarios se ve que la mayor parte se encuentran entre 0 y 1000 milisegundos, habiendo muchos valores atípicos y siendo más altos en el dispositivo Moto (figuras 4.13 y 4.14). Sin embargo, el valor medio de tiempos de adquisición, eliminando estos máximos, es de 96 milisegundos en el dispositivo Micro y 108 milisegundos en el dispositivo Moto (coincidiendo los mismos valores en los dos sensores como se puede ver en la tabla 4.4).

A la hora de automatizar el valor umbral de las 4 métricas (autocorrelación de la señal para la eliminación del ruido y las métricas de los 3 ítems que se acaban de explicar para eliminar la presencia de zonas planas), se han hecho pruebas, observado las señales y calculado el porcentaje de ventanas eliminadas y una tasa de error comparando una limpieza manual y automática de la señal de 3 usuarios. Si la ventana no cumple una o más de las métricas, se ha eliminado. Los resultados se muestran en la tabla 4.5. El umbral se ha fijado en función del valor máximo, medio o la mediana de los valores de la métrica en dicho usuario y muestra. La opción elegida ha sido la 7 (P7-max) que, aunque no obtiene la mejor tasa de error, sí los mejores resultados, eliminando un 4.85 % de las ventanas, por las siguientes razones:

- Al estar eliminando zonas innecesarias de la señal, es mejor utilizar el valor máximo de la muestra porque ayuda a centrarse en los valores realmente malos. En la figura 4.15 se muestra una señal con la zona eliminada en color rojo, utilizando los criterios elegidos con el máximo y los mismos con la media y se ve que el máximo funciona mejor. La media elimina poco ruido en usuarios con ruido esporádico y mucho en usuarios con ruido frecuente, quedándose en estos últimos prácticamente sin señal de interés. Lo mismo ocurre con la mediana en la figura 4.16.

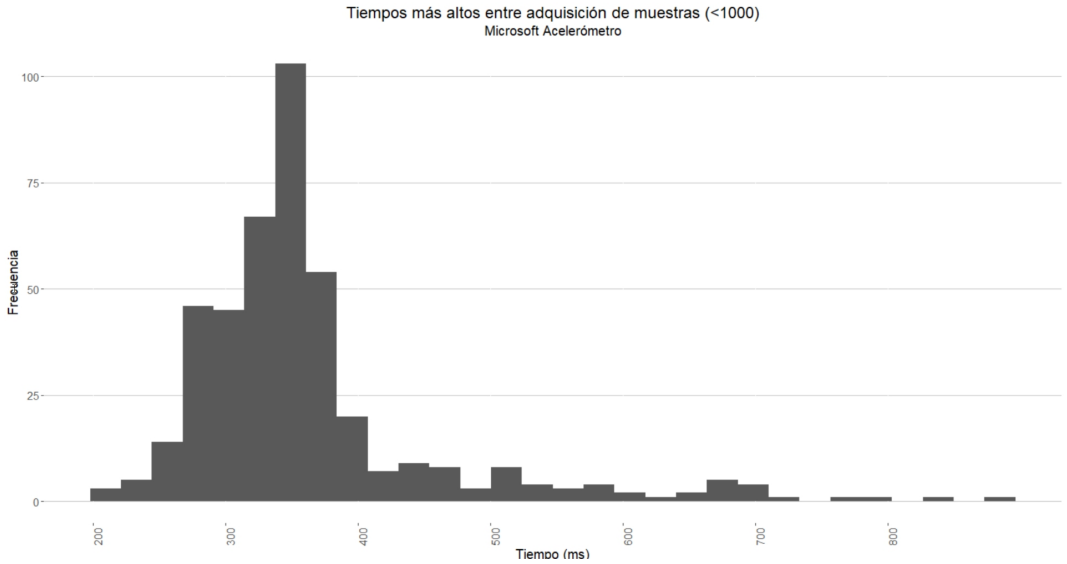


(a) Micro

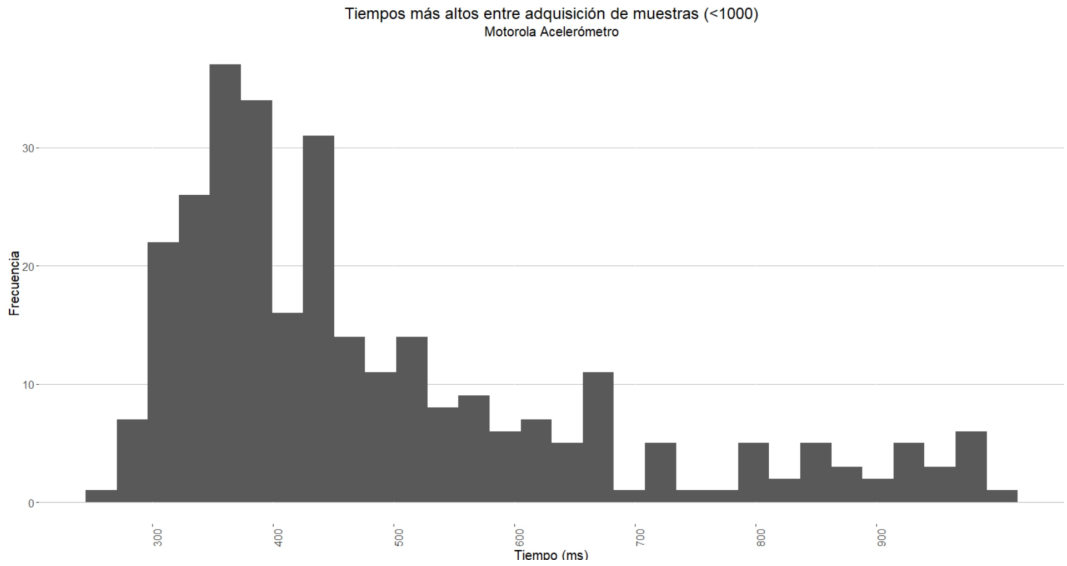


(b) Moto

Figura 4.13: Distribución de valores altos en el tiempo de adquisición de muestras consecutivas.



(a) Micro



(b) Moto

Figura 4.14: Distribución de los valores altos de interés en el tiempo de adquisición de muestras consecutivas.

- Umbral de cortes por 0 del 25 %. Parece un valor apropiado que funciona bien. No queremos eliminar mucha señal, únicamente las partes planas.

- Umbral de autocorrelación del 25 %. Valores altos han demostrado eliminar mucha señal y no beneficiar a los resultados (P1). De todas formas, se ha visto que está relacionado con el criterio elegido, y al utilizar el máximo benefician umbrales más bajos, pero al utilizar la media o la mediana son necesarios umbrales más altos.

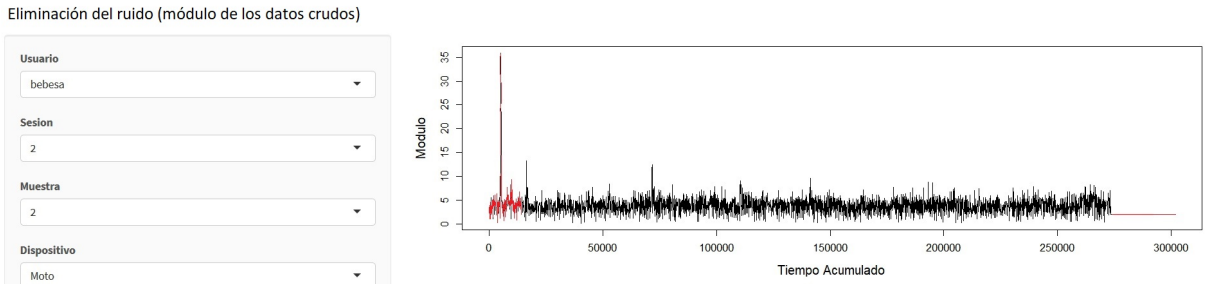
- Umbral de energía del 10 %. Se ha visto que la ventana tenía que ser realmente mala para obtener valores de energía bajos. Por ello, fijar umbrales más altos no beneficiaba a los resultados y eliminaba ventanas que realmente eran buenas.

- Eliminar aquellas ventanas que tienen al menos un tiempo de adquisición (separación entre dos muestras) superior a 550ms. Aunque puede parecer un valor alto comparado con la media (100 ms), no lo es tanto, ya que lo que se busca es eliminar ventanas realmente malas; la presencia de algo de ruido en la señal puede ser característico del usuario y ayudar en su reconocimiento.

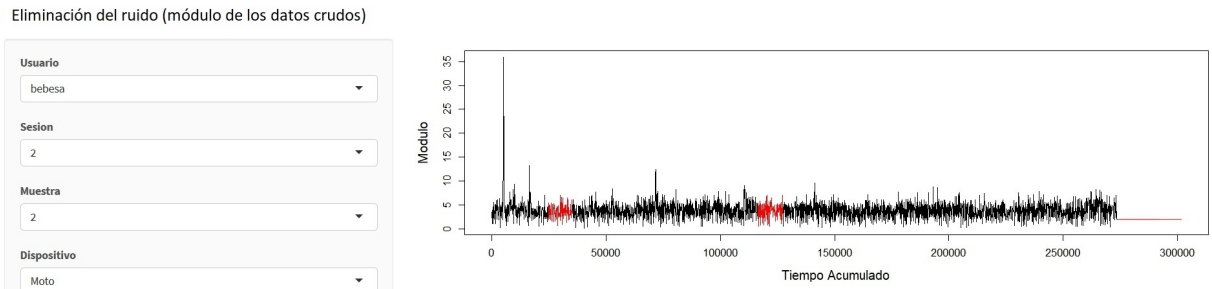
Prueba	Cortes por 0	Autocorrelación	Energía	T. Adquisición	% Eliminadas	Tasa Error
P1-max	0.25	0.75	0.3	450	53.23	66.14
P2-max	0.25	0.6	0.25	450	30.13	42.88
P3-max	0.25	0.5	0.2	450	18.76	25.48
P4-max	0.3	0.4	0.2	450	11.84	14.78
P5-max	0.25	0.4	0.15	450	10.79	14.92
P6-max	0.25	0.3	0.15	550	5.94	8.96
P7-max	**0.25**	**0.25**	**0.1**	**550**	**4.85**	**7.12**
P8-mean	0.25	0.25	0.1	550	3.87	5.72
P9-mean	0.25	0.5	0.1	550	5.64	8.11
P10-mean	0.25	0.4	0.1	450	5.69	7.97
P11-mean	0.25	0.3	0.1	500	4.54	6.98
P12-median	0.25	0.3	0.1	500	4.58	6.88

Tabla 4.5: Análisis de umbrales para automatizar la limpieza de la señal.

La matriz de confusión con el funcionamiento de la opción elegida se muestra en la tabla 4.6 donde “SI” es ruido y “NO” señal.



(a) Máximo



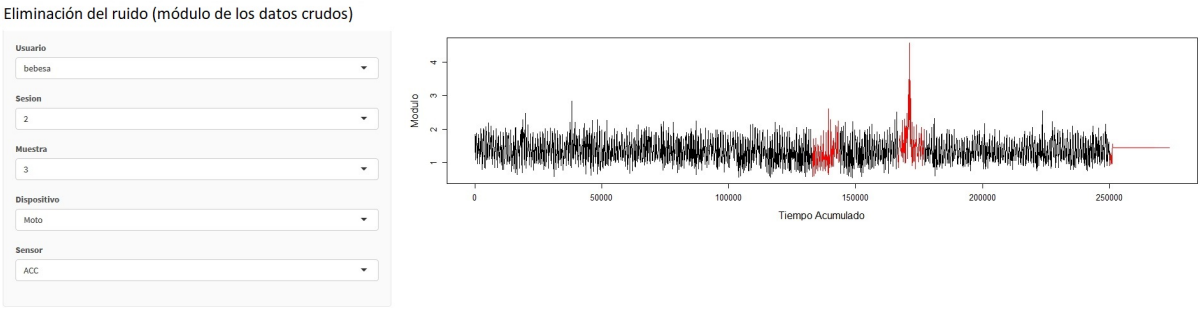
(b) Media

Figura 4.15: Comparación del máximo y la media como criterio de eliminación del ruido. En color rojo el ruido y en negro, la señal.

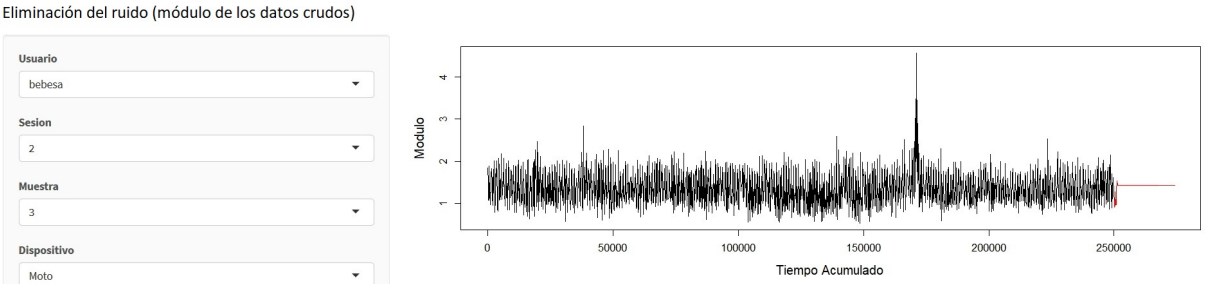
En la figura 4.17 se pueden ver ejemplos de señales con el sistema de limpieza automática construido (en color rojo ruido y en negro señal), sistema final de limpieza que queda como sigue:

- **Corrección de valores negativos** en el tiempo de adquisición de dos datos consecutivos. Se ha corregido de manera automática con la lectura de los datos siguiendo el procedimiento que se detalla a continuación y que se seguirá únicamente en el caso de que se detecte alguna diferencia de tiempos negativos en la muestra (el software R mostrará un mensaje informativo como el de la figura 4.18).

    - Crear una columna con el valor de tiempo relativo acumulado.



(a) Máximo



(b) Mediana

Figura 4.16: Comparación del máximo y la mediana como criterio de eliminación del ruido. En color rojo el ruido y en negro, la señal.

- Ordenar los datos de la tabla de menor a mayor usando la columna anteriormente creada.
  - Si existen valores negativos, son puntos espurios, se eliminan todas las filas que los contienen.
  - Si no existen valores negativos, no se elimina ninguna fila.
- Eliminar la columna que contiene el tiempo relativo original.
- Crear una nueva columna de tiempo relativo, pero ahora sin puntos erróneos. Empezando en valor 0 y restando los valores acumulados en ese momento con el anterior

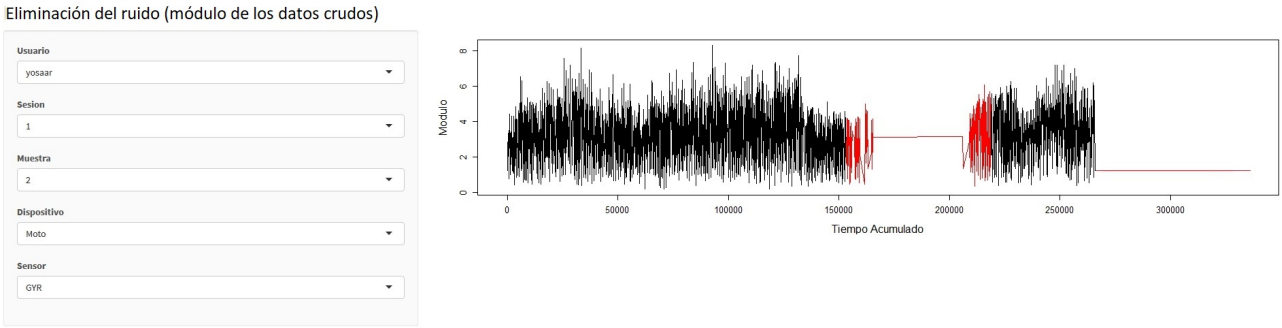
- **Eliminación del ruido.** Eliminación de aquellas ventanas que cumplen al menos uno de los siguientes criterios.

  - Número de cortes por 0 inferior al 25 % del valor máximo de dicho usuario y muestra.
  - Valor de autocorrelación inferior al 25 % del valor máximo de dicho usuario y muestra.
  - Valor de energía inferior al 10 % del valor máximo de dicho usuario y muestra.
  - Presencia de al menos un tiempo de adquisición entre datos consecutivos superior a 550ms.

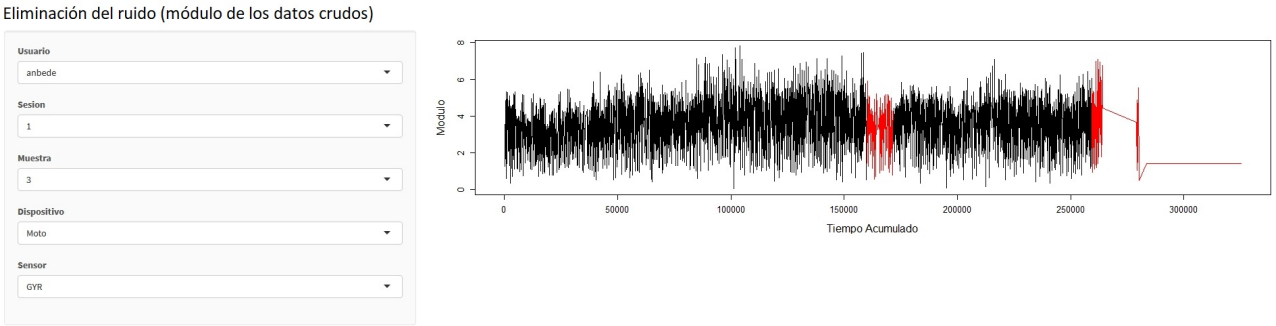
- **Análisis final de la señal resultante.** Comprobación del número de ventanas sin ruido en todas las muestras disponibles del mismo usuario. Si existe alguna con menos del 50 % que la muestra que más tiene del mismo usuario (máximo), se elimina dicho usuario completo porque tiene una señal demasiado corta para poder abordar su reconocimiento. Así se eliminarían los 4 usuarios mencionados en el apartado anterior (sección 4.2).

	NO	SI
NO	2568	104
SI	105	159

Tabla 4.6: Limpieza automática de la señal. Matriz de confusión del sistema automático de limpieza de la señal comparado con la limpieza manual, suponiendo éste último como el correcto. “SI” es ruido y “NO” señal.



(a)



(b)

Figura 4.17: Ejemplos de funcionamiento del sistema de limpieza automática de la señal.

```
Fin del usuario  anqude , crear tabla
El usuario bebesa , en la sesion 1 , muestra 2 , dispositivo Moto  tiene 1 valores negativos. Se eliminan.
El usuario bebesa , en la sesion 1 , muestra 2 , dispositivo Moto  tiene 1 valores negativos. Se eliminan.
Fin del usuario  bebesa , crear tabla
Fin del usuario  cabeca , crear tabla
```

Figura 4.18: Estado de los datos crudos. Presencia de valores negativos en la muestra (detección).

Capítulo 5

Configuración Experimental

La realización de un experimento requiere tomar muchas decisiones, las cuales pueden terminar siendo un éxito o un fracaso. A lo largo de este capítulo se explican los parámetros que se han fijado y utilizarán en los siguientes capítulos junto con las razones por las cuáles se han elegido. En esto se basaba el trabajo previo [57], en abordar una primera aproximación al problema, analizando los distintos elementos que entran en juego en el sistema, para entenderlos mejor y ver su relación en el rendimiento del reconocimiento del usuario. En definitiva, plantar unas bases sólidas sobre las que seguir trabajando, ya sí, en un sistema eficiente de reconocimiento, como se hará en capítulos posteriores.

Un usuario realiza dos operaciones en un sistema biométrico:

- **Inscripción:** proceso en el que el usuario proporciona los datos biométricos para construir su modelo o plantilla biométrica, que se almacenará en el sistema.
- **Autenticación:** proceso en el que una muestra biométrica de origen desconocido debe ser clasificada (autenticada) como perteneciente o no a la identidad del reclamante, mediante la comparación o emparejamiento de la muestra de entrada con el modelo o plantilla del reclamante.

Desde el momento en que se adquiere una muestra biométrica sin procesar hasta que se crea o se compara la plantilla, dependiendo de la operación, se debe transformar para extraer características adecuadas que se procesarán en el algoritmo de coincidencia o clasificador utilizado. Esta etapa se llama extracción de características y se compone de dos pasos principales:

- **Preprocesamiento:** donde la señal sin procesar se adapta y modifica para ser procesada de una manera más adecuada en el siguiente paso.
- **Extracción de características:** donde se extraen las características que se utilizarán para crear el modelo o para comparar.

A continuación, se describirá cada parte con mayor detalle.

5.1. Preprocesamiento

Junto con la limpieza de la señal, extrayendo la región de interés (ROI) que ya se ha explicado en la sección 4.3, en [57] se probaron las siguientes técnicas de preprocesamiento.

- **Normalización del periodo:** la señal adquirida tiene diferentes separaciones de tiempo entre dos datos consecutivos, es decir, no se adquiere con una frecuencia de muestreo fija. La alternativa para conseguir diferencias de tiempo fijas es remuestrear la señal [39, 48], que se puede hacer a través de un método de interpolación o analizando las componentes de frecuencia de los datos, donde aquí podemos ver que los componentes mayores de 6Hz son insignificantes. Luego, siguiendo el teorema de Nyquist-Shannon, se fijó una frecuencia de muestreo de 12Hz (un período de 83,3ms). Valor de acuerdo con el mostrado en [79], donde se demuestra que movemos el brazo a un máximo de 8,6 Hz, haciendo el movimiento más rápido posible. Como los datos se recopilan caminando, entonces una frecuencia de muestreo de 12 Hz (frecuencia máxima en la señal de 6 Hz) parecería razonable.
- **Normalización de la amplitud:** El objetivo es cambiar el valor de los datos a una escala común. La necesidad de realizar esta operación depende del algoritmo de aprendizaje automático.
- **Filtrado:** El objetivo es suavizar la señal. Las ventajas de su uso no están claras, por eso se analizó. A partir de las técnicas aplicadas en la bibliografía, se probó uno de los algoritmos más explotados [48, 91, 72], la media móvil ponderada (WMA), donde D_t, D_{t-1}, D_{t+1} son los datos adquiridos en el instante t, y en el anterior y posterior a t:

$$wma = \frac{D_{t-1} + D_t + D_{t+1}}{3} \tag{5.1}$$

5.2. Análisis de parámetros

Para la extracción de características, en [57] se realizó un análisis de los siguientes parámetros y técnicas.

- **Tamaño de la ventana:** la señal se divide en ventanas, cuyo tamaño se mide en número de ciclos de caminata. Se probó la influencia del tamaño de la señal (tamaño de la ventana) con tamaños de 2 a 15.
- **Calidad de la señal:** en lugar de utilizar toda la señal, considerar sólo las piezas de calidad. Como medida se utilizó la máxima autocorrelación de la ventana, por lo que solo las ventanas cuya autocorrelación es superior a un umbral se utilizarán en la plantilla y autenticación del usuario. Se probaron diferentes valores umbrales.
- **Fusión de ventanas:** muchos clasificadores pierden la relación temporal entre ventanas, es decir, tratan cada ventana por separado. Por lo que se probó a incluir esa información temporal para mejorar el reconocimiento. La fusión se puede realizar en la etapa de extracción de características o en la etapa de clasificación, que es la que se utiliza aquí, fusionando scores $s_i(F_i^t/\lambda_C)$ (s_i en resumen), de varias ventanas consecutivas, n, como salida final de la etapa de comparación, s_j^*, como muestra la figura 5.1. F hace referencia al vector de características (representación matemática de la muestra biométrica) y λ_c representa la plantilla del usuario C (modelo). En el trabajo indicado se probaron varias técnicas de fusión (estadísticos como la media, la mediana, el máximo y el mínimo), obteniendo los mejores resultados con la media y la mediana y la influencia del número de ventanas fusionadas, n y la superposición entre ellas.

5.3. Extracción de características

Las señales del acelerómetro y del giroscopio son series temporales, por lo que la bibliografía muestra dos formas de realizar la extracción de características: en el dominio del tiempo [81, 19] y en el dominio de la frecuencia [76, 88].

Extracción de características del Dominio del Tiempo

El objetivo es analizar la capacidad de las señales del acelerómetro y del giroscopio en la caracterización de la marcha del usuario. Aquí se

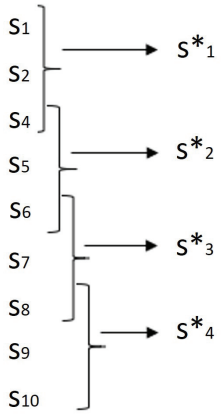


Figura 5.1: Ejemplo de fusión de ventanas a nivel de scores, para $n = 3$ y solapamiento 1. s_i es el score original (salida del clasificador para cada ventana de muestra de prueba), mientras que s_i^* es el nuevo score como resultado de fusionar n scores originales.

van a utilizar características basadas en las empleadas en la literatura [81, 19, 91, 45]. Más específicamente, la extracción de características de una muestra se logra de la siguiente manera:

1. Cada coordenada de la muestra (X/Y/Z) se divide en segmentos llamados ventanas (Figura 5.2) [81, 88] que se superponen.
2. De cada ventana, se extraen las siguientes características de cada coordenada del sensor: media, mediana, máximo, mínimo, desviación estándar, rango máximo (máximo-mínimo), curtosis, percentil 25, percentil 75, coeficiente de asimetría, energía, valor máximo de la autocorrelación, la relación media de los componentes XY, XZ y YZ.

Se muestra una breve descripción de cada una de las características.

- Medidas estadísticas:
  - Media: medida de tendencia central que representa el centro de gravedad de la distribución de la variable.
  - Mediana: medida de tendencia central que representa el valor de la variable en la posición central entre un conjunto de datos ordenados.
  - Máximo: el valor más grande entre el conjunto de valores.

- Mínimo: el valor más pequeño entre el conjunto de valores.
- Desviación estándar: medida de dispersión que se utiliza para cuantificar la variación de un conjunto de datos.
- Rango: intervalo entre el valor máximo y el valor mínimo, proporcionando una idea de la dispersión de los datos.
- Curtosis: característica de forma de la distribución de probabilidad/frecuencias de los datos que indica cómo de apuntada o achatada se encuentra una distribución respecto a un comportamiento normal (distribución normal). Valores grandes indican mayor concentración de valores de la variable tanto muy cerca de la media de la distribución (pico) como muy lejos de ella (colas), al tiempo que existe una relativamente menor frecuencia de valores intermedios, no implicando con ello una mayor varianza [84].
- Cuantil 25 % y 75 %: puntos tomados a intervalos regulares de la función de distribución de la variable aleatoria. Lo que se ha utilizado ha sido dividir la distribución en cuatro partes correspondientes a los cuantiles 25 %, 50 % (media) y 75 %.
- Coeficiente de asimetría: representa el grado de simetría (o asimetría) de la distribución de probabilidad de la variable aleatoria. Considerando como eje de simetría la recta paralela al eje de ordenadas que pasa por la media de la distribución. Una distribución es simétrica si existe el mismo número de valores a la derecha que a la izquierda de la media y por tanto el mismo número de desviaciones con signo positivo que con signo negativo. Mientras que hay asimetría positiva si hay valores más separados de la media por la derecha y asimetría negativa si hay valores más separados de la media por la izquierda [85].

- Energía conjunta de las 3 componentes: proporciona una idea de la disposición tridimensional de los datos.

$$Energia = \frac{1}{N} \cdot \sum_{n=1}^{N} (\sqrt{X[n]^2 + Y[n]^2 + Z[n]^2})^2 \tag{5.2}$$

- Valor máximo de la autocorrelación en cada una de las componentes: representa la relación de la señal consigo mismo.

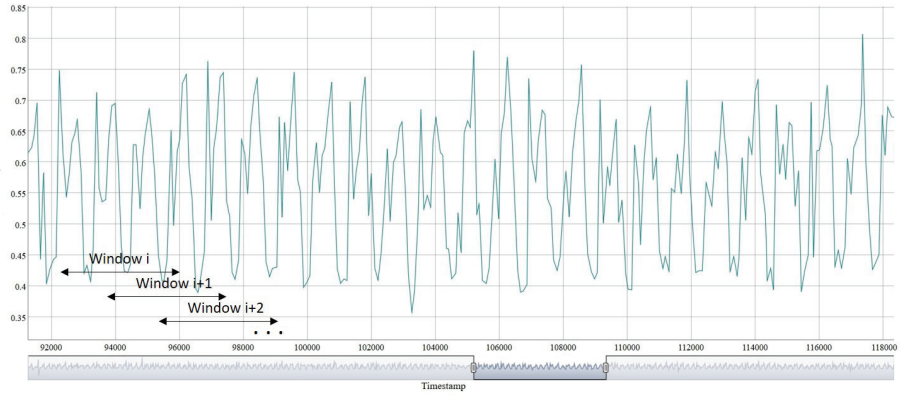


Figura 5.2: Ejemplo de ventana de muestra.

- Ratio medio de las componentes XY: media de todos los cocientes de X e Y. Proporciona una medida de la gravedad de la distribución conjunta de ambas componentes de la variable.

- Ratio medio de las componentes XZ: media de todos los cocientes de X y Z.

- Ratio medio de las componentes YZ: media de todos los cocientes de Y y Z.

En [57] se realizaron las siguientes pruebas:

1. Usando cada coordenada del sensor (X/Y/Z) por separado.

2. Creando un vector de características que une los vectores de cada coordenada del sensor, es decir todas las características de todas las coordenadas.

3. Construyendo el módulo [19]: $\sqrt{X^2 + Y^2 + Z^2}$ y extrayendo las mismas características sobre esta nueva señal fusionada.

En este proyecto solo se van a realizar experimentos con las opciones 1 y 3, ya que la opción 2 mostró malos resultados, con, además, un incremento grande del tamaño del vector de características.

Extracción de características del Dominio de la Frecuencia

Como en el dominio del tiempo, la señal se divide en ventanas. La Transformada Rápida de Fourier (FFT) se aplica a cada ventana para convertir la señal en una representación en el dominio de la frecuencia. A partir del resultado, un conjunto de números complejos, solo se usa el módulo, es decir, las amplitudes de los componentes de frecuencia de la señal. Después de eliminar la componente cero, se extraen las siguientes características:

- Primera y segunda amplitud dominante: representa los dos valores más altos obtenidos entre las amplitudes resultantes del Análisis de la transformada de Fourier en cada una de las componentes de los datos.
- Primera y segunda frecuencia dominante: representa los dos valores de la frecuencia correspondientes a los dos puntos donde se consiguen las amplitudes anteriores.
- Área bajo la curva de Fourier (AUC) basado en splines: utiliza una interpolación de splines para calcular la cantidad de área bajo la curva formada por las amplitudes del Análisis de Fourier.
- Las mismas medidas estadísticas que en el dominio del tiempo, quitando el máximo y el mínimo.

La ecuación (5.3) muestra la expresión de la Transformada Rápida de Fourier (FFT) [52].

$$X_k = \sum_{n=0}^{N-1} X_n exp\left(\frac{-i2\pi kn}{N}\right) \tag{5.3}$$

Donde:

- X_k: cantidad de frecuencia k en la señal; cada valor *k-ésimo* es un número complejo que incluye amplitud (fuerza) y cambio de fase.
- N: número de muestras.
- n: muestra, $n\epsilon\{0...N-1\}$.
- k: frecuencia entre *0 Hz* y *N-1 Hz*.
- $1/N$: tamaño real de los picos de tiempo.

- n/N: porcentaje de tiempo que ha pasado.
- $2\pi k$: velocidad en *radianes/segundo*.
- $\exp^{-ix}$: camino circular hacia atrás que indica cuánto nos hemos movido, para esta velocidad y tiempo.

En este dominio se han realizado las mismas pruebas que se indicaban en el apartado de las características del dominio del tiempo.

5.4. Medición del error

Otra decisión importante es cómo evaluar los modelos implementados con el objetivo de poder compararlos y buscar la mejor solución final.

Entre las medidas más utilizadas en los sistemas biométricos se encuentran las curvas ROC *(Receiver Operating Characteristic)*, éstas son una representación gráfica de la sensibilidad frente a la especificidad para un sistema de clasificación binario según se varía el umbral de decisión.

Nuestro problema se corresponde con el de clasificación binaria, dado que, para cada usuario, se considera a dicho usuario como auténtico y al resto como usuarios impostores.

Las medidas de error básicas usadas en este tipo de problemas son:

- Falsos positivos (False Positives o FP) o falsa aceptación: ocurre cuando se identifica a una persona no autorizada como autorizada. De manera que, si el sistema trata de verificar la identidad de una persona, un usuario impostor podría acceder de forma no autorizada.
- Falsos negativos (False Negatives o FN) o falso rechazo: ocurre cuando se impide el acceso a una persona autorizada.
- Verdaderos positivos (True Positives o TP): ocurre cuando el sistema trata de verificar la identidad de una persona y un usuario auténtico (verdadero) accede de forma correcta y es autorizada.
- Negativos verdaderos (True Negatives o TN): ocurre cuando el sistema trata de verificar la identidad de una persona y un usuario impostor es rechazado.

- Sensibilidad (True Positive Rate o TPR): proporción de usuarios auténticos que se consideran correctamente como autorizados, con respecto a todos los usuarios auténticos. En función de los términos anteriores, se puede calcular con la fórmula (5.4).

$$Sensibilidad = \frac{TruePositive}{FalseNegative + TruePositive} = \frac{TP}{FN + TP} \tag{5.4}$$

- Especificidad (False Positive Rate o FPR): proporción de usuarios impostores que se consideran erróneamente como autorizados con respecto a todos los usuarios impostores, cuyo resultado se puede obtener con la fórmula (5.5).

$$Especificidad = \frac{FalsePositive}{FalsePositive + TrueNegative} = \frac{FP}{FP + TN} \tag{5.5}$$

Tanto la sensibilidad como la especificidad tienen valores en el rango [0,1], generando una curva ROC en estos rangos donde su área se denomina AUC. Los valores de AUC se interpretan de manera que cuanto mayor sea el valor del AUC, mejor es el rendimiento del modelo.

Otra medida del rendimiento muy utilizada en biometría es la tasa de equierror, que es el punto de intersección entre ambas tasas: sensibilidad y especificidad, conocido como *Equal Error Rate (EER)*. Cuanto menor sea su valor, mejor será el sistema.

La figura 5.3 muestra la especificidad en el eje de abscisas y la sensibilidad en el eje de ordenadas, generando la curva sobre su área (AUC) marcado en gris. El valor de la tasa de equierror se produce con FPR=0.2 y TPR=0.8.

Como resultado final, tenemos dos maneras de mostrar el error:

- De manera gráfica: como se muestra en la figura 5.3, los valores de la sensibilidad y la especificidad para distintos valores umbrales.
- Mediante valor numérico: utilizando el área bajo la curva ROC o la tasa de equierror explicada. Estos valores se pueden calcular de manera individual para cada usuario o de manera global como la media de todos los usuarios disponibles.

Dado que nuestro objetivo aquí es comparar resultados, la opción gráfica es poco práctica, por lo que se utilizarán directamente los valores

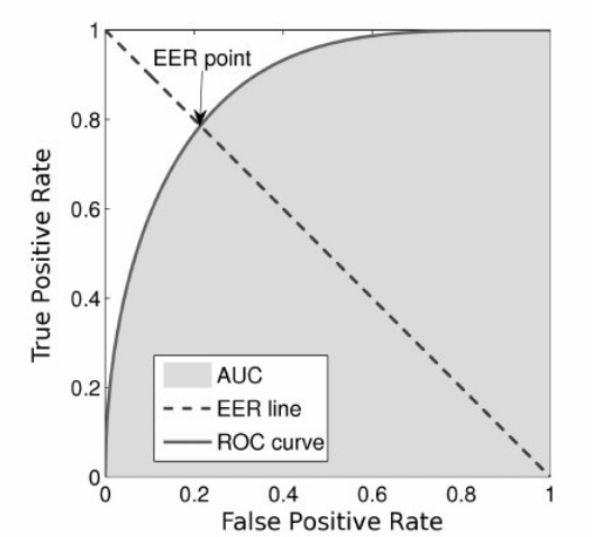


Figura 5.3: Ejemplo de EER a partir de la curva ROC y el AUC (medición del error).

numéricos. En [57] se consideraron ambas métricas (AUC y EER), en este proyecto nos centraremos únicamente en la tasa de equierror, por ser la más utilizada en la bibliografía sobre biometría. Para la toma de decisiones se utilizará su valor medio con respecto a todos los usuarios, aunque en ciertas ocasiones se aprovechará la ventaja de tener pocos usuarios para hacer un estudio detallado de cada uno de ellos.

5.5. Experimentos

Teniendo en cuenta el corpus biométrico que estamos usando, se tienen:

- Diversos **usuarios**.
- Dos **sesiones** posibles en que se recogieron datos. S1 y S2 hacen referencia a la sesión 1 y 2 respectivamente.
- Dos **muestras** de datos tomadas por sesión y pulsera a cada usuario, representándose como M1 y M2 para referenciar a la muestra 1 y 2 respectivamente.

En [57] se trabajó con el clasificador K-vecinos más próximos y clasificación no supervisada, donde para cada usuario i se tienen los siguientes conjuntos de datos:

- Conjunto de entrenamiento *(train)*: contiene los datos del usuario auténtico que se usarán para crear su patrón (fase de *inscripción*).
- Conjunto de prueba *(test)*: distinguiendo entre:
  - **Muestras auténticas:** Serán muestras del usuario distintas a las usadas para el entrenamiento. Se usarán para calcular la tasa de falsos negativos.
  - **Muestras impostores:** Serán muestras de otros usuarios distintos al usuario *i*. Simularán ataques al sistema, por lo tanto, se usarán para calcular la tasa de falsos positivos.

Con respecto a la sesión y muestra, se tienen los siguientes escenarios:

1. Monosesión-Monomuestra (MonoMono_M1): compara los datos dentro de la misma sesión, es decir, las muestras usadas para entrenamiento y para *prueba auténtica* del usuario son tomadas en la misma sesión. Es el caso más favorable.
   - **Train:** S1, M1, usuario i
   - **Test Auténticos:** S1, M2, usuario i
   - **Test Impostores:** S1, M2, usuario j $\neq$ i
2. Multisesión-Monomuestra (MultiMono_M1): Las muestras usadas para entrenamiento y *prueba auténtica* son tomadas en distintas sesiones. Aquí se quiere probar la variabilidad del rasgo biométrico con el tiempo.
   - **Train:** S1, M1, usuario i
   - **Test Auténticos:** S2, M1, usuario i
   - **Test Impostores:** S1, M2, usuario j $\neq$ i
3. Multisesión-Multimuestra (MultiMulti_M1): En biometría se ha demostrado que la variabilidad del rasgo con el tiempo es un problema que afecta al rendimiento del sistema. Una forma de paliarlo es intentar incluir en el modelo del usuario esta variabilidad. Una manera de hacerlo es usar para entrenamiento muestras de distintas sesiones, teniendo de esta manera una mayor cantidad de muestras.
   - **Train:** S1 y S2, M1, usuario i
   - **Test Auténticos:** S1 y S2, M2, usuario i

- **Test Impostores:** S1 y S2, M2, usuario j ≠ i

El protocolo experimental seguido para cada caso, *Monosesión-Monomuestra*, *Multisesión-Monomuestra* y *Multisesión-Multimuestra*, es el indicado. Ahora bien, otra forma que puede parecer más razonable de actuar es considerar las distintas posibilidades dentro de *Monosesión-Monomuestra*, que serían la indicada junto con las siguientes tres:

1. MonoMono_M2:
   - **Train:** S1, M2, usuario i
   - **Test Auténticos:** S1, M1, usuario i
   - **Test Impostores:** S1, M1, usuario j ≠ i
2. MonoMono_M3:
   - **Train:** S2, M1, usuario i
   - **Test Auténticos:** S2, M2, usuario i
   - **Test Impostores:** S2, M2, usuario j ≠ i
3. MonoMono_M4:
   - **Train:** S2, M2, usuario i
   - **Test Auténticos:** S2, M1, usuario i
   - **Test Impostores:** S2, M1, usuario j ≠ i

De la misma forma, en *Multisesión-Monomuestra* existirían además de la mencionada las siguientes tres:

1. MultiMono_M2:
   - **Train:** S1, M2, usuario i
   - **Test Auténticos:** S2, M2, usuario i
   - **Test Impostores:** S1, M1, usuario j ≠ i
2. MultiMono_M3:
   - **Train:** S2, M1, usuario i
   - **Test Auténticos:** S1, M1, usuario i
   - **Test Impostores:** S2, M2, usuario j ≠ i

3. MultiMono_M4:

   - **Train:** S2, M2, usuario i
   - **Test Auténticos:** S1, M2, usuario i
   - **Test Impostores:** S2, M1, usuario j ≠ i

Y por último, en *Multisesión-Multimuestra* existiría, además de la mencionada otra más que dividiría los datos de la siguiente manera:

1. MultiMulti_M2:

   - **Train:** S1 y S2, M2, usuario i
   - **Test Auténticos:** S1 y S2, M1, usuario i
   - **Test Impostores:** S1 y S2, M1, usuario j ≠ i

Utilizando todas las posibilidades se haría una especie de validación cruzada. Sin embargo, no es una situación realista en biometría, donde la manera de actuar es utilizar la primera opción del procedimiento experimental mencionado, que utiliza como entrenamiento la primera sesión, intentando simular el comportamiento de la vida real. Interpretando que los datos se utilizan en orden y que la primera sesión, S1, y muestra, M1, que se obtiene es la que forma parte del conjunto de *train*, la que se usa para lo que en biometría se denomina *inscribir* al usuario. Por ejemplo, si el acceso a una oficina de trabajo funciona a través de la huella dactilar, el primer día se recoge la muestra del nuevo usuario y esa es la que se utiliza como entrenamiento para los accesos en días posteriores.

Sin embargo, en este proyecto, también se utilizará clasificación supervisada para tratar de mejorar los resultados obtenidos con la clasificación no supervisada. Se tratará de clasificación supervisada binaria con dos clases: usuario auténtico y usuario impostor. De esta manera, el conjunto de entrenamiento (train) de los datos cambia para incluir además de los datos del usuario auténtico, datos de otros usuarios impostores con los que será entrenado el clasificador y construido el modelo de aprendizaje. Ahora, en clasificación supervisada, los escenarios experimentales, para cada usuario *i*, considerarán el conjunto de datos de **train** mostrado en la clasificación no supervisada como **train auténtico**, y como **train impostores**, los datos de todos los usuarios adquiridos en [4], no utilizando en ningún momento datos de entrenamiento como prueba.

5.6. Clasificación

Como partida para analizar los parámetros del modelo y poder sentar unas bases sólidas dentro de esta biometría [57], se utiliza un algoritmo de coincidencia que transforma los vectores de características lo menos posible. De los propuestos en la literatura, K-vecinos más próximo (K-NN) es uno de los más utilizados, logrando rendimientos similares a alternativas más complejas [48, 88]. Entre las pruebas realizadas, $K = 1$ muestra el mejor rendimiento y es el valor de K que se va a utilizar siempre que se mencione este clasificador. Su funcionamiento es el siguiente:

1. **Inscripción:** De las muestras del usuario C en el corpus, se seleccionan una o dos (dependiendo del escenario probado) para construir su plantilla, λ_C. La plantilla está construida de la siguiente manera:
   - La muestra se divide en ventanas.
   - De cada ventana, W_i, se extrae el vector de características en el dominio de la frecuencia o en el dominio del tiempo, F_i^e; utilizamos el superíndice para diferenciar los vectores de características en las fases de inscripción (superíndice e) y autenticación (superíndice a).
   - El conjunto de todos los vectores de características extraídos forma la plantilla de usuario, $\lambda_C = \{F_i^e\}\ 0 \leq i \leq N$, siendo N el número de ventanas de las muestras de inscripción .
2. **Autenticación:** En primer lugar, se divide en ventanas la muestra de prueba. De cada ventana, W_j, se extrae el vector de características correspondiente, F_j^a. Su score, es decir, la salida de la etapa de comparación, se calcula como se muestra en (5.6).

$$s(F_j^a/\lambda_C) = min_i(EuclideanDistance(F_j^a, F_i^e)) \tag{5.6}$$

Con el objetivo de mejorar los resultados, se van a utilizar algoritmos más complejos de clasificación supervisada. En todos ellos, la fase de inscripción será la misma que se acaba de describir con K-vecinos más próximos, pero construyendo la plantilla del usuario con vectores característicos tanto suyos, considerados como clase "auténtica" como de usuarios impostores considerados como clase "impostora" (clasificación supervisada binaria). En la fase de autenticación, se continúa dividiendo en ventanas la muestra de prueba, extrayendo de cada ventana, W_j, el vector de características correspondiente, F_j^a y obteniendo un score como

salida del clasificador, el cual se calculará de manera específica en cada algoritmo. Los clasificadores probados en este trabajo son:

- **Árboles de decisión:** es una estructura similar a un diagrama de flujo donde un nodo interno representa una característica, la rama representa una regla de decisión y cada nodo hoja representa el resultado/valor. Su funcionamiento es el siguiente:

  1. Selecciona la característica más significativa utilizando medidas de selección de características (ganancia de información, entropía, índice de Gini) para dividir las ventanas.
  2. Esa característica se convierte en nodo de decisión y divide el conjunto de ventanas en subconjuntos más pequeños de manera que se cumpla una regla de decisión.
  3. Realiza el paso anterior recursivamente hasta que se cumpla una de las siguientes condiciones: todas las ventanas pertenecen al mismo valor de la característica, no quedan más características o no hay más ventanas.

- **Bagging:** es un método híbrido que va construyendo más de un árbol de decisión haciendo repetidamente remuestreo de los datos de entrenamiento con sustitución, y votando los árboles para hallar una predicción de consenso. Se aplica de la siguiente forma [69]:

  1. Genera B conjuntos de pseudo-entrenamiento a partir de la ventana de entrenamiento original.
  2. Entrena un árbol con cada una de las B ventanas del paso 1. Cada árbol se crea sin apenas restricciones y no se somete a poda, por lo que tiene varianza alta pero poco bias. En la mayoría de los casos, la única regla de parada es el número mínimo de ventanas que deben tener los nodos terminales. El valor óptimo de este hiperparámetro puede obtenerse comparando el "out of bag error" que puede interpretarse de la misma forma que un error de validación cruzada.
  3. Para cada nueva ventana, obtiene la predicción de cada uno de los B árboles. El valor final de la predicción se obtiene como la media de las B predicciones en el caso de variables cuantitativas o como la clase predicha más frecuente (moda) para las variables cualitativas.

- **Random Forest:** utiliza una serie de árboles de decisión, con el fin de mejorar la tasa de clasificación. Es una modificación del proceso de bagging que suele conseguir mejores resultados gracias a que decorrelaciona los árboles generados en el proceso. Cada árbol se construye de la siguiente manera [5]:

  1. Dado que el número de casos en el conjunto de entrenamiento es N, una ventana de esos N casos se toma aleatoriamente con reemplazo. Esta ventana será el conjunto de entrenamiento para construir el árbol i.
  2. Si existen M variables de entrada (características), un número m<M se especifica tal que, para cada nodo, m características se seleccionan aleatoriamente de M. La mejor división de estas m características es usada para ramificar el árbol. El valor m se mantiene constante durante la generación de todo el bosque.
  3. Cada árbol crece hasta su máxima extensión posible y no hay proceso de poda.
  4. Nuevas ventanas se predicen a partir de la agregación de las predicciones de los x árboles (i.e., mayoría de votos para clasificación que es lo que aquí nos interesa, promedio para regresión).

- **Reglas RIPPER:** va explorando cada clase desde la menos frecuente a la más frecuente y creando un conjunto de reglas. A diferencia de los árboles de decisión, es un procedimiento iterativo, no jerárquico, más robusto, que tiende a generar reglas más simples [60]. Repite los siguientes pasos [25] hasta que la longitud de descripción del conjunto de reglas alcance los umbrales que se le especifiquen, no haya ventanas pertenecientes a la clase positiva o la tasa de error sea superior al 50 %.

  - Fase de crecimiento: hace crecer una regla agregando antecedentes (o condiciones) a la regla hasta que sea perfecta (es decir, 100 % precisa). El procedimiento prueba todos los valores posibles de cada característica y selecciona la condición con mayor ganancia de información
  - Fase de poda: poda gradualmente cada regla.

  Después de generar el conjunto de reglas inicial R_i, existe una etapa de optimización, que genera y poda dos variantes de cada regla a partir de datos aleatorios usados en las fases de crecimiento y poda.

Una variante se genera a partir de una regla vacía, mientras que la otra se genera agregando antecedentes a la regla original. Por último, elimina las reglas del conjunto de reglas que aumentarían la longitud de descripción de la lista de todo el conjunto de reglas si estuviera en él.

- **Naive Bayes:** se basa en el Teorema de Bayes. Supone que una característica particular en una clase es independiente de otras características. El procedimiento sería el siguiente.

  1. Calcula la probabilidad previa para las etiquetas de las clases, aquí auténtico e impostor.
  2. Determina la probabilidad con cada característica para cada clase.
  3. Calcula la probabilidad posterior con el Teorema de Bayes.
  4. La ventana pertenece a la clase con la probabilidad más alta.

- **Máquinas de Vectores Soporte (SVM):** funciona correlacionando datos a un espacio de características de grandes dimensiones de forma que los puntos de datos se puedan categorizar, incluso si los datos no se puedan separar linealmente de otro modo. Aquí los datos son conjuntos de ventanas. Se detecta un separador entre las clases y los datos se transforman de forma que el separador se puede extraer como un hiperplano. Tras ello, las características de los nuevos datos se pueden utilizar para predecir la clase a la que pertenece una nueva ventana. La función matemática utilizada para la construcción del hiperplano se conoce como función kernel. Las más conocidas son: lineal, polinómico, función de base radial y sigmoide.

- **Perceptrón Multicapa:** es una red neuronal artificial (ANN) formada por múltiples capas, de tal manera que tiene capacidad para resolver problemas que no son linealmente separables. Está compuesto por una capa de entrada, una capa de salida y n capas ocultas entremedias, tal y como se puede ver en la figura 5.4. Se caracteriza por tener salidas disjuntas pero relacionadas entre sí, de tal manera que la salida de una neurona es la entrada de la siguiente. Se diferencian dos fases:

  - **Propagación:** calcula el resultado de salida de la red desde los valores de entrada hacia delante.

- **Aprendizaje:** los errores obtenidos a la salida del perceptrón se van propagando hacia atrás (backpropagation) con el objetivo de modificar los pesos de las conexiones para que el valor estimado de la red se asemeje cada vez más al real, realizando la aproximación mediante la función gradiente del error.

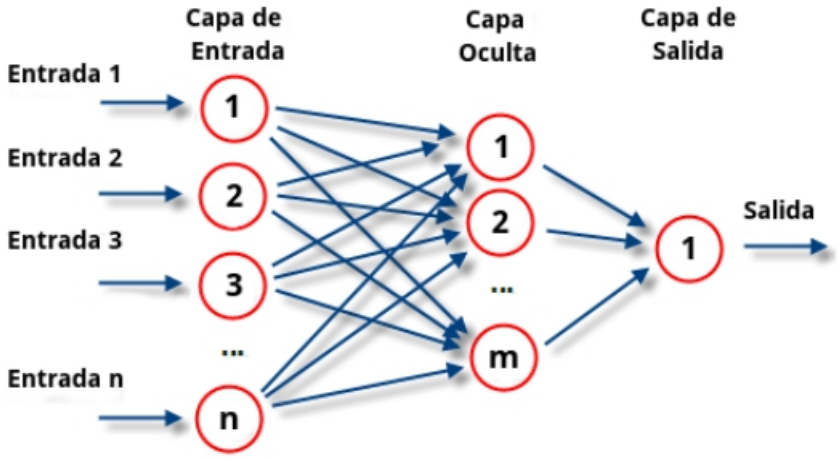


Figura 5.4: Esquema del funcionamiento del algoritmo perceptrón multicapa.

Capítulo 6

Pruebas experimentales

Utilizando el trabajo realizado, tanto en este proyecto, tras la adquisición y el análisis de nuevos datos, como en previos [4, 57], se va a continuar explorando el problema y buscando alternativas que consigan un sistema eficiente de reconocimiento biométrico.

En este capítulo se describen las pruebas realizadas, en el siguiente, se mostrarán los resultados.

6.1. Situación inicial. Comparativa de bases de datos.

En [57] se realizó el análisis de diversos parámetros hasta obtener, entre todas las posibilidades probadas, el mejor valor de cada uno de ellos, lo cual llevó a sentar unas bases sólidas para esta biometría y a construir un sistema final cuyos parámetros se han explicado en los apartados 5.1 y 5.2 y cuyos valores se muestran a continuación.

- **Preprocesamiento:** en el dominio del tiempo funciona mejor utilizar los datos originales filtrados, sin normalizar. El filtro aplicado, tal y como se explica en la sección 5.1 es el de la media móvil ponderada de orden 3. En el dominio de la frecuencia y para tener un significado físico de las componentes extraídas del Análisis de Fourier, se utilizan los datos interpolados linealmente sin normalizar las amplitudes, asegurándonos de esta manera, que los datos tienen una frecuencia de muestreo fija. De todas formas, la decisión de normalizar e interpolar los datos en cada dominio se deja con cierta incertidumbre, ya que no se obtienen resultados suficientemente concluyentes.

- **Tamaño de las ventanas/número de ciclos:** en este caso, los resultados son claros, en el estudio de un tamaño apropiado frente a utilizar el más pequeño, de 2, o el más grande, de 15, resulta mejor utilizar un tamaño intermedio de 8, o incluso mejor, una fusión de los resultados obtenidos con los tamaños entre 6 y 10 y el estadístico de la media.
- **Calidad de la señal:** en [57], siguiendo la premisa de que, cuanto mayor es la autocorrelación de una ventana, mejor será el patrón incluido en ella, se toma la decisión de eliminar el ruido haciendo una segmentación automática y eliminando aquellas ventanas con autocorrelación inferior a 3/4 del valor de máxima autocorrelación en valor absoluto de cada usuario/sesión/muestra y componente. Se tendrá en cuenta que, en caso de perder usuarios, el criterio se irá reduciendo en 5 centésimas de forma sucesiva hasta que no se pierdan. De todas formas, se hicieron pocas pruebas y la limpieza de la señal automática se adoptó de manera temporal. En este trabajo, se ha profundizado más en ello y sobre los nuevos datos, se aplicará el sistema de limpieza explicado en 4.3, que incluye la eliminación de ventanas con baja autocorrelación.
- **Fusión de varias ventanas:** esta propuesta original tiene tres parámetros que fueron explicados en el apartado 5.2, tomando la decisión de aplicar tamaño de ventana 4 scores, solapamiento 2 y estadístico de la mediana. Se obtiene la conclusión clara de que la fusión de ventanas funciona, y que es mejor tener muchos datos de cada usuario para poder fusionar sus ventanas, ya que cuanto mayor es la fusión, mejores son los resultados.

El funcionamiento del sistema final, desde que se adquirieron los datos crudos en cada usuario, hasta que se obtiene su tasa de error, se puede ver en la figura 6.1.

Se pretende aplicar el mismo sistema a las dos bases de datos para comparar los resultados. Se hará sobre las características extraídas tanto en el dominio del tiempo como en el dominio de la frecuencia, en los dos dispositivos (Micro y Moto) y en sus dos sensores (ACC y GYR).

6.2. Análisis de estabilidad de la señal

Las conclusiones alcanzadas con los experimentos realizados siempre han llevado a que, dependiendo del usuario, unas veces funciona mejor una

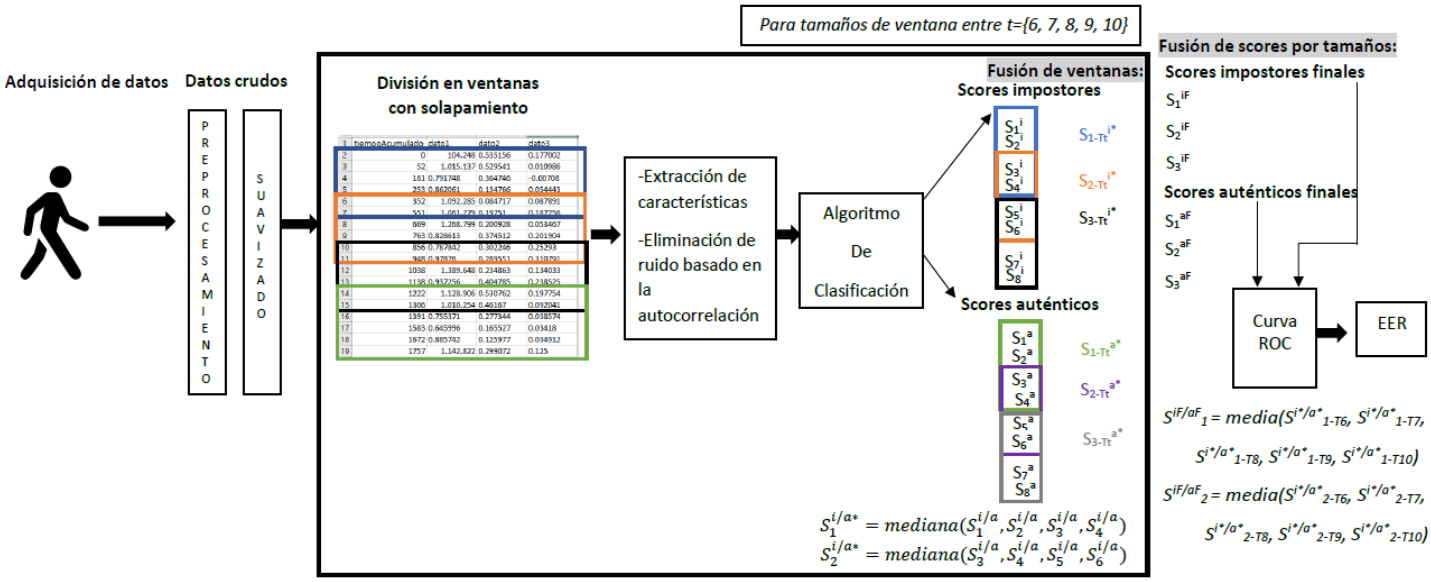


Figura 6.1: Esquema del funcionamiento del sistema final del trabajo [57] tomado como partida.

componente de los datos (X/Y/Z/módulo) y un sensor (ACC/GYR) u otro. Pero siempre se ha trabajado con los mismos escenarios experimentales (MonoMono_M1 y MultiMono_M1 explicados en el apartado 5.5) debido a que, en biometría, no se hacen pruebas cruzadas y los sistemas se entrenan con una sola muestra, la primera.

Pero para que la construcción de un sistema de predicción característico de cada usuario tenga sentido tiene que existir estabilidad en la señal, es decir, si se intercambia la muestra de datos en un usuario, las componentes características se mantengan iguales. Por ello, se va a realizar un análisis de la señal a corto y largo plazo a través de los siguientes escenarios experimentales.

- **Escenarios experimentales a corto plazo** (Monosesión-Monomuestra, MonoMono): compara los datos dentro de la misma sesión. Es el escenario más favorable que simula el momento temporal "a corto plazo" porque las muestras usadas para entrenamiento y para *prueba auténtica* del usuario son tomadas en la misma sesión. Como se tienen 2 sesiones de datos y 2 muestras por sesión, existen 4 divisiones de los datos, que se corresponden con MonoMono_M1, MonoMono_M2, MonoMono_M3 y MonoMono_M4, explicadas en el apartado 5.5.

- **Escenarios experimentales a largo plazo** (Multisesión-Monomuestra, MultiMono): las muestras usadas para entrenamiento y *prueba auténtica* son tomadas en distintas sesiones. Permite probar la variabilidad del rasgo biométrico con el tiempo, simulando el momento temporal "a largo plazo". Corresponde a las divisiones de datos MultiMono_M1, MultiMono_M2, MultiMono_M3, MultiMono_M4 explicadas en el apartado 5.5.

6.3. Aplicación de clasificadores

Una alternativa a la construcción de un sistema de predicción característico de cada usuario, con el objetivo de mejorar los resultados, es la aplicación de diferentes clasificadores, más complejos, sobre los datos. De manera que se encuentre un algoritmo eficiente, capaz de detectar el patrón de cada usuario.

Se va a seguir un modelo de clasificación binaria supervisada con clases auténtico e impostor y a probar los clasificadores explicados en el apartado 5.6. Un problema en estos datos es que el conjunto de ventanas auténticas tiene una representación mucho menor que las impostoras (mayoritarias). Esto se conoce con el nombre de *desbalanceo de clases* y no es algo nuevo en Machine Learning. Por el contrario, se sabe desde hace mucho tiempo que afecta de forma considerable a diferentes clasificadores, habiéndose desarrollado técnicas para tratar con este problema. Estas técnicas se basan en dos aspectos fundamentales [16, 51]:

- **Agregar patrones (oversampling o sobremuestreo):** se realiza sobre las clases minoritarias tratando de emparejar el número de patrones con las clases mayoritarias.
- **Eliminar patrones (undersampling o submuestreo):** eliminar patrones de las clases mayoritarias con la finalidad de que el clasificador pueda entrenarse con la misma cantidad de patrones en cada clase.

Las reglas básicas para hacer sobremuestreo o submuestreo son las siguientes:

- **Al azar:** agregar/quitar patrones de forma aleatoria.
- **Duplicados:** realizar copias de patrones dentro de su misma clase para incrementar el número total o eliminar patrones repetidos dentro de una misma clase.

- **Patrones cercanos:** muy similar al funcionamiento del clasificador KNN, buscando los vecinos más cercanos para poder incrementar o eliminar patrones.

Para solucionar este problema, se va a utilizar la función "ovun.sample" del paquete "ROSE" en R [73], el cual ya está desarrollado y probado.

- **Oversampling:** duplica los datos de la clase minoritaria hasta tener la misma cantidad de ventanas que la clase mayoritaria. Se hará referencia con la palabra *over*.
- **Undersampling:** elimina ventanas de la clase mayoritaria hasta tener la misma cantidad que la clase minoritaria. Se hará referencia con la palabra *under*.
- **Balanced:** es una combinación de las dos técnicas anteriores. Duplica datos en la clase minoritaria y los elimina en la mayoritaria hasta conseguir un equilibrio, cuando la probabilidad de volver a muestrear la clase minoritaria (auténtica) es del 50 %. Se hará referencia con la palabra *balanced*.
- También se va a probar cada clasificador con los datos tal cual están, no balanceados. Se hará referencia con la palabra *ninguna*.

En ningún momento se van a crear muestras artificiales (no idénticas) porque no se quiere alterar la distribución "natural" de las clases y confundir al modelo en su clasificación. Esto, además, se desaconseja, en general, en biometría.

Se desea destacar que se ha probado el clasificador *perceptrón multicapa*, pero no se van a incluir sus resultados debido a que presenta mucha variabilidad dentro de una misma ejecución, dependiendo de sus parámetros de inicialización.

6.4. Profundizar en el clasificador SVM

Una vez probado un primer subconjunto de clasificadores más simples y rápidos como los árboles de decisión, algoritmos basados en reglas como RIPPER y basados en probabilidades como Naive Bayes, también otros más complejos como bagging, Random Forest y máquinas de vectores soporte (SVM), se ha decidido profundizar en aquel que genera los mejores

resultados, como se verá más adelante: SVM. Todos ellos han sido probados con los valores por defecto, a modo experimentación.

Random Forest, como se mostrará en el capítulo correspondiente, también genera buenos resultados. La decisión de profundizar en SVM se debe a que tiene una mayor cantidad de parámetros que se pueden configurar y adaptar para conseguir mejorar los resultados. Además, genera un modelo final de menor tamaño, lo que es importante pensando en la escalabilidad de la solución y respecto a tiempo de ejecución son similares.

Por defecto, se ha aplicado un kernel lineal, es decir, la frontera de división entre los datos en la dimensión original será hiperplano. Siguiendo la referencia [68], un hiperplano se define como un subespacio plano y afín, es decir, el subespacio no tiene por qué pasar por el origen. Dados los parámetros $\beta_0, \beta_1, \beta_2$, todos los pares de valores $x = (x_1, x_2)$ para los que se cumple la igualdad son puntos del hiperplano, cuya definición matemática se define con la ecuación 6.1. El punto x cae a un lado o al otro del hiperplano. Así pues, se puede entender que un hiperplano divide un espacio p-dimensional en dos mitades [68, 66, 43].

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0 \tag{6.1}$$

La solución a este problema consiste en seleccionar como clasificador óptimo, *maximal margin hyperplane* o hiperplano óptimo de separación, al hiperplano que se encuentra más alejado de todas las observaciones de entrenamiento. Para obtenerlo, se tiene que calcular la distancia perpendicular de cada observación a un determinado hiperplano. La menor de estas distancias (conocida como margen) determina cómo de alejado está el hiperplano de las observaciones de entrenamiento. El *maximal margin hyperplane* se define como el hiperplano que consigue un mayor margen, es decir, que la distancia mínima entre el hiperplano y las observaciones es lo más grande posible.

La figura 6.2 muestra el *maximal margin hyperplane* para un conjunto de datos de entrenamiento. Las tres observaciones equidistantes respecto al hiperplano óptimo de separación se encuentran a lo largo de las líneas discontinuas que indican la anchura del margen. A estas observaciones se les conoce como vectores soporte, ya que son vectores en un espacio p-dimensional y definen el *maximal margin hyperplane*. Cualquier modificación en estas observaciones (vectores soporte) conlleva cambios en el *maximal margin hyperplane*. Sin embargo, modificaciones en observaciones

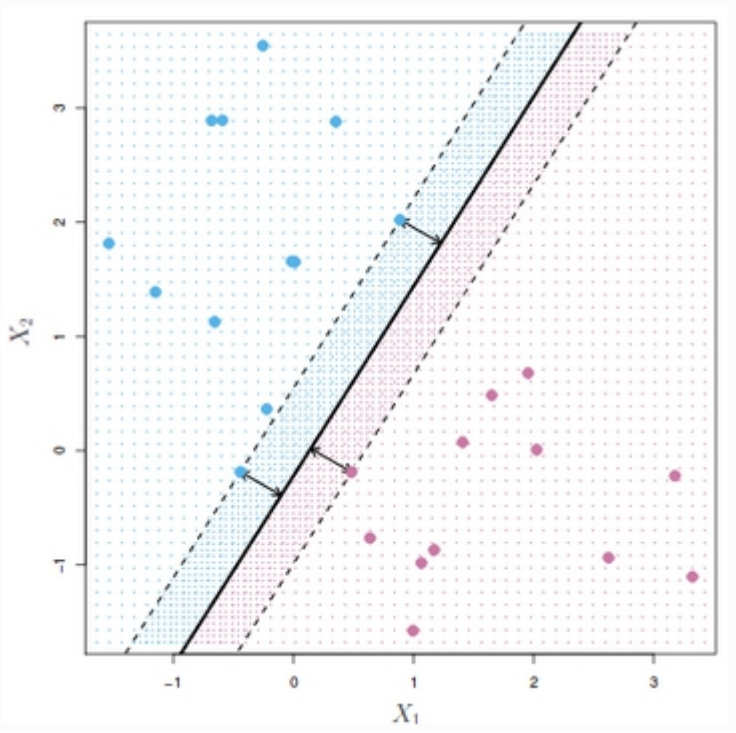


Figura 6.2: Hiperplano óptimo de separación en el clasificador Máquinas de Vectores Soporte (SVM), extraída de [68].

que no son vector soporte no tienen impacto alguno en el hiperplano. Esto hace que sea un clasificador menos afectado por el desbalanceo de las clases.

Sin embargo, rara vez se encuentran datos en los que las clases sean perfecta y linealmente separables. De hecho, incluso cumpliéndose estas condiciones ideales, en las que exista un hiperplano capaz de separar perfectamente las observaciones en dos clases, esta aproximación sigue presentando dos inconvenientes [68]:

- **Poca robustez:** dado que el hiperplano tiene que separar perfectamente las observaciones, es muy sensible a variaciones en los datos. Incluir una nueva observación puede suponer cambios muy grandes en el hiperplano de separación.
- **Sobreajuste:** que el maximal margin hyperplane se ajuste perfectamente a las observaciones de entrenamiento para separarlas todas correctamente suele conllevar problemas de sobreajuste (overfitting).

Por estas razones, es preferible crear un clasificador basado en un hiperplano que, aunque no separe perfectamente las dos clases, sea más robusto y tenga mayor capacidad predictiva al aplicarlo a nuevas observaciones (menos problemas de overfitting). Esto es exactamente lo que consiguen los clasificadores de vector soporte, también conocidos como soft margin classifiers o Support Vector Classifiers. Para lograrlo, en lugar de buscar el margen de clasificación más ancho posible que consigue que las observaciones estén en el lado correcto del margen; se permite que ciertas observaciones estén en el lado incorrecto del margen o incluso del hiperplano, incluyendo un hiperparámetro, llamado coste o C. C controla el número y severidad de las violaciones del margen (y del hiperplano) que se toleran en el proceso de ajuste. Si $C = \infty$, no se permite ninguna violación del margen y, por lo tanto, el resultado es equivalente al Maximal Margin Classifier. Cuanto más se aproxima C a cero, menos se penalizan los errores y más observaciones pueden estar en el lado incorrecto del margen o incluso del hiperplano. C es a fin de cuentas el hiperparámetro encargado de controlar el balance entre bias y varianza del modelo. Se va a utilizar, en todos los casos, un coste de 15 unidades, obtenido mediante validación cruzada en múltiples pruebas.

El kernel lineal consigue buenos resultados cuando el límite de separación entre clases es aproximadamente lineal. Si no lo es, su capacidad decae drásticamente. Una estrategia para enfrentarse a escenarios en los que la separación de los grupos es de tipo no lineal consiste en expandir las dimensiones del espacio original. Este parámetro se conoce como kernel. Un kernel (K) es una función que devuelve el resultado del producto escalar *(dot product)* entre dos vectores realizado en un nuevo espacio dimensional distinto al espacio original en el que se encuentran los vectores. Los más utilizados y los que se van a probar son los siguientes.

- **Kernel lineal:** como ya se ha explicado.

$$K(x, x') = x \cdot x' \tag{6.2}$$

- **Kernel polinómico:** Cuando se emplea $d = 1$ y $c = 0$, el resultado es el mismo que el de un kernel lineal. Si $d > 1$, se generan límites de decisión no lineales, aumentando la no linealidad a medida que aumenta d. No suele ser recomendable emplear valores de d mayores que 5 por problemas de overfitting. Se van a probar como valores de d 3, 4 y 5 con $c = 0$.

$$K(x, x') = (x \cdot x' + c)^d \tag{6.3}$$

- **Kernel radial o gaussiano (RBF):** el valor de γ controla el comportamiento del kernel. Cuando es muy pequeño, el modelo final es equivalente al obtenido con un kernel lineal, a medida que aumenta su valor, también lo hace la flexibilidad del modelo. Este kernel tiene dos ventajas: que solo tiene dos hiperparámetros que optimizar: γ y la penalización C común a todos los SVM, y que su flexibilidad puede ir desde un clasificador lineal a uno muy complejo. Aquí se va a utilizar el valor de gamma 2, obtenido mediante validación cruzada en múltiples pruebas.

$$K(x, x^{'}) = exp(-\gamma||x - x^{'}||^2) \tag{6.4}$$

- **Kernel sigmoide:** es una variación del kernel que se aproxima al funcionamiento de una red neuronal. Se utiliza valor de gamma 2 y $c = 0$.

$$K(x, x^{'}) = tanh(\gamma x \cdot x^{'} + c) \tag{6.5}$$

6.5. Sistema final

Una vez probadas las alternativas indicadas en los apartados anteriores sobre preprocesamiento de lo datos, balanceo de datos en entrenamiento y pruebas de distintos clasificadores, se obtendrá una configuración final del sistema. Como el objetivo es lograr la mejor clasificación posible, se intentará mejorar aún más el rendimiento de ese sistema final mediante las siguientes pruebas:

- **Fusión de varias ventanas:** parámetro explicado en el apartado 5.2. Para las pruebas iniciales se va a usar la configuración del trabajo anterior [57] en el que se fusionan los scores de n=4 ventanas consecutivas con solapamiento 2. Aquí vamos a incrementar el número n, de ventanas a solapar, para ver si con esto mejoramos la caracterización del usuario.

- **Fusión total:** como continuación de la mejora anterior, se propone fusionar todas las ventanas de las muestras de tests en sólo una. Es decir, una vez obtenidos todos los scores, fusionarlos en sólo uno a través de su mediana.

- **Fusión de scores por dominios:** se aplica el sistema de manera independiente para las características de cada usuario en el dominio del tiempo y de la frecuencia. Aquí se propone que, una vez obtenidas, se fusionen a través de los estadísticos: suma, media o producto.

Para verificar la idoneidad del sistema construido, se van a realizar las pruebas en las dos bases de datos disponibles, lo que equivale a un total de 32 usuarios (12 de la base de datos del trabajo [4] y 20 adquiridos en el presente proyecto).

6.6. Influencia de la mano en el uso del dispositivo

Hasta este apartado, todas las pruebas realizadas se han hecho con la mano de uso habitual del reloj en cada usuario, esto es, generalmente, la izquierda en un usuario diestro y la derecha en un usuario zurdo. Tratando de simular el escenario más realista, en el que el usuario se encuentra más cómodo. Con la mano de uso habitual del reloj se han realizado 4 recorridos por sesión (2 con cada dispositivo). Pero se adquirieron 2 recorridos más (1 con cada dispositivo), en los que el usuario llevaba el reloj ubicado en la mano opuesta. Se va a llamar M1 al primer recorrido de cada usuario con cada dispositivo, M2 al segundo recorrido con cada dispositivo y M3 al tercer recorrido utilizando la mano no habitual.

Se van a probar dos escenarios. El primero de ellos tiene como objetivo analizar las consecuencias de cambiar de mano, es decir, qué pasa si una vez creado el modelo, el usuario decide ponerse el dispositivo en la otra mano, bien por error o bien por necesidad. En este caso, existirían dos posibilidades:

1. MonoMono_ManoNoHabitual_Escenario1: compara los datos dentro de la misma sesión, es decir, las muestras usadas para entrenamiento y para *prueba auténtica* del usuario son tomadas en la misma sesión. Es el caso más favorable.
   - **Train:** S1, M1, usuario i
   - **Test Auténticos:** S1, M3, usuario i
   - **Test Impostores:** S1, M2, usuario j $\neq$ i
2. MultiMono_ManoNoHabitual_Escenario1: Las muestras usadas para entrenamiento y *prueba auténtica* son tomadas en distintas sesiones, permitiendo probar la variabilidad del rasgo biométrico con el tiempo.
   - **Train:** S1, M1, usuario i

- **Test Auténticos:** S2, M3, usuario i
- **Test Impostores:** S1, M2, usuario j ≠ i

Un segundo escenario en el que el usuario usa siempre la mano no habitual, tanto para crear el modelo como para probarlo (autenticarse). Con el objetivo de comprobar si el movimiento de las manos es simétrico y, por tanto, si influye en el rendimiento del sistema la muñeca en la que el usuario lleva colocado el wearable. En este caso, se tiene una única posibilidad.

1. MultiMono_ManoNoHabitual_Escenario2: Las muestras usadas para entrenamiento y *prueba auténtica* son tomadas en distintas sesiones, permitiendo probar la variabilidad del rasgo biométrico con el tiempo.
   - **Train:** S1, M3, usuario i
   - **Test Auténticos:** S2, M3, usuario i
   - **Test Impostores:** S1, M3, usuario j ≠ i

Como en el apartado 5.5 lo que se especifica como *train* es *train auténtico*, utilizando como *train impostor* los datos de todos los usuarios adquiridos en [4], siendo clasificación supervisada y no utilizando en ningún momento datos de entrenamiento como prueba.

Capítulo 7

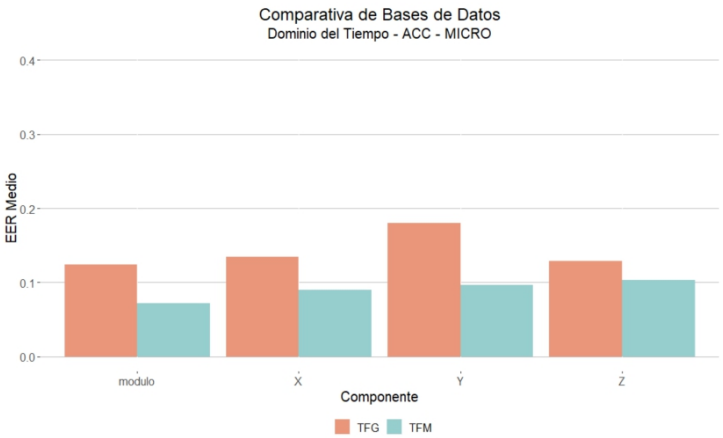
Resultados

Tras haber explicado, en el capítulo anterior, las pruebas realizadas, se van a mostrar aquí los resultados obtenidos.

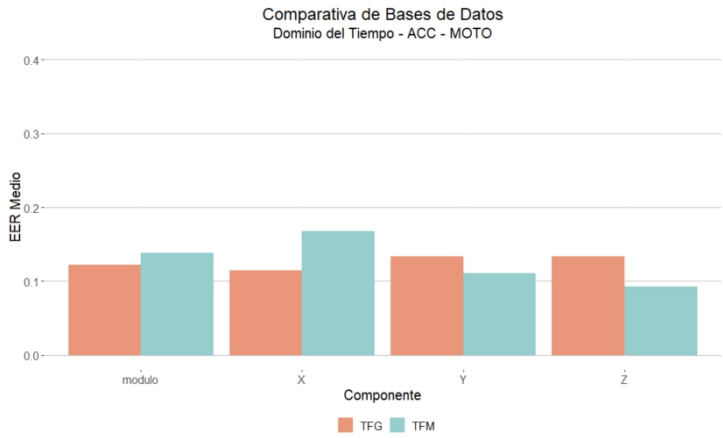
7.1. Situación inicial. Comparativa de bases de datos.

Durante todo el apartado se va a utilizar el escenario experimental *Multisesión-Multimuestra* (MultiMulti_M1) que tiene en cuenta la variabilidad del rasgo biométrico con el tiempo e incluye datos de las dos sesiones, para tener una mayor cantidad, especialmente en la primera base de datos, del TFG, donde los recorridos de los usuarios tienen una menor duración temporal.

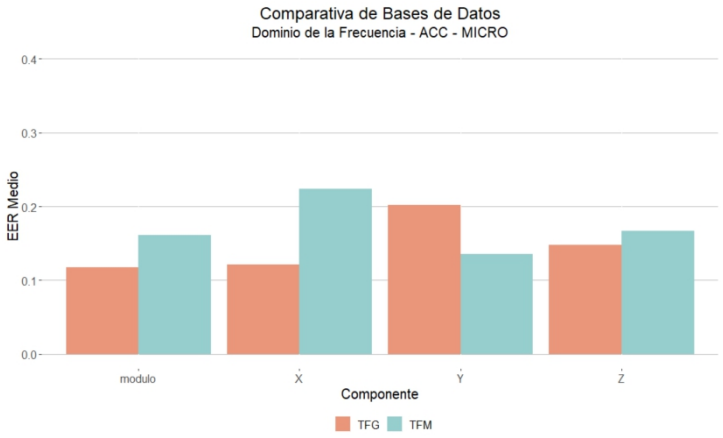
En la figura 7.1 se puede ver una comparación de las tasas de error medias de todos los usuarios en las dos bases de datos, teniendo en cuenta que en el TFG se tenían 14 usuarios y ahora (TFM) se tienen un total de 20. Se están utilizando las características extraídas con el sensor acelerómetro y aplicándolo a los dos dispositivos (MICRO y MOTO). Los nuevos datos (TFM marcados en azul) funcionan ligeramente mejor que los anteriores en el Dominio del Tiempo. Sin embargo, en el Dominio de la Frecuencia ocurre justo lo contrario. Entre dispositivos (MICRO y MOTO en las figuras a y b para el Dominio del Tiempo, y c y d para el Dominio de la Frecuencia) existen pocas diferencias, variando la componente con la que se consiguen los mejores resultados entre dispositivos y bases de datos, aunque en el Dominio del Tiempo con el dispositivo Microsoft se mantiene, en ambas bases de datos, el módulo con errores ligeramente inferiores.



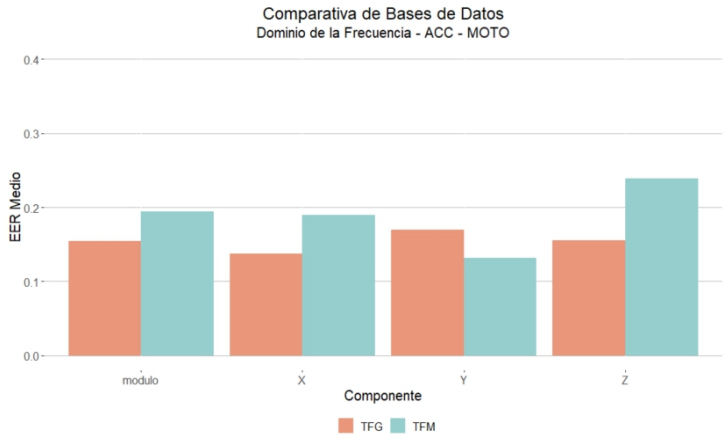
(a) Dominio del Tiempo - MICRO



(b) Dominio del Tiempo - MOTO



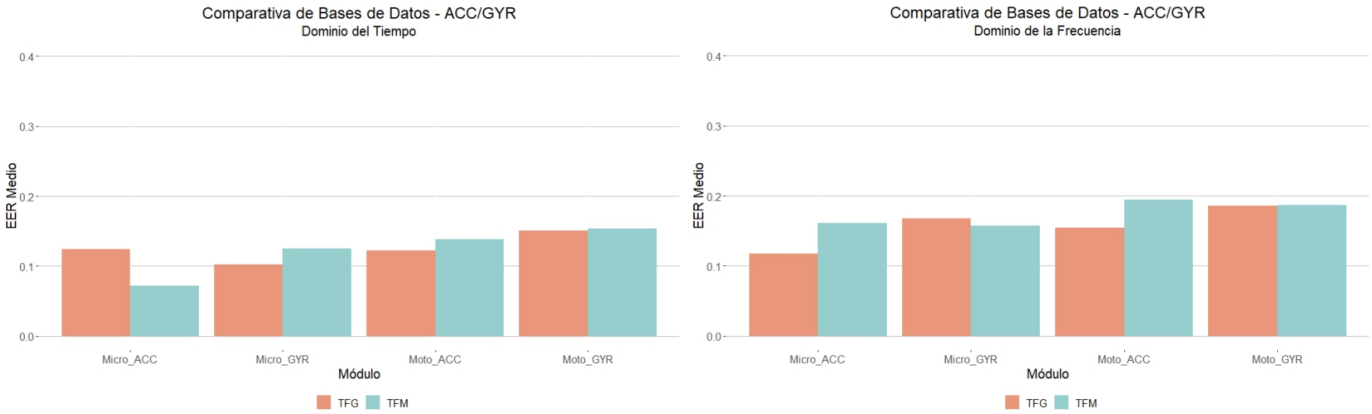
(c) Dominio de la Frecuencia - MICRO



(d) Dominio de la Frecuencia - MOTO

Figura 7.1: Comparación por componentes del sistema de reconocimiento biométrico inicial en las dos Bases de Datos disponibles

En la figura 7.1 se está utilizando únicamente el acelerómetro como sensor. Si se utiliza también el giroscopio y se comparan los resultados medios de todos los usuarios en la componente Módulo, se pueden ver diferencias muy pequeñas (figura 7.2). Por tanto, se puede decir que los resultados son similares en ambas Bases de Datos, lo cual demuestra el correcto funcionamiento del sistema construido con independencia de los datos.



(a) Dominio del Tiempo (b) Dominio de la Frecuencia

Figura 7.2: Comparación por sensores del sistema de reconocimiento biométrico inicial en las dos Bases de Datos disponibles

Con el objetivo de mejorar los resultados, se han probado los siguientes escenarios para ambos dispositivos en los Dominios del Tiempo y de la Frecuencia, utilizando el módulo como componente (figura 7.3).

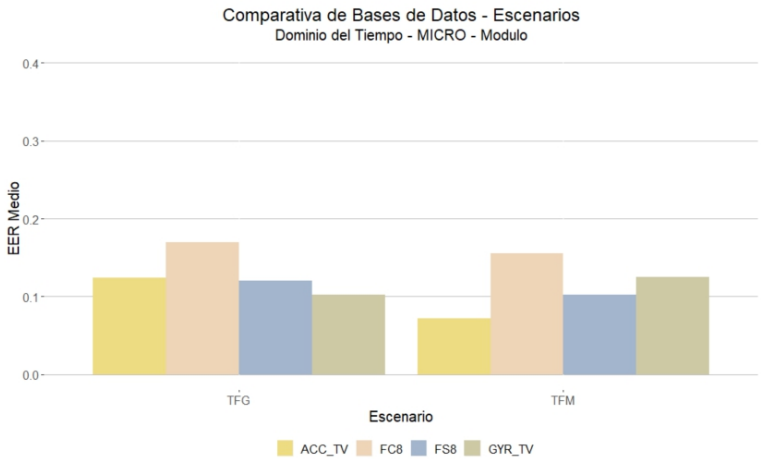
- **Sistema construido en el sensor acelerómetro (ACC_TV):** resultados obtenidos utilizando únicamente las características del sensor acelerómetro.
- **Sistema construido en el sensor giroscopio (GYR_TV):** resultados obtenidos utilizando únicamente las características del sensor giroscopio.
- **Fusión de características con tamaño de ventana 8 (FC8):** en lugar de utilizar un sólo sensor, se ha probado a construir, para cada usuario, ventanas de datos iguales en ambos sensores para poder extraer sus características y utilizarlas de manera conjunta.

- **Fusión de scores con tamaño de ventana 8 (FS8):** en lugar de utilizar las características de ambos sensores, se ha probado a construir sus sistemas de manera independiente. Pero una vez aplicado el clasificador y obtenidos sus scores, se ha tratado de fusionarlos aplicando diversos estadísticos (mínimo, máximo y media, que resulta equivalente a la mediana por ser dos sensores).

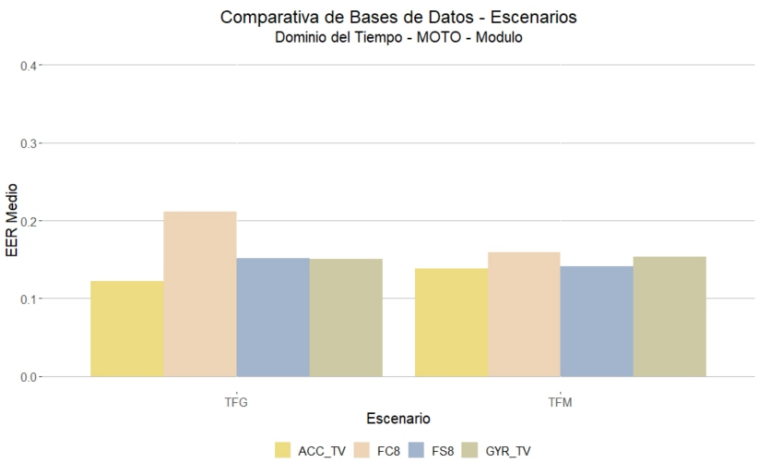
La figura 7.3 muestra que el camino de fusionar características (FC8 en color naranja claro) no merece la pena ya que sus tasas de error son más altas. Esto tiene sentido debido a que se duplica el número de características utilizadas en el clasificador y para que esta vía tuviera más sentido, habría que centrar esfuerzos en la selección de características. Fusionar scores (FS8 en color morado) también obtiene tasas de error más altas, funcionando mejor la utilización de un sólo sensor, o bien acelerómetro o giroscopio (ACC_TV o GYR_TV, respectivamente). También se puede apreciar como los resultados en ambas bases de datos son muy similares, coincidiendo con las conclusiones anteriores.

Las figuras 7.1, 7.2 y 7.3 muestran los errores medios de todos los usuarios disponibles y, aunque los resultados son parecidos entre bases de datos, dispositivos y dominios, sí se puede apreciar que unas veces funciona mejor una componente u otra y un sensor u otro. Centrándonos en los nuevos datos y en el sensor acelerómetro, en la figura 7.4 se muestra la tasa de error de cada usuario en cada una de las componentes. Se utiliza un gráfico de barras apiladas al 100 % para mostrar el porcentaje relativo de cada una de las componentes y poder ver que, dependiendo del usuario, unas veces funciona mejor una u otra y que quedarse con el módulo no es siempre la mejor opción.

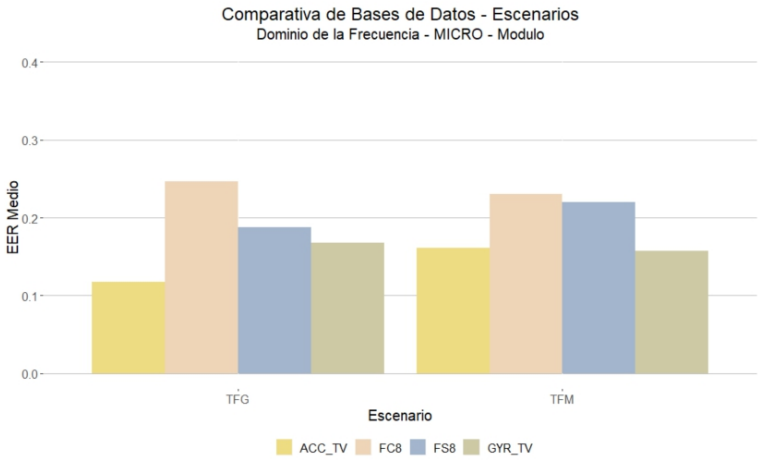
No obstante, quedándonos únicamente con la componente módulo, en la figura 7.5 se puede ver que en el Dominio del Tiempo generalmente funciona mejor el acelerómetro y en el Dominio de la Frecuencia, el giroscopio, pero que dependiendo del usuario unas veces funciona mejor uno u otro. A raíz de los resultados mostrados, parece que cada usuario tiene un componente y/o sensor característico, con el que se obtienen muy buenos resultados. Si eso fuera cierto, y lo pudiéramos predecir de antemano, parece que nos permitiría construir sistemas con un buen rendimiento. En este punto del trabajo nos planteamos esta línea de investigación, pero antes de nada quisimos comprobar la certeza de la premisa de partida, es decir, que cada usuario tiene un sensor y componente que le caracteriza mejor que el resto y con el que se obtienen buenos resultados siempre. Esto se muestra en el siguiente apartado.



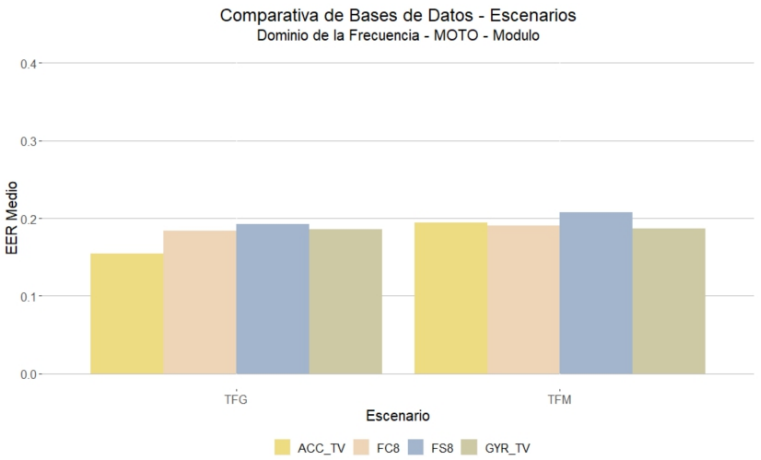
(a) Dominio del Tiempo - MICRO



(b) Dominio del Tiempo - MOTO

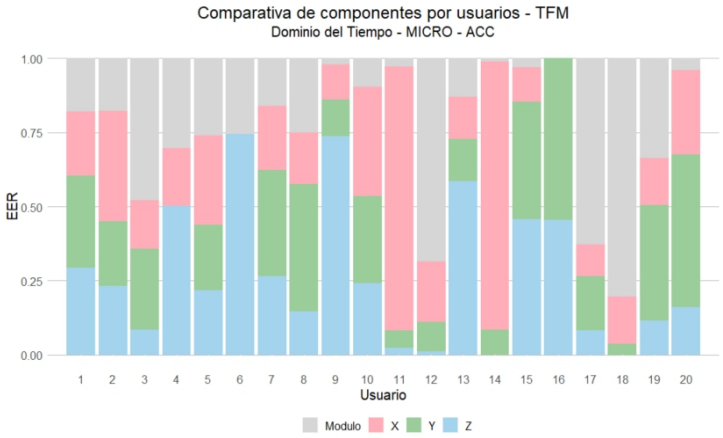


(c) Dominio de la Frecuencia - MICRO

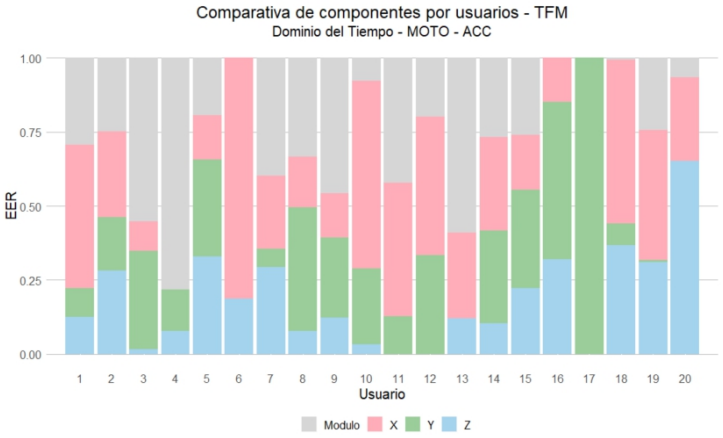


(d) Dominio de la Frecuencia - MOTO

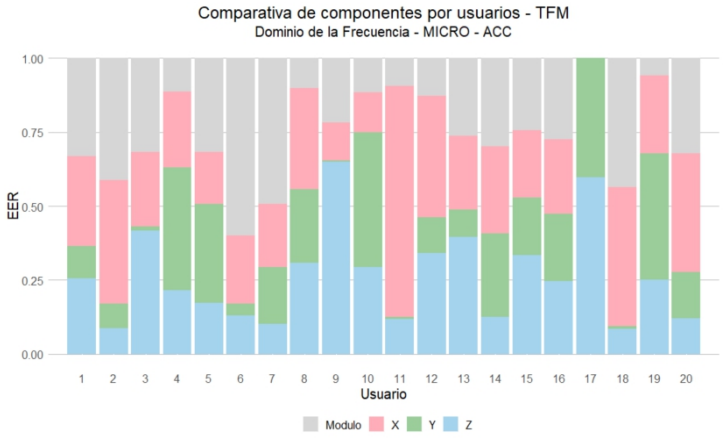
Figura 7.3: Comparación del sistema de reconocimiento biométrico inicial en diversos escenarios de las dos Bases de Datos disponibles



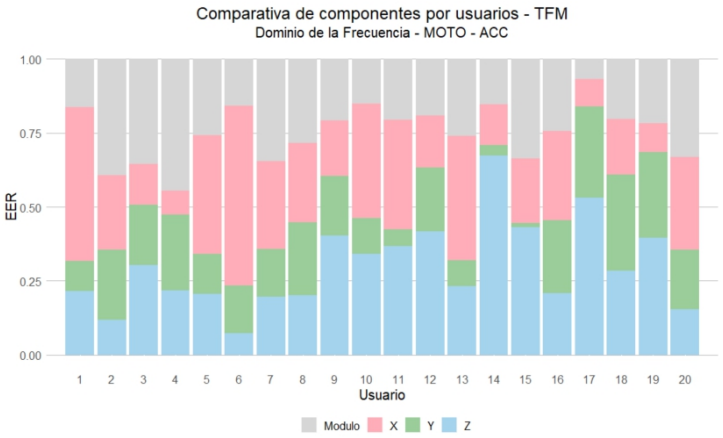
(a) Dominio del Tiempo - MICRO



(b) Dominio del Tiempo - MOTO

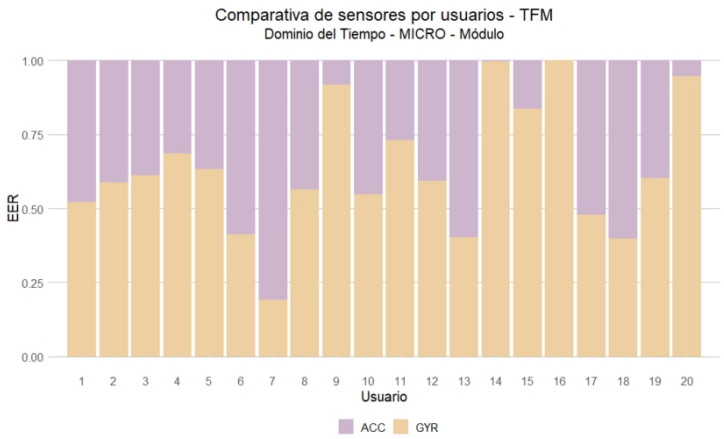


(c) Dominio de la Frecuencia - MICRO

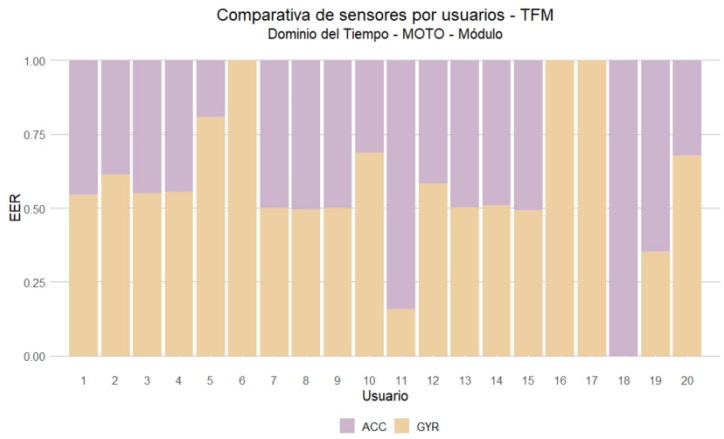


(d) Dominio de la Frecuencia - MOTO

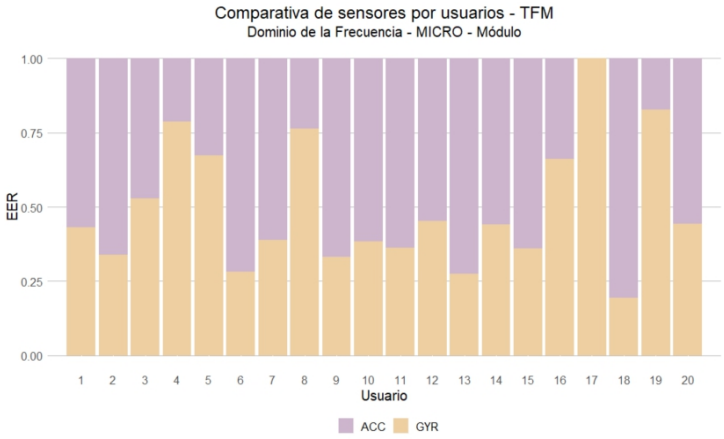
Figura 7.4: Comparación de componentes por usuarios (Nueva Base de Datos). Sistema de reconocimiento biométrico inicial.



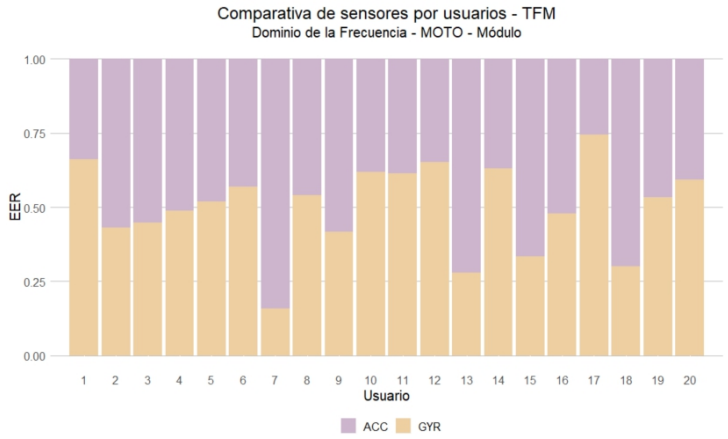
(a) Dominio del Tiempo - MICRO



(b) Dominio del Tiempo - MOTO



(c) Dominio de la Frecuencia - MICRO



(d) Dominio de la Frecuencia - MOTO

Figura 7.5: Comparación de sensores por usuarios (Nueva Base de Datos). Sistema de reconocimiento biométrico inicial.

Dominio	Dispositivo	Sensor	#Usuarios 1 comp.	#Usuarios 2 comp.
DT	Micro	ACC	2	10
DT	Micro	GYR	3	14
DT	Moto	ACC	2	10
DT	Moto	GYR	3	8
DF	Micro	ACC	3	5
DF	Micro	GYR	2	10
DF	Moto	ACC	6	8
DF	Moto	GYR	1	9

Tabla 7.1: Análisis de la estabilidad a corto plazo (Nueva Base de Datos). # Usuarios 1 comp. es el número de usuarios que tienen siempre resultados buenos en la misma componente en los cuatro escenarios. # Usuarios 2 comp. es el número de usuarios cuyos buenos resultados se alternan siempre entre las mismas 2 componentes (de un total de cuatro) en los cuatro escenarios.

7.2. Análisis de estabilidad de la señal

Con los datos capturados, hay disponibles 4 escenarios a corto plazo y 4 a largo plazo. Para que exista estabilidad en el comportamiento del usuario con respecto a su sensor y componente característico, en cada dominio, dispositivo y sensor debe existir un número alto de usuarios en los que coincida la misma predicción en sus cuatro escenarios. Se está utilizando la base de datos nueva, con 20 usuarios. En las tablas 7.1 y 7.2 se indica esta información a corto y largo plazo, respectivamente. Los resultados muestran que, en ambos casos, muy pocos usuarios mantienen un comportamiento estable. Si nos fijamos en aquellos usuarios que mantienen dos coordenadas estables en los 4 escenarios, se obtiene, más o menos, el 50 % de los usuarios.

La figura 7.6 muestra un usuario concreto en el Dominio del Tiempo con el dispositivo Microsoft y el sensor Acelerómetro. Cada columna contiene las tasas de error de cada una de las componentes, correspondiendo las 4 primeras a los escenarios a corto plazo y el resto a largo plazo. Es un ejemplo de usuario en el que tanto a corto como a largo plazo la coordenada Z es una buena opción, aunque no siempre la mejor. La figura 7.7 muestra otro usuario en el que los mejores resultados se alternan entre la coordenada Y y Z. El usuario de la figura 7.8 muestra cómo a corto

Dominio	Dispositivo	Sensor	#Usuarios 1 comp.	#Usuarios 2 comp.
DT	Micro	ACC	0	13
DT	Micro	GYR	4	13
DT	Moto	ACC	0	12
DT	Moto	GYR	1	12
DF	Micro	ACC	2	14
DF	Micro	GYR	3	10
DF	Moto	ACC	7	8
DF	Moto	GYR	3	10

Tabla 7.2: Análisis de la estabilidad a largo plazo (Nueva Base de Datos). # Usuarios 1 comp. es el número de usuarios que tienen siempre resultados buenos en la misma componente en los cuatro escenarios. # Usuarios 2 comp. es el número de usuarios cuyos buenos resultados se alternan siempre entre las mismas 2 componentes (de un total de cuatro) en los cuatro escenarios.

plazo, cada vez funciona mejor una coordenada y el de la figura 7.9 cómo la coordenada Z pasa de ser la peor a la mejor.

En biometría existe un problema, bien conocido desde 1998 como "animalario" [26], que describe que los humanos se pueden clasificar en 4 categorías:

- **Ovejas**: usuarios buenos, fáciles de reconocer.
- **Cabras**: usuarios difíciles de reconocer.
- **Corderos**: usuarios fáciles de imitar, siendo muy probable que una persona elegida al azar sea aceptada como cordero.
- **Lobos**: personas particularmente exitosas en imitar a otras personas, siendo muy probable que su patrón sea aceptado como el de otra persona.

Basándonos en ese animalario, podríamos decir que hay demasiados usuarios cabras y muy poco ovejas, demostrando un comportamiento no estable con respecto a la componente y sensor con mejor rendimiento. Esto nos hizo abandonar la línea de investigación centrada en intentar predecir el sensor y coordenada característicos de cada individuo, para ser usados solo estos en su reconocimiento.

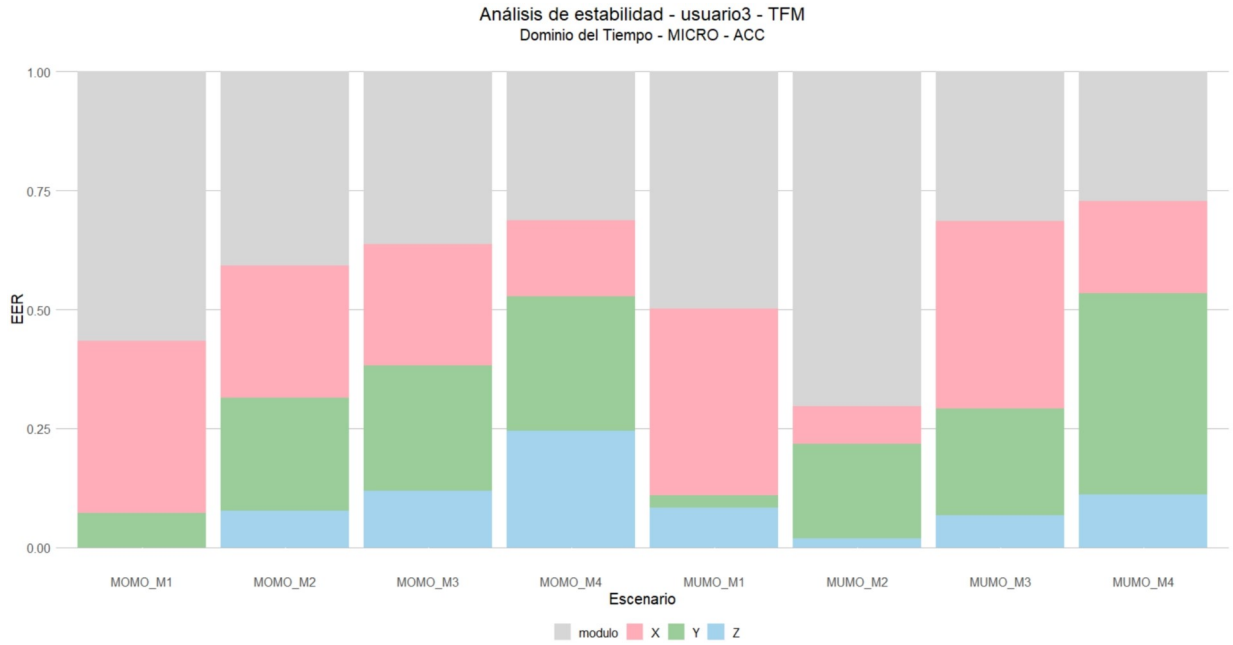


Figura 7.6: Análisis de la estabilidad en el usuario 3 (Dominio del Tiempo, Micro, ACC)

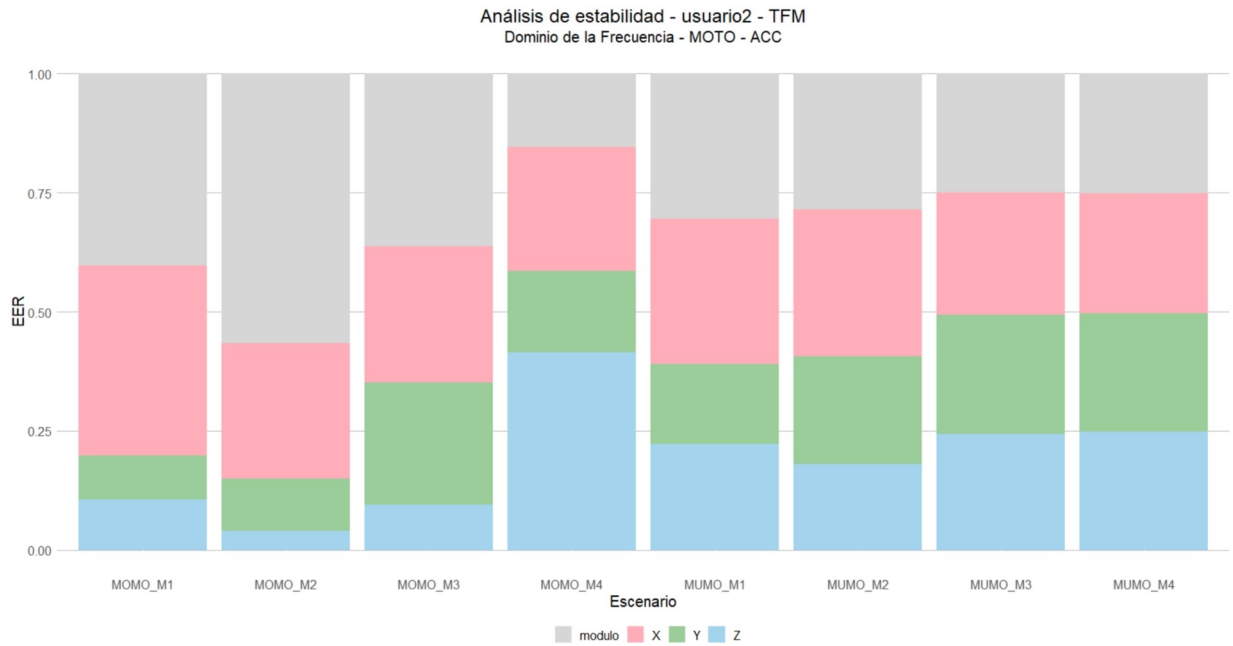


Figura 7.7: Análisis de la estabilidad en el usuario 2 (Dominio de la Frecuencia, Moto, ACC)

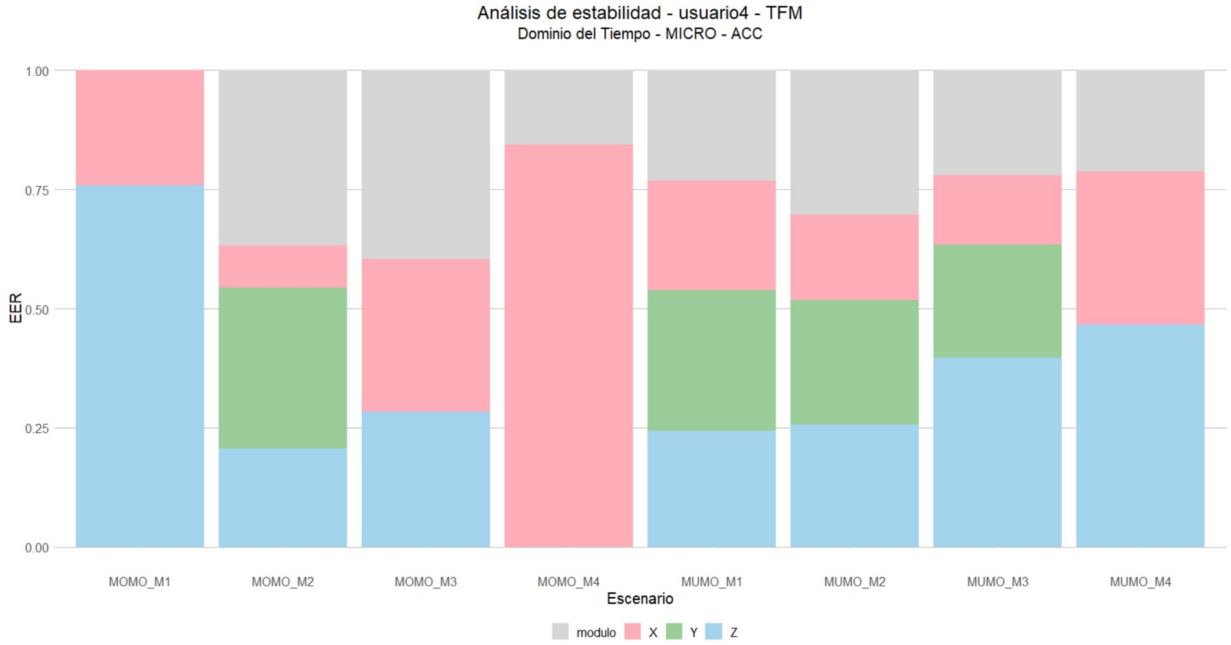


Figura 7.8: Análisis de la estabilidad en el usuario 4 (Dominio del Tiempo, Micro, ACC)

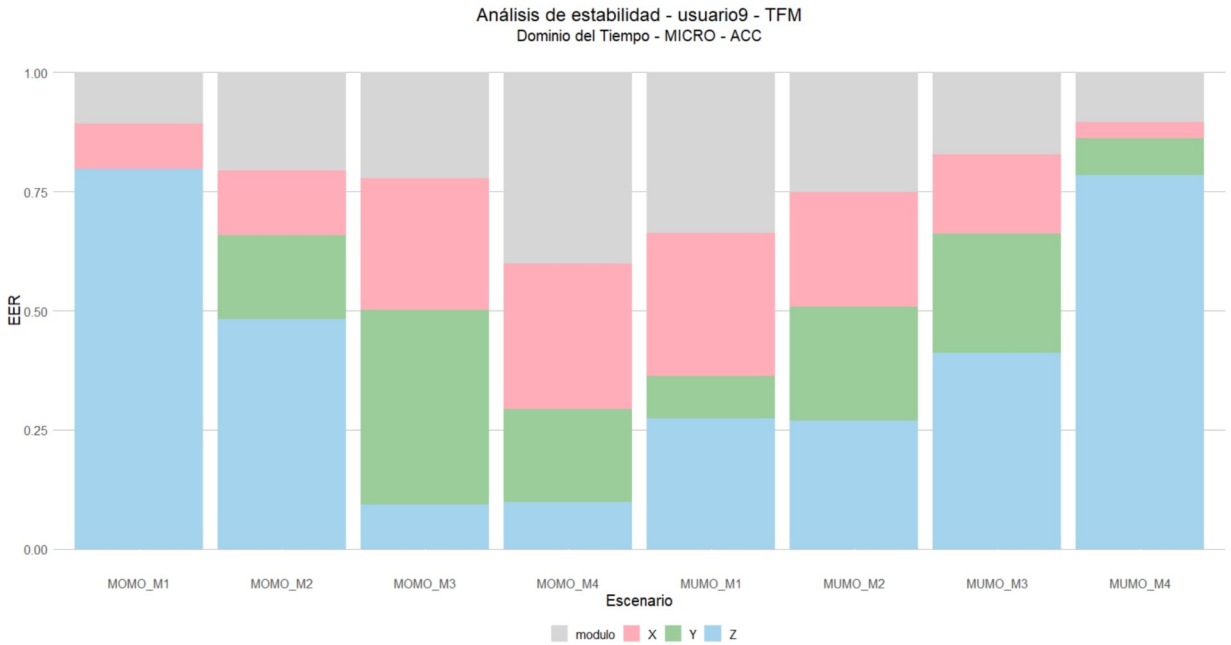


Figura 7.9: Análisis de la estabilidad en el usuario 9 (Dominio del Tiempo, Micro, ACC)

7.3. Aplicación de clasificadores

A la hora de aplicar nuevos clasificadores, se va a tener siempre en cuenta los resultados obtenidos con el modelo no supervisado inicial de 1-vecino más próximo (en las figuras se referencia como "1-NN").

En las gráficas de la figura 7.10 se ve la aplicación de los 6 clasificadores (árboles, random Forest, bagging, naive Bayes, JRip y SVM Lineal) sobre el escenario Monosesión-Monomuestra (MonoMono_M1) en el dispositivo Microsoft, el sensor acelerómetro con las características del dominio del tiempo en cada una de las coordenadas (X/Y/Z/módulo). Eliminar ventanas de la clase mayoritaria (under) e incluso buscar un equilibrio entre añadir patrones a la clase minoritaria y eliminarlos de la mayoritaria (balanced) no funciona bien. Siendo mejor añadir patrones a la clase minoritaria (over) o incluso dejar los datos no balanceados con el algoritmo SVM.

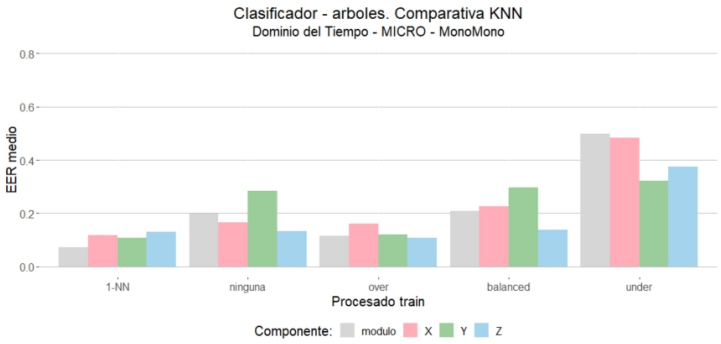
Si se observa la tasa de equierror de cada uno de los usuarios, se puede ver que los árboles de decisión, bagging y JRip presentan sobreajuste, es decir, existen usuarios que sólo son capaces de aprender el comportamiento de la clase mayoritaria, consiguiendo un EER de 1, el peor valor posible.

El mal funcionamiento de los árboles de decisión y el algoritmo basado en reglas (JRip) puede estar relacionado con el desequilibrio entre las clases. Mientras que el sobreajuste en bagging puede deberse a una mala configuración de los parámetros del algoritmo. Por otro lado, el algoritmo basado en probabilidades Naive Bayes tampoco parece funcionar bien en estos datos.

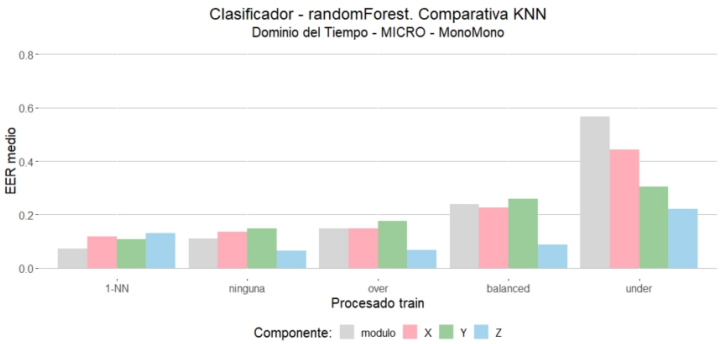
Lo mismo se puede ver cuando se extraen características en el dominio de la frecuencia (figura 7.11). En este caso, se muestran los resultados en el escenario Multisesión-Monomuestra del dispositivo Microsoft y el sensor acelerómetro. Funcionando mejor Random Forest y SVM.

Sin duda, en ambos dominios el mejor funcionamiento lo consiguen los algoritmos Random Forest y SVM, siendo mejor, de manera global, entre los dos, SVM. No resulta preocupante que en este clasificador los mejores resultados se consigan dejando los datos no balanceados, pues en SVM los vectores soporte se construyen solo para los casos de las clases que se encuentran en la frontera. Los casos que están lejos de la frontera pueden ser eliminados sin afectar mucho la calidad de la clasificación, que resulta ser bastante exacta en datos con cierto desbalance.

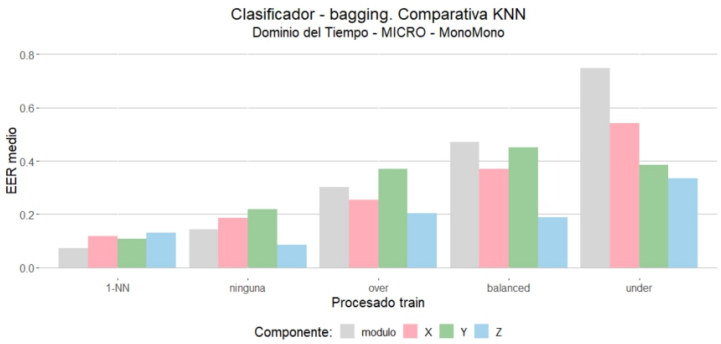
En las figuras 7.12 y 7.13 se muestran los resultados del dominio del tiempo y de la frecuencia, respectivamente, en los 2 escenarios experimen-



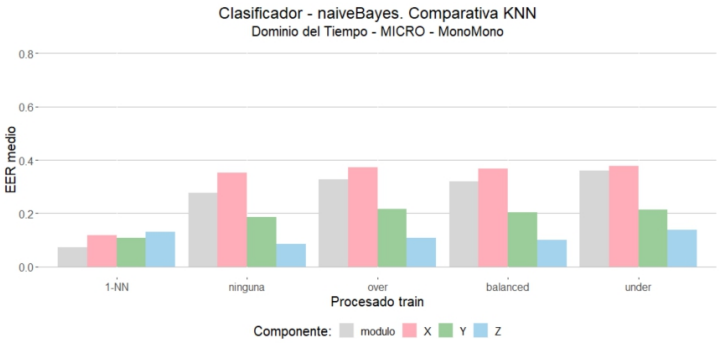
(a) Árboles de decisión



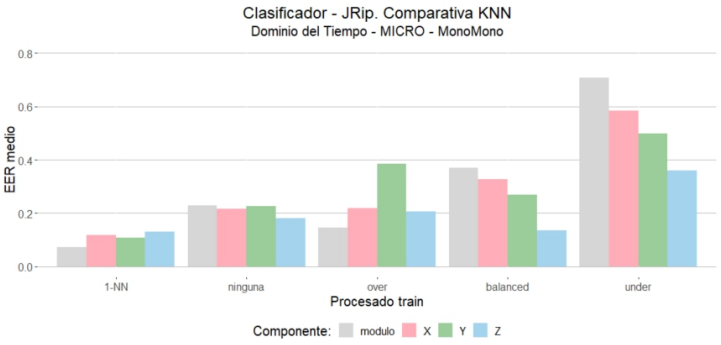
(b) Random Forest



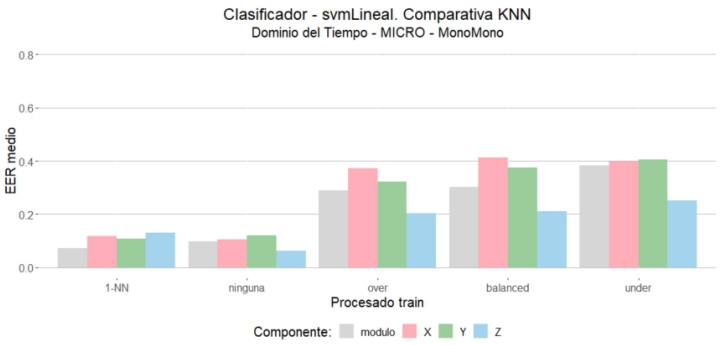
(c) Bagging



(d) Naive Bayes

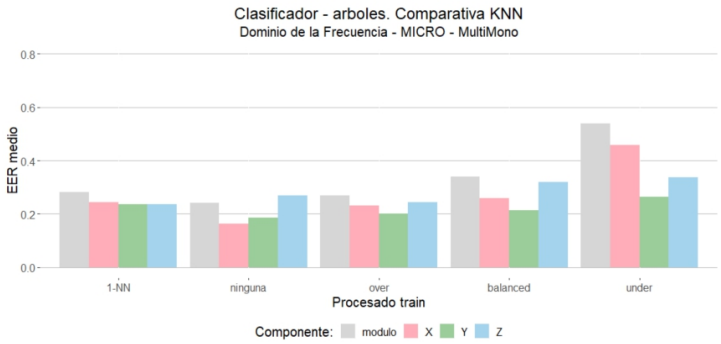


(e) JRip

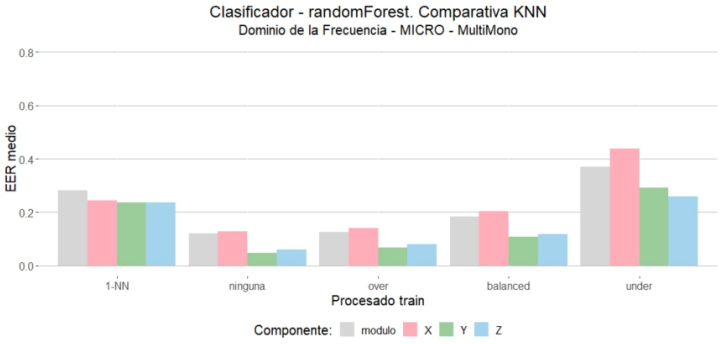


(f) SVM Lineal

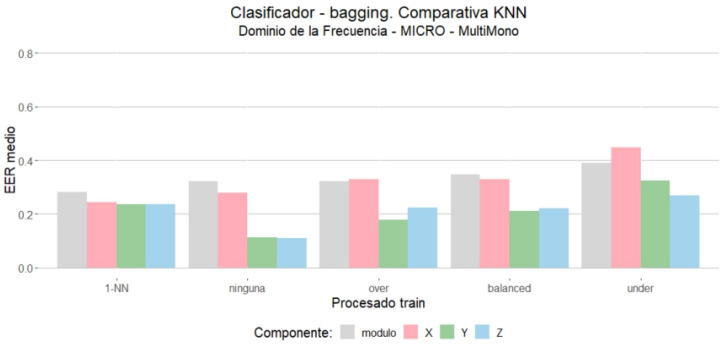
Figura 7.10: Aplicación de clasificadores - Microsoft ACC - Dominio del Tiempo - Monosesión-Monomuestra



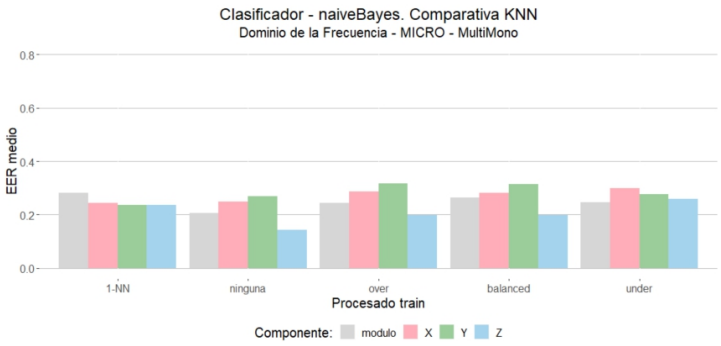
(a) Árboles de decisión



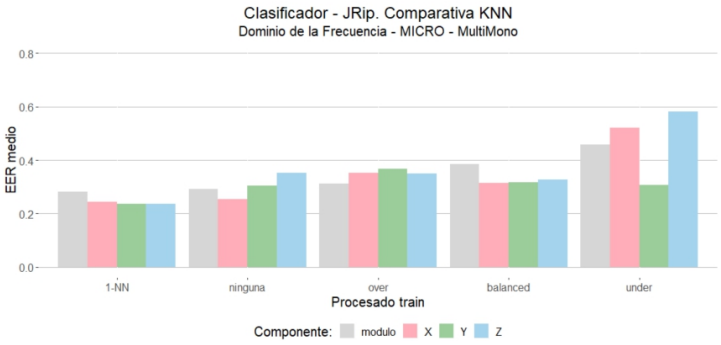
(b) Random Forest



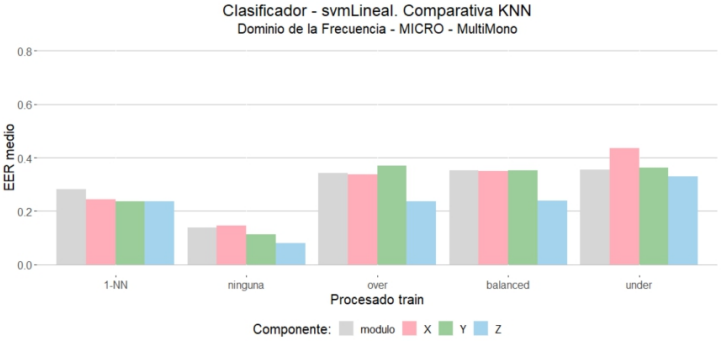
(c) Bagging



(d) Naive Bayes



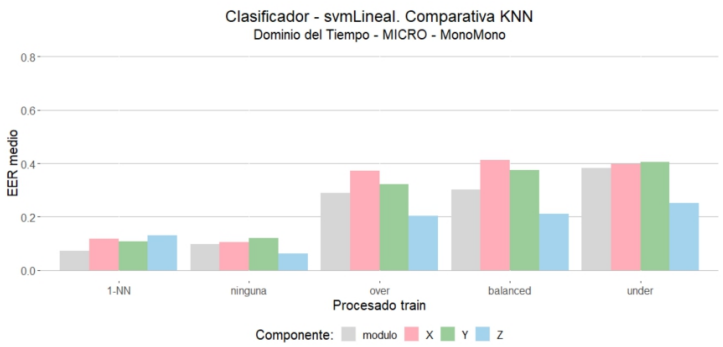
(e) JRip



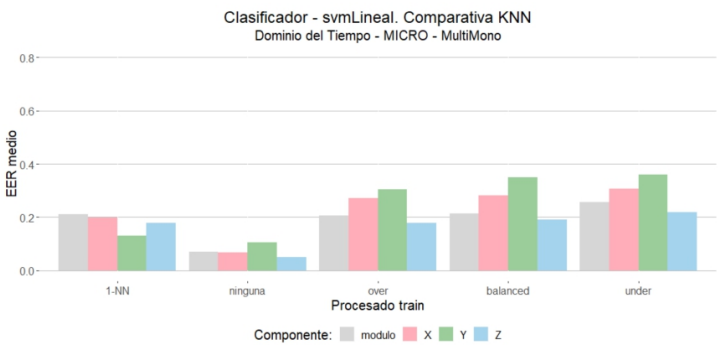
(f) SVM Lineal

Figura 7.11: Aplicación de clasificadores - Microsoft ACC - Dominio de la Frecuencia - Multisesión-Monomuestra

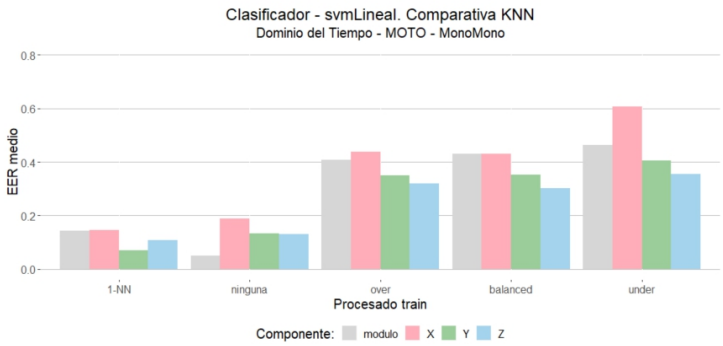
tales más relevantes (MonoMono y MultiMono), en los dos dispositivos (Micro y Moto) con el sensor acelerómetro y el algoritmo SVM. Las conclusiones entre dominios son comunes, funcionan mejor los datos desbalanceados, consiguiendo mejorar los resultados que se tenían con 1-NN. Existiendo una tendencia a que en el dispositivo Micro funcione mejor la coordenada Z y en el dispositivo Moto el módulo, siendo las dos coordenadas que, de manera global, mejor funcionan. No sorprende, ya que en trabajos relacionados [78] ya se menciona que utiliza solamente la dimensión Z del acelerómetro afirmando que existe una asociación entre la fuerza de reacción del suelo y la fuerza de la señal en este eje, generándose picos de gran magnitud que parecen beneficiar al reconocimiento biométrico. Por tanto, se puede decir que el buen funcionamiento del clasificador se mantiene en los dos dominios y en los dos dispositivos, aunque no la coordenada donde se obtienen los mejores resultados.



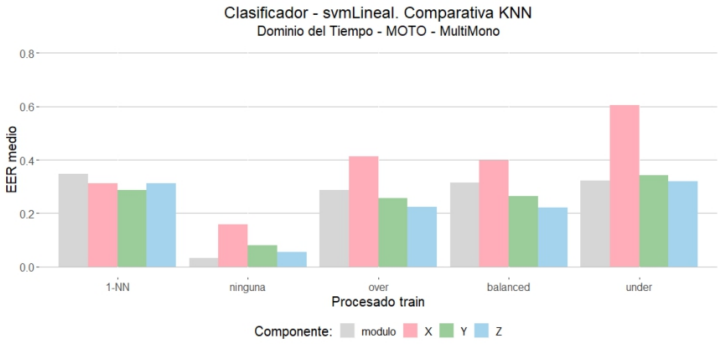
(a) MonoMono - MICRO



(b) MultiMono - MICRO

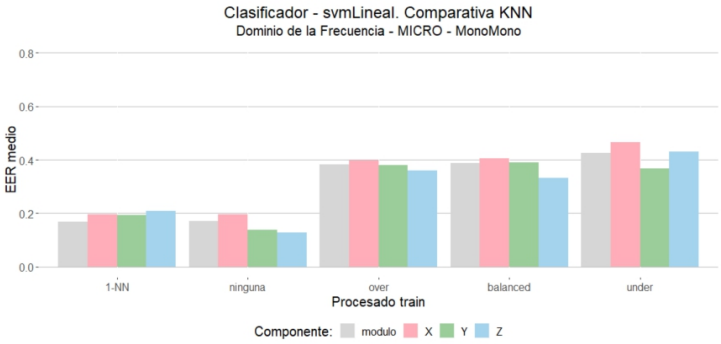


(c) MonoMono - MOTO

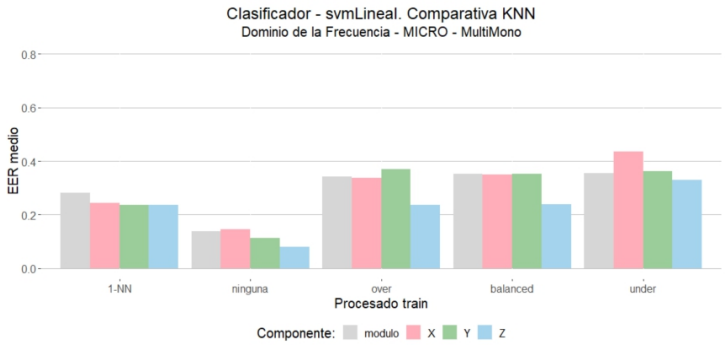


(d) MultiMono - MOTO

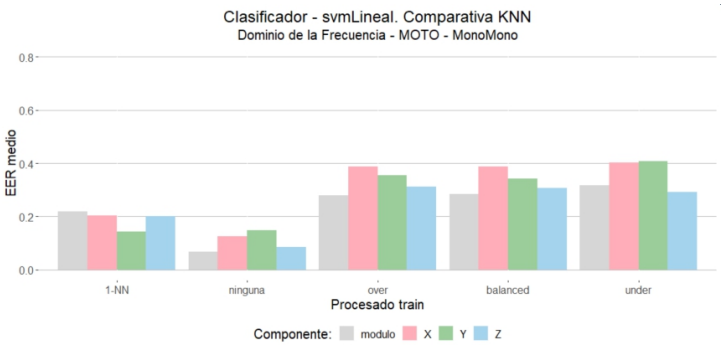
Figura 7.12: Aplicación de clasificadores por escenarios y dispositivos - ACC - Dominio del Tiempo.



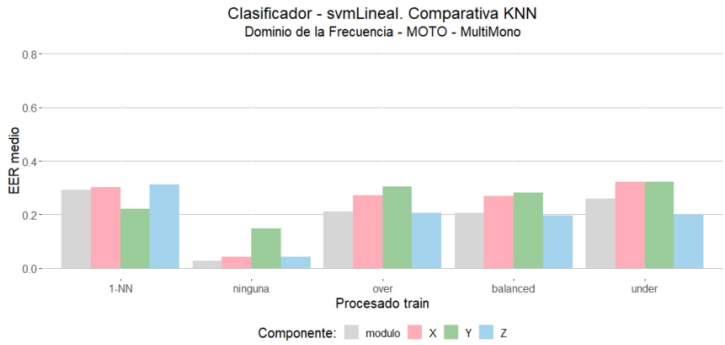
(a) MonoMono - MICRO



(b) MultiMono - MICRO



(c) MonoMono - MOTO



(d) MultiMono - MOTO

Figura 7.13: Aplicación de clasificadores por escenarios y dispositivos - ACC - Dominio de la Frecuencia.

7.4. Profundizar en el clasificador SVM

Todos los resultados mostrados en este apartado van a utilizar el escenario experimental *Multisesión-Monomuestra* por ser el más realista en biometría, al tener en cuenta la variabilidad del rasgo biométrico con el tiempo e incluir los datos de una sola sesión como entrenamiento. Por otro lado, se va a utilizar siempre el sensor acelerómetro por generar, de manera global, los mejores resultados.

Se ha elegido el clasificador, SVM, dejando las incertidumbres que se explicarán a continuación, aunque en algunos casos se ha tomado una decisión clara, basada en los resultados.

- **Normalización del periodo:** a través de la interpolación lineal de los datos, tal y como se explica en el apartado 5.1. En el dominio de la frecuencia, para que las componentes extraídas del análisis de

Fourier tuvieran significado físico, se había decidido si aplicar está normalización, pero en el dominio del tiempo, ante la duda, se había descartado.

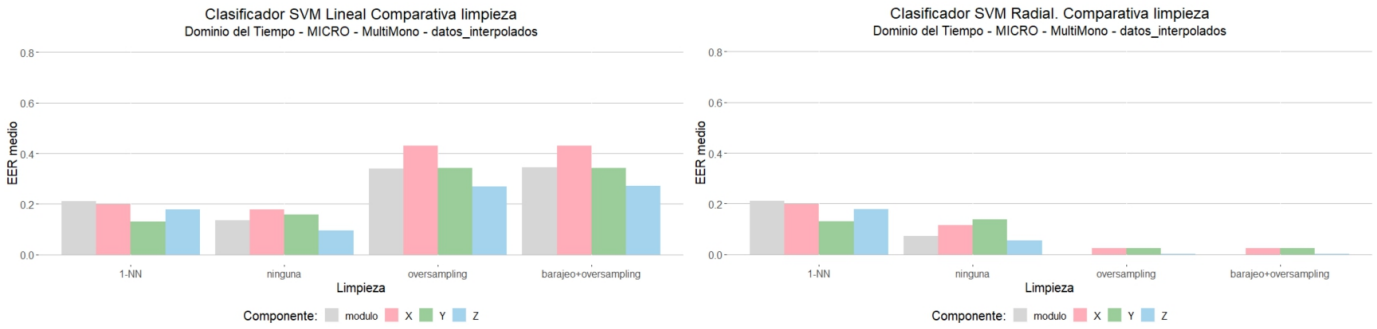
- **Normalización de la amplitud:** para llevar a los datos a una escala común, tal y como se explica en el apartado 5.1. Esta decisión depende del algoritmo de aprendizaje automático, donde se ha visto que no aplicarlo termina generando problemas en el aprendizaje, obteniendo los mismos scores en todas las ventanas del mismo usuario. Para evitar este comportamiento extraño, se ha tomado la decisión de aplicar siempre está normalización.

- **Muestreo de los datos de entrenamiento:** se ha visto un mejor funcionamiento sobremuestreando los datos, pero que dejarlos sin balancear generaba los mismos o mejores resultados.

- **Barajear los datos de entrenamiento:** dependiendo del algoritmo de aprendizaje automático, para evitar problemas de sobreajuste, puede ser necesario barajear los datos. Se ha aplicado una técnica de sobremuestreo barajeando que consiste en ir cogiendo, en orden, ventanas de usuarios auténticos e impostores alternativamente, hasta que no quede ninguna ventana de usuarios impostores (clase mayoritaria). Las ventanas auténticas se van duplicando, empezando por la primera y terminando en la última.

La figura 7.14 del dispositivo micro en el dominio del tiempo demuestra que sobremuestreando los datos, es indiferente barajarlos o no. La figura de la izquierda, 7.14 (a), corresponde a un kernel lineal, el cual tiene un comportamiento malo sobre los datos que hace que sea mejor no balancearlos. Esto no ocurre en la figura de la derecha, 7.14 (b), donde se muestra un kernel radial, que como se verá más adelante es el que muestra mejor rendimiento. Las conclusiones son las mismas en los dos dispositivos y dominios, por lo que de aquí en adelante se probarán los datos sin balancear y sobremuestreados con barajeo.

La figura 7.15 muestra en el dispositivo Micro el efecto, en el dominio del tiempo, de utilizar los datos interpolados y sin interpolar, no balanceados y balanceados con sobremuestreo con barajeo en todos los kernels probados. Cuando no se balancean los datos, el kernel sigmoide consigue la mejor tasa de error, del 0 % en todos los usuarios y componentes. Sin embargo, al sobremuestrearlos pasa a funcionar como uno de los peores kernels, con tasas de error medias en todos los usuarios del 40 %. Mientras que



(a) SVM Lineal - DT - Micro (b) SVM Radial - DT - Micro

Figura 7.14: Comparación en SVM del efecto que provoca barajear los datos, una vez se han sobremuestreado.

el kernel radial muestra un comportamiento bueno y estable en todos los casos, tanto si se muestrean los datos como si se interpolan o no. Lo mismo ocurre en el dispositivo Moto, como se puede ver en la figura 7.16 y en el dominio de la frecuencia, figura 7.17, donde no existen dudas del buen funcionamiento de los datos interpolados y el kernel radial.

Por lo tanto, de aquí en adelante se van a tomar las decisiones que a continuación se indican, consiguiendo por escenarios, dominios y componentes, las tasas de error medias en los 20 usuarios que se muestran en la tabla 7.3 para el dispositivo Micro y 7.4 para el dispositivo Moto.

- Se va a seguir aplicando la interpolación lineal en el dominio de la frecuencia y se va a empezar a interpolar también el dominio del tiempo **(normalización del periodo)**.
- Se van a escalar los datos en ambos dominios por generar mejores resultados **(normalización de la amplitud)**.
- Se va a utilizar el clasificador SVM con kernel radial por obtener buenos resultados y ser el más estable **(algoritmo de aprendizaje automático)**.
- Se van a utilizar los datos balanceados para evitar problemas de sobreajuste. Además, con el kernel elegido, genera resultados ligeramente mejores **(muestreo de los datos de entrenamiento)**.
- Igual que en el punto anterior, para evitar problemas de sobreajuste, se van a barajear los datos **(barajeo de los datos de entrenamiento)**.

Escenario	Componente	EER medio DT (%)	EER medio DF (%)
MonoMono	X	6.87	0.19
MonoMono	Y	2.28	0.28
MonoMono	Z	0.28	0.17
MonoMono	módulo	0.01	0.04
MultiMono	X	2.56	0.06
MultiMono	Y	2.55	0.09
MultiMono	Z	0.12	0.19
MultiMono	módulo	0	0.08

Tabla 7.3: Resultados del dispositivo Micro con el clasificador SVM, kernel radial, datos interpolados para la extracción de características, sobremuestreados y barajeados en el conjunto de entrenamiento (Nueva Base de Datos).

Las tablas 7.3 y 7.4 muestran mejores tasas de error en la componente módulo y en el dominio de la frecuencia. También permiten ver cómo el escenario Multisesión-Monomuestra (MultiMono) genera mejores resultados que Monosesión-Monomuestra (MonoMono). Una explicación para esto es que, al principio, el comportamiento de los usuarios es variable, ya que no es algo natural que te estén grabando mientras caminas. Pero que a medida que se recogen más datos y conocen el experimento, es decir, saben lo que tienen que hacer, su comportamiento pasa a ser más estable.

7.5. Sistema final

Aunque la configuración elegida genera buenos resultados con tasas de error medias en los 20 usuarios inferiores al 7 % en el dominio del tiempo y al 1 % en el dominio de la frecuencia, se van a tratar de seguir mejorando los resultados.

La figura 7.18 muestra la fusión de scores obtenidos con el clasificador SVM, en el escenario *multisesión-monomuestra* en los dos dispositivos y dominios, partiendo de no realizar ninguna fusión y llegando hasta la fusión total (se obtiene un único score de toda la muestra de prueba). En definitiva estamos aumentando el tamaño de la señal usada para reconocer al usuario. Las conclusiones siguen siendo claras, el módulo consigue los mejores resultados y a mayor fusión de ventanas, tal y como se veía en [57], mejores resultados. Llegando a obtener tasas de error medias del

Escenario	Componente	EER medio DT (%)	EER medio DF (%)
MonoMono	X	5.39	0
MonoMono	Y	1.02	0.28
MonoMono	Z	1.18	0.11
MonoMono	módulo	0	0
MultiMono	X	3.59	0.02
MultiMono	Y	0.40	0.10
MultiMono	Z	0.01	0.01
MultiMono	módulo	0	0

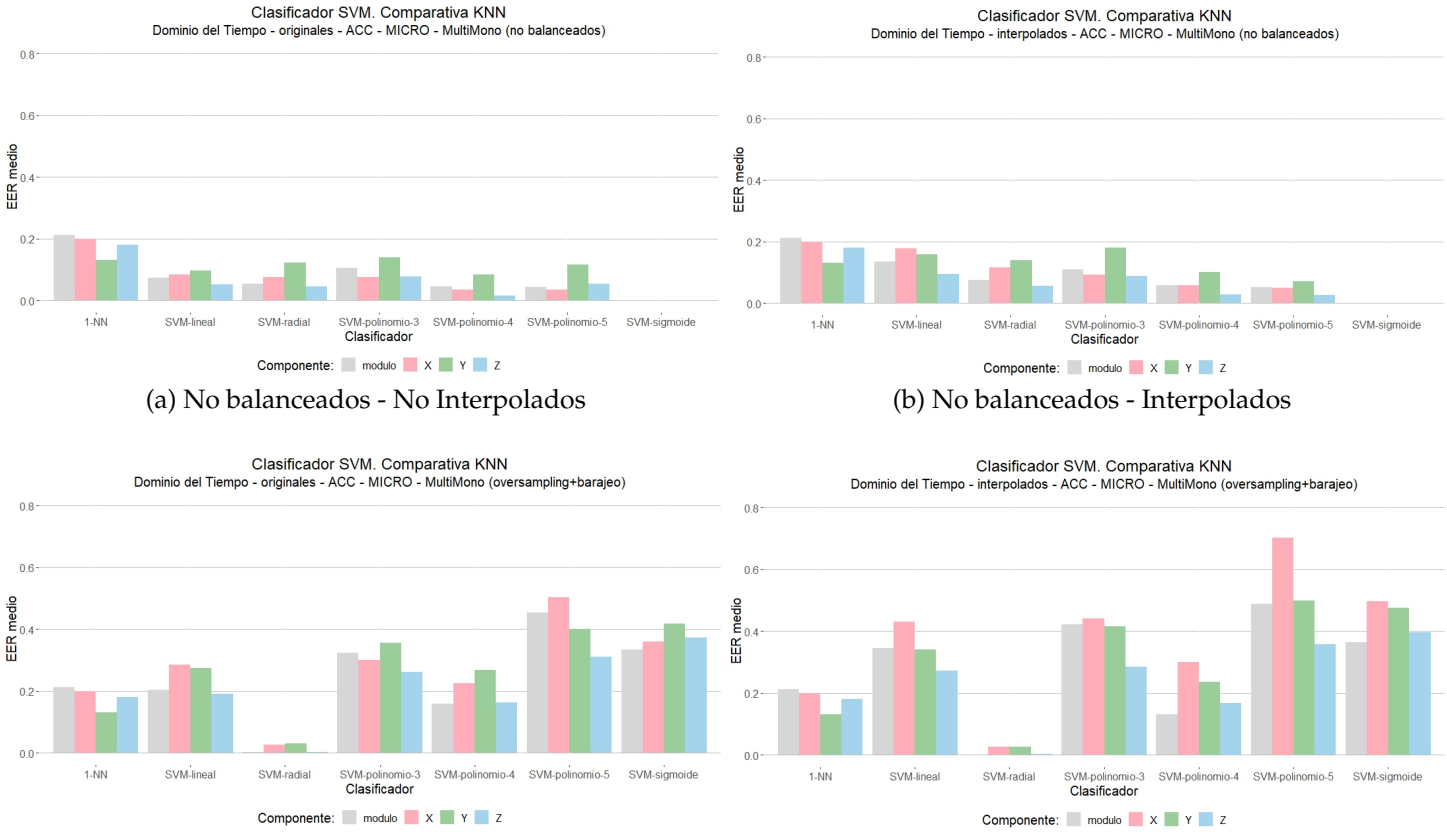
Tabla 7.4: Resultados del dispositivo Moto con el clasificador SVM, kernel radial, datos interpolados para la extracción de características, sobremuestreados y barajeados en el conjunto de entrenamiento (Nueva Base de Datos).

0 % en todas las componentes al hacer una fusión de todas las ventanas. Las mismas conclusiones se obtienen al aplicar la fusión de scores con el mismo sistema sobre la base de datos adquirida en [4] con 12 usuarios (figura 7.19).

Por último, la figura 7.20 muestra los resultados del dominio del tiempo y de la frecuencia, de fusionar los scores de ambos dominios con diferentes estadísticos (suma, media y producto) y de hacer la fusión total de scores que se mostraba en las figuras 7.18 y 7.19, en ambos dispositivos para la base de datos con 20 usuarios en (a) y (b) y de 12 usuarios en (c) y (d). Resulta indiferente el estadístico que aplicar a la fusión de scores de ambos dominios, los resultados son los mismos, mejorando los del dominio del tiempo, pero no los del dominio de la frecuencia. No obstante, sigue siendo mejor utilizar un único dominio y fusionar todos los scores (fusión total) ya que, en este escenario, como ya se veía, en todas las componentes se obtienen tasas de error del 0 %.

7.6. Influencia de la mano en el uso del dispositivo

Como las conclusiones en *monosesión-monomuestra* y *multisesión-monomuestra* son las mismas, se van a mostrar únicamente los gráficos para el segundo caso (MultiMono).



(a) No balanceados - No Interpolados

(b) No balanceados - Interpolados

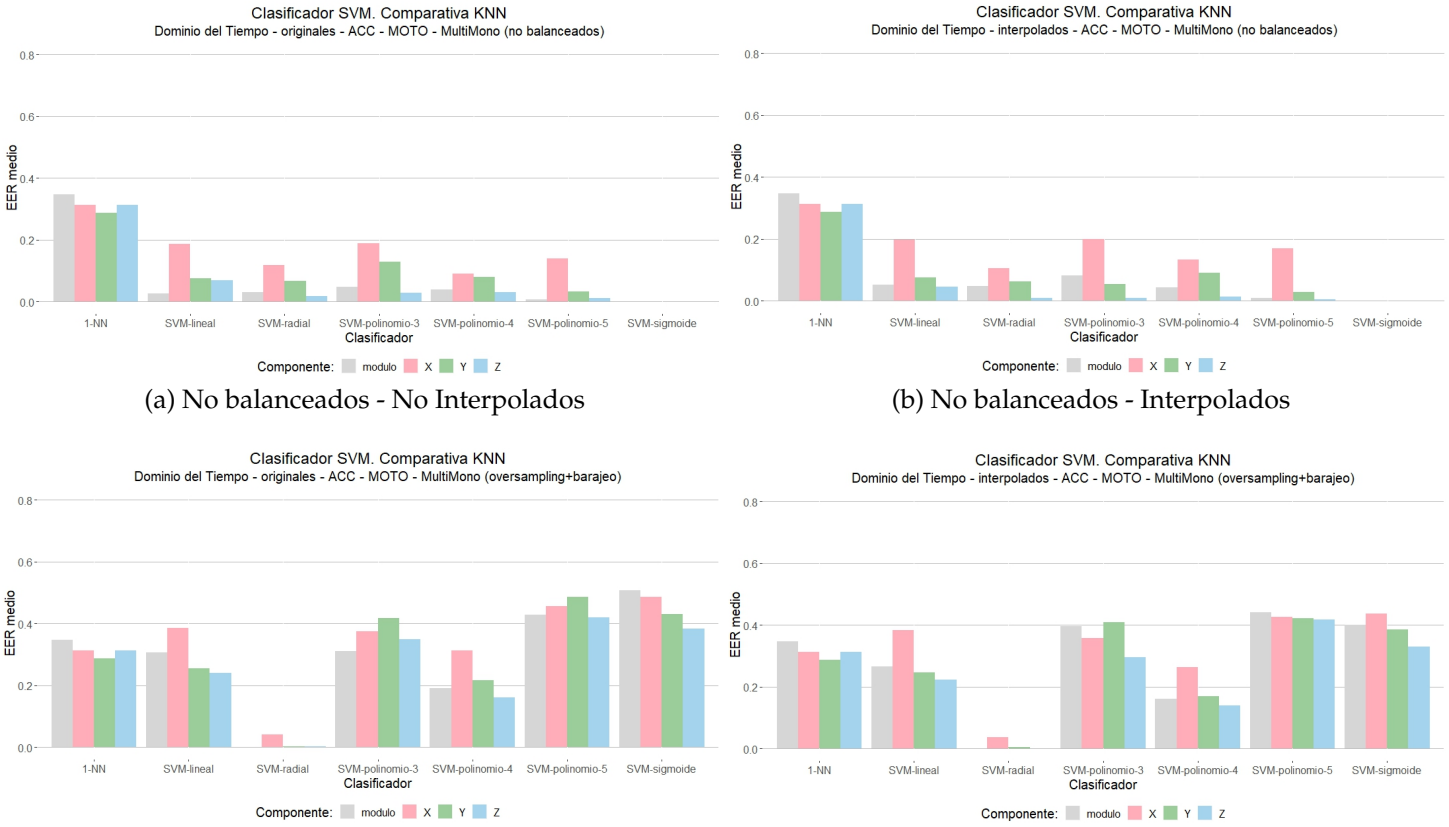
(c) Balanceados - No Interpolados

(d) Balanceados - Interpolados

Figura 7.15: Efecto en el dominio del tiempo y el dispositivo Micro de interpolar y muestrear los datos en el clasificador SVM.

En base a los resultados vistos en los apartados anteriores se han considerado los siguientes sistemas:

- **No Fusión:** no realizando ninguna fusión de scores, como situación inicial.
- **Fusión con 4 scores y 2 de solapamiento:** situación intermedia que funciona bien.
- **Fusión con 10 scores y 2 de solapamiento:** situación intermedia que funciona mejor que la anterior.
- **Fusión total:** sistema que conseguía los mejores resultados.



(a) No balanceados - No Interpolados

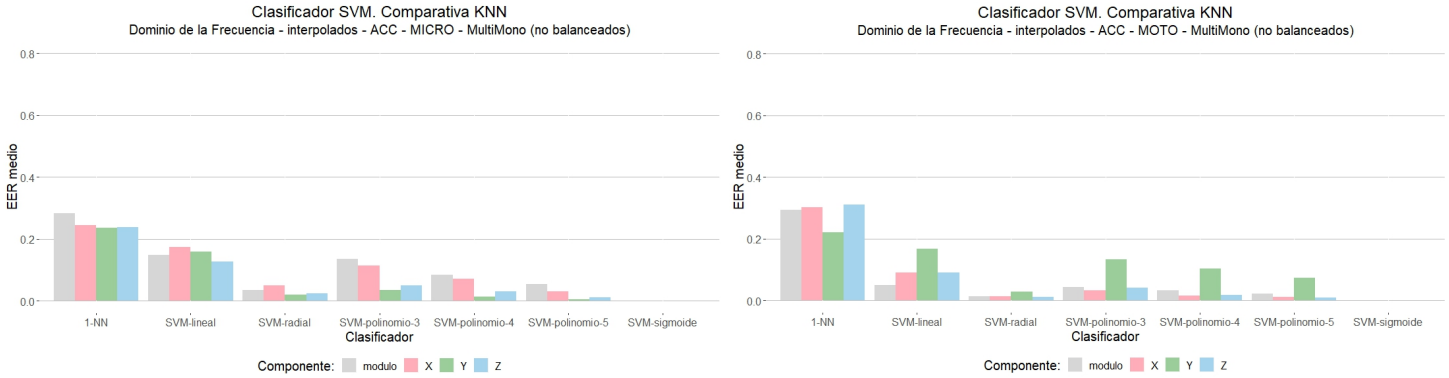
(b) No balanceados - Interpolados

(c) Balanceados - No Interpolados

(d) Balanceados - Interpolados

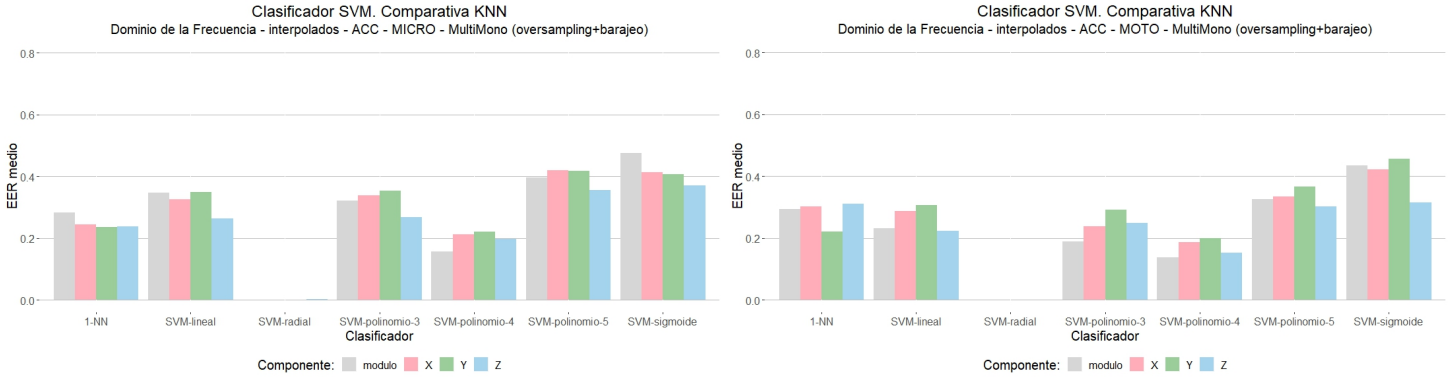
Figura 7.16: Efecto en el dominio del tiempo y el dispositivo Moto de interpolar y muestrear los datos en el clasificador SVM.

La figura 7.21 muestra los resultados del dispositivo Micro en el dominio del tiempo, con cada sistema utilizando la mano de uso habitual del reloj y la opuesta en los dos escenarios planteados. De igual manera, la figura 7.22 muestra los resultados del dispositivo Moto en el dominio de la frecuencia. Las conclusiones son claras y coinciden entre dominios y dispositivos. Indicando que es irrelevante la mano en la que el usuario lleva colocado el dispositivo y, por tanto, que el movimiento de las manos es simétrico. La obtención de buenos resultados también parece indicar lo que se veía en el apartado anterior, que el usuario se siente más cómodo a medida que lleva más tiempo andando y su señal resulta más estable, siendo más fácil de predecir y generando mejores resultados.



(a) Micro - No balanceados

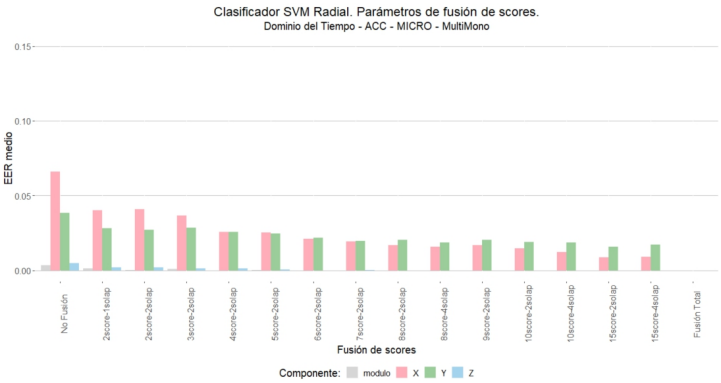
(b) Moto - No balanceados



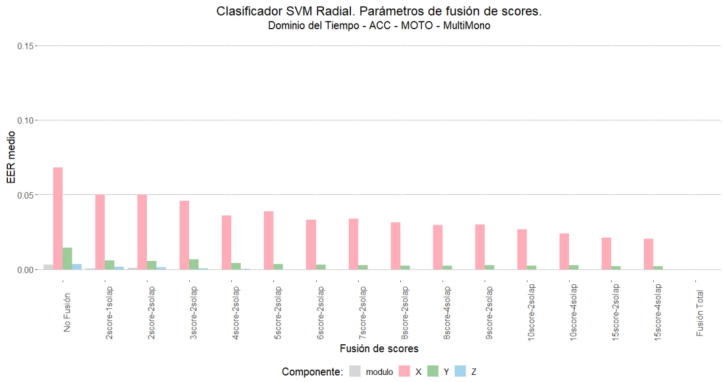
(c) Micro - Balanceados

(d) Moto - Balanceados

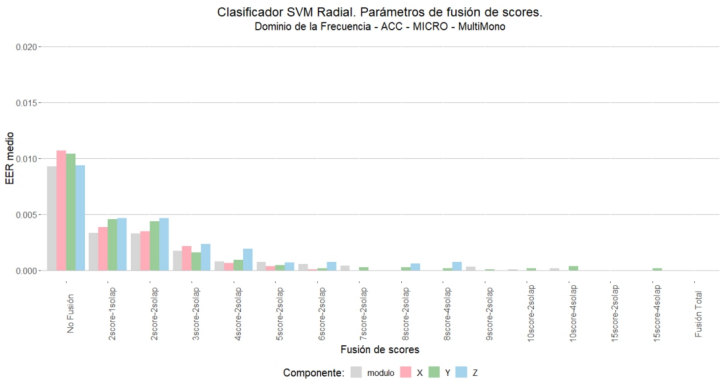
Figura 7.17: Efecto en el dominio de la frecuencia de muestrear los datos en el clasificador SVM.



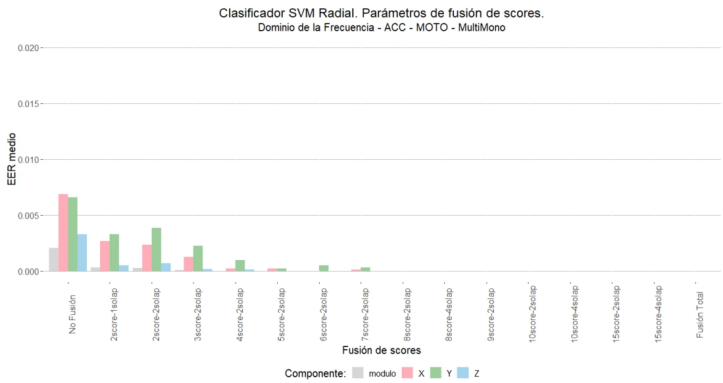
(a) Micro - DT - MultiMono



(b) Moto - DT - MultiMono

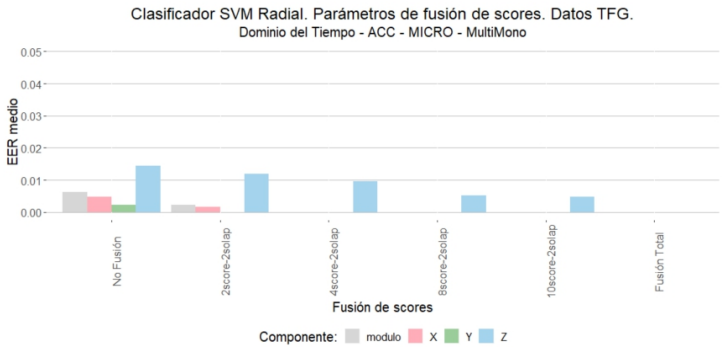


(c) Micro - DF - MultiMono

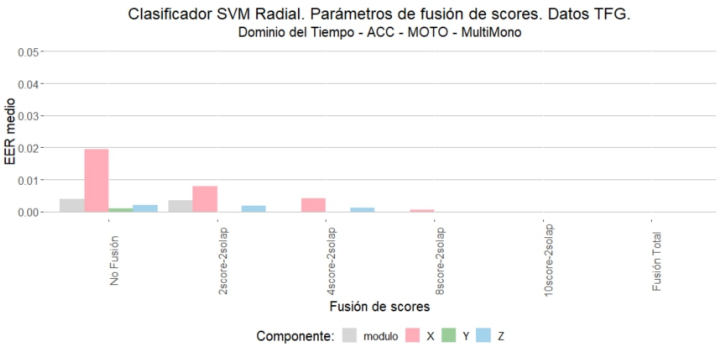


(d) Moto - DF - MultiMono

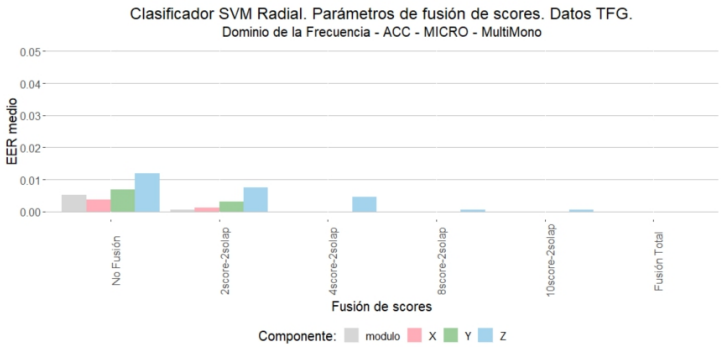
Figura 7.18: Efecto de variar el número de scores fusionados con el clasificador SVM y la base de datos adquirida en este proyecto con 20 usuarios. **Para una mejor visualización de los resultados la escala del eje Y es especifica de cada dominio debido a que el dominio de la frecuencia funciona mejor que el del tiempo.*



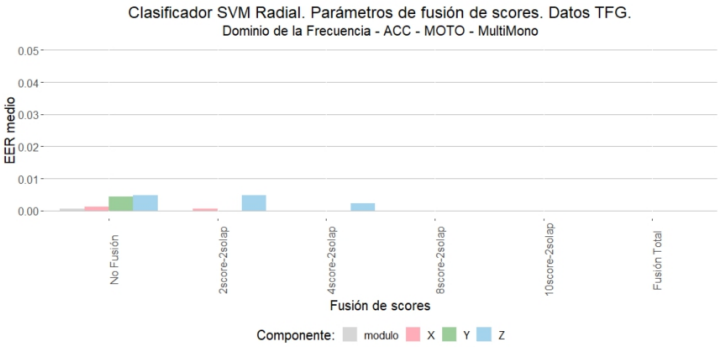
(a) Micro - DT - MultiMono



(b) Moto - DT - MultiMono

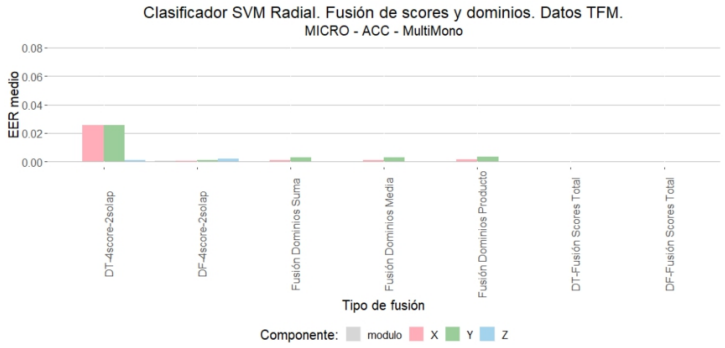


(c) Micro - DF - MultiMono

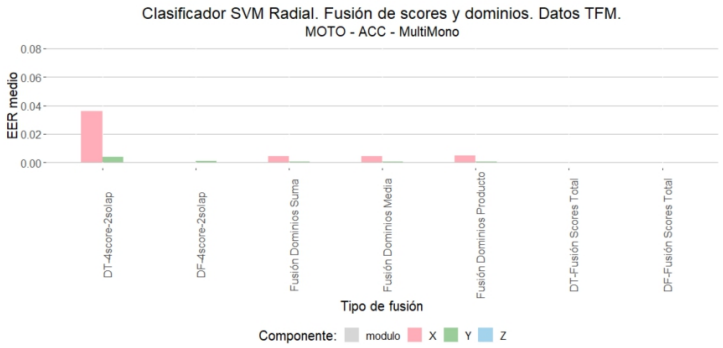


(d) Moto - DF - MultiMono

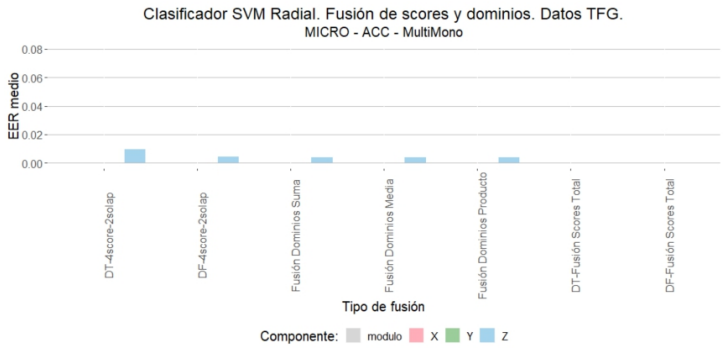
Figura 7.19: Efecto de variar el número de scores fusionados con el clasificador SVM y la base de datos adquirida en [4] con 12 usuarios.



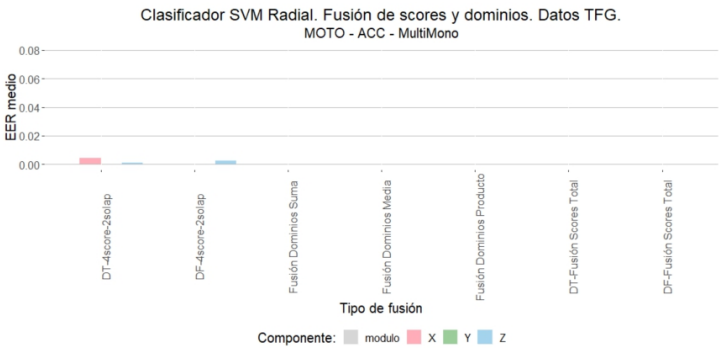
(a) Micro - BD1



(b) Moto - BD1

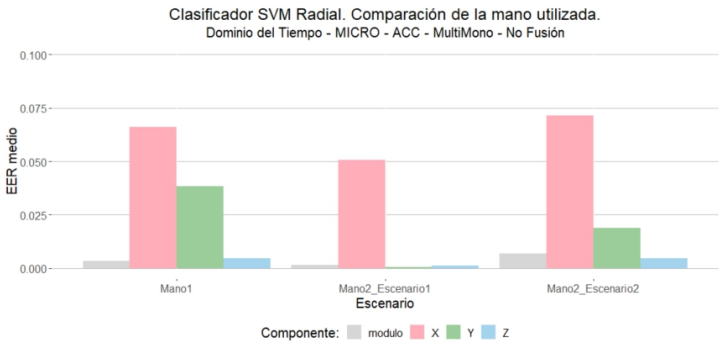


(c) Micro - BD2

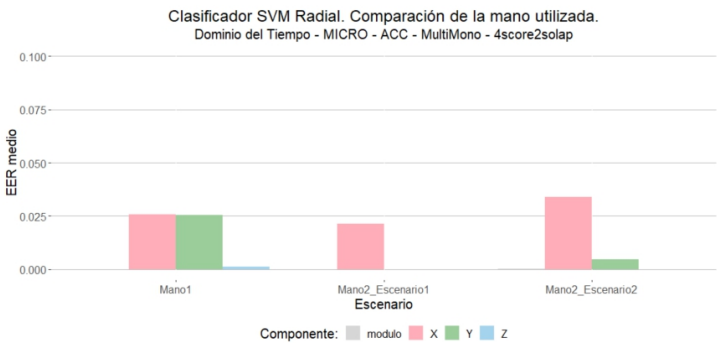


(d) Moto - BD2

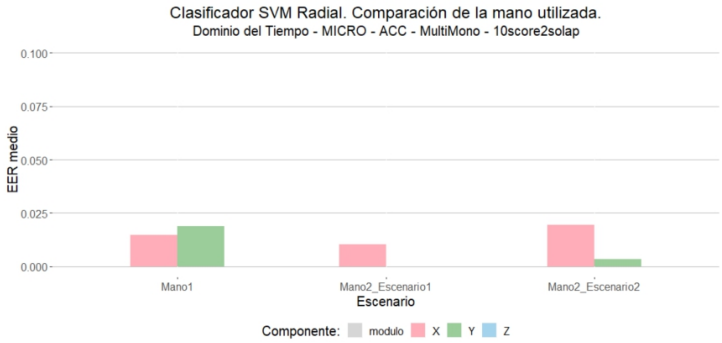
Figura 7.20: Efecto de fusionar dominios y scores con el clasificador SVM y las dos bases de datos disponibles. BD1 es la base de datos adquirida para este proyecto con 20 usuarios y BD2 es la adquirida en [4] con 12 usuarios.



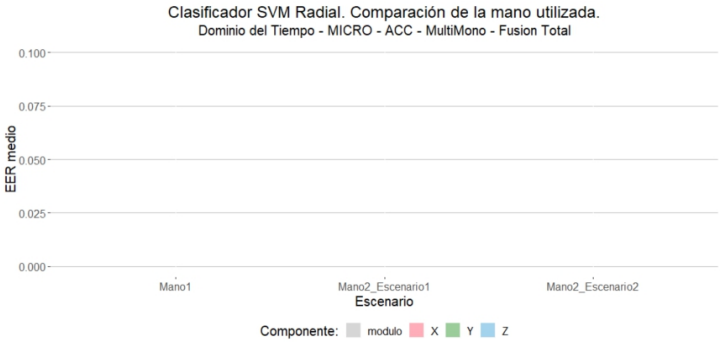
(a) No Fusión - DT - Micro



(b) Fusión 4 scores / 2 solap - DT - Micro

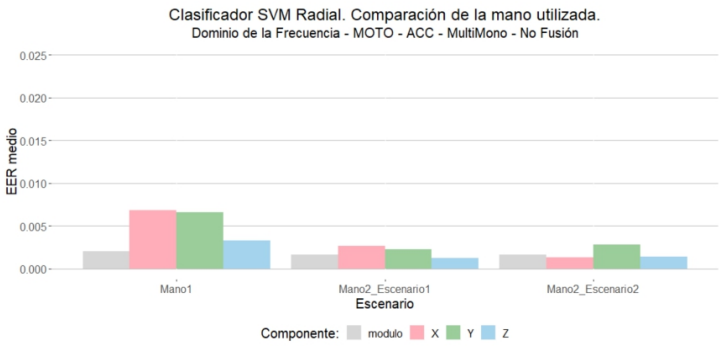


(c) Fusión 10 scores / 2 solap - DT - Micro

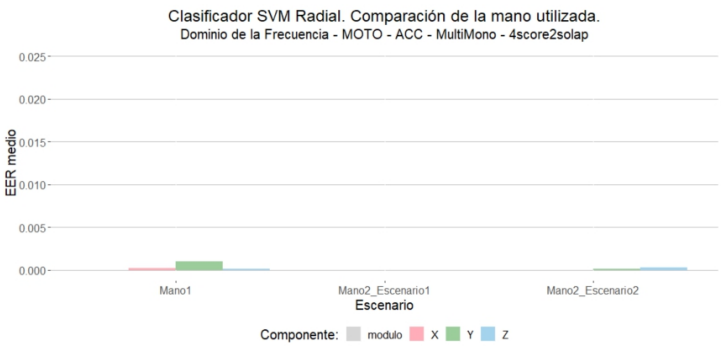


(d) Fusión total - DT - Micro

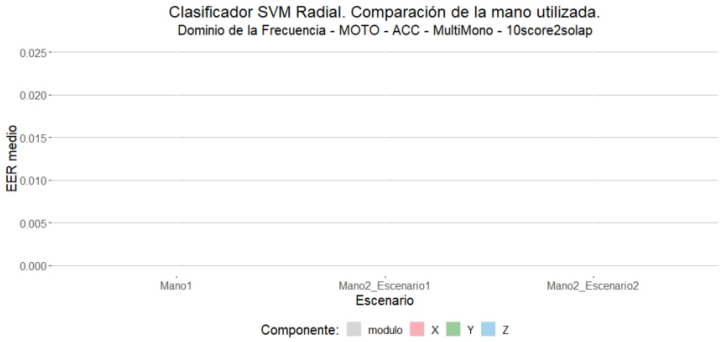
Figura 7.21: Influencia de la mano de uso del dispositivo utilizando la pulsera de Microsoft y la extracción de características en el dominio del tiempo.



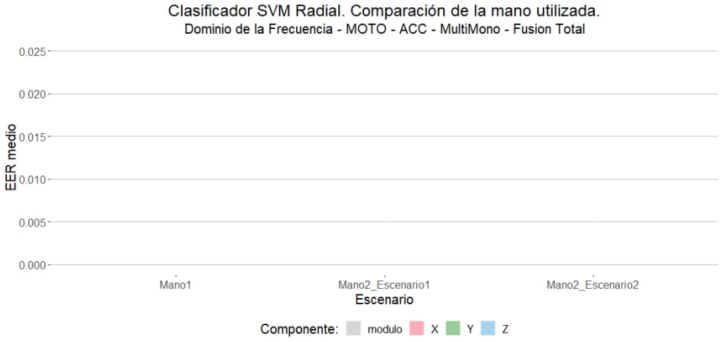
(a) No Fusión - DF - Moto



(b) Fusión 4 scores / 2 solap - DF - Moto



(c) Fusión 10 scores / 2 solap - DF - Moto



(d) Fusión total - DF - Moto

Figura 7.22: Influencia de la mano de uso del dispositivo utilizando el reloj de Motorola y la extracción de características en el dominio de la frecuencia.

Capítulo 8

Arquitectura

Las conclusiones experimentales alcanzadas parecen indicar que se ha logrado un buen sistema de reconocimiento que podría ser implementable. Esto conlleva múltiples vías de trabajo futuro, donde la más importante es capturar más datos, incrementando el tamaño experimental. Para lo que es necesario implementar una arquitectura escalable.

Los datos con los que actualmente se ha trabajado y hecho pruebas experimentales no son Big Data. No obstante, cuando se capturan datos y el usuario lleva el dispositivo colocado en su muñeca, se adquiere un nuevo dato de las coordenadas X, Y, Z de cada sensor cada 83ms, aproximadamente. Lo que significa, por sensor y usuario.

- 720 adquisiciones de datos al minuto.
- 43200 adquisiciones de datos en una hora.

En los datos actuales, adquiridos para este trabajo, cada usuario se mantiene andando, en total, una hora y se adquieren datos de 2 sensores, lo que significan 86400 datos adquiridos con cada usuario. Datos de tamaño pequeño, que crecen muy rápido.

Mientras el usuario anda, una aplicación Android instalada en el teléfono móvil vinculado al ponible (wearable) almacena los datos de las coordenadas X, Y, Z de cada sensor. Una vez adquiridos, la propia aplicación tiene un botón, "Sincronizar", que realiza, a través de un protocolo SOAP estándar, la conexión a una base de datos relacional (MySQL del grupo de investigación [89]). Para la explotación de esos datos, se descargan, a través de un programa Java, a un fichero de texto

plano (CSV). Existiendo un fichero para cada usuario, sesión, muestra de datos, dispositivo y sensor.

Como próximos pasos, se proponen dos escenarios, para los cuales esta manera de trabajar es inabordable, ya que resultaría un sistema poco eficiente.

8.1. Escenario experimental de investigación

Consiste en capturar más datos, tanto de más usuarios como de aquellos ya disponibles. Hasta ahora, la adquisición ha sido supervisada, en un recorrido fijo. Aquí se propone, una adquisición de datos no supervisada, en la que el usuario tenga los dispositivos y durante varios días, de manera continua, se capturen sus datos, realizando una variedad de actividades, no solo caminar. De esta manera, se tendrían millones de datos.

Estos datos hay que almacenarlos y utilizar la base de datos relacional actual generaría un bajo rendimiento. Como alternativas, se podrían usar bases de datos especializadas en datos temporales (timeseries, TSDB) o directamente un almacenamiento optimizado en HDFS (Hadoop Distributed File System).

Una base de datos de series temporales es una base de datos vertical con propiedad de marca de tiempo. Surgieron, principalmente, enfocadas al entorno de la bolsa, pero rápidamente se incorporaron más aplicaciones, ya que no existe casi ningún sector que no incluya la componente temporal en los datos de su día a día. DB-Engines [22], un sitio web de clasificación de popularidad de bases de datos, clasifica y cuenta este tipo como un directorio independiente, cuya tasa de crecimiento, en los últimos meses, ocupa el primer lugar en todas las clasificaciones de bases de datos (figura 8.1). Usan la marca de tiempo (timestamp), definida por el momento temporal en que fue tomada la métrica en cuestión, pudiendo tener parámetros adicionales que puedan aportar información sobre la marca, por lo que son aptas para nuestros datos, los cuales están organizados de la siguiente manera. Se muestra un ejemplo en la figura 8.2:

- Marca temporal: momento en el que se produjo la extracción de la métrica, campo *fecha* en 8.2. En este tipo de bases de datos, este campo funciona como clave primaria ya que sirve para posteriormente, realizar búsquedas en rangos temporales de manera muy optimizada. Puede expresarse en diferentes formatos en función del nivel de precisión que se necesite.

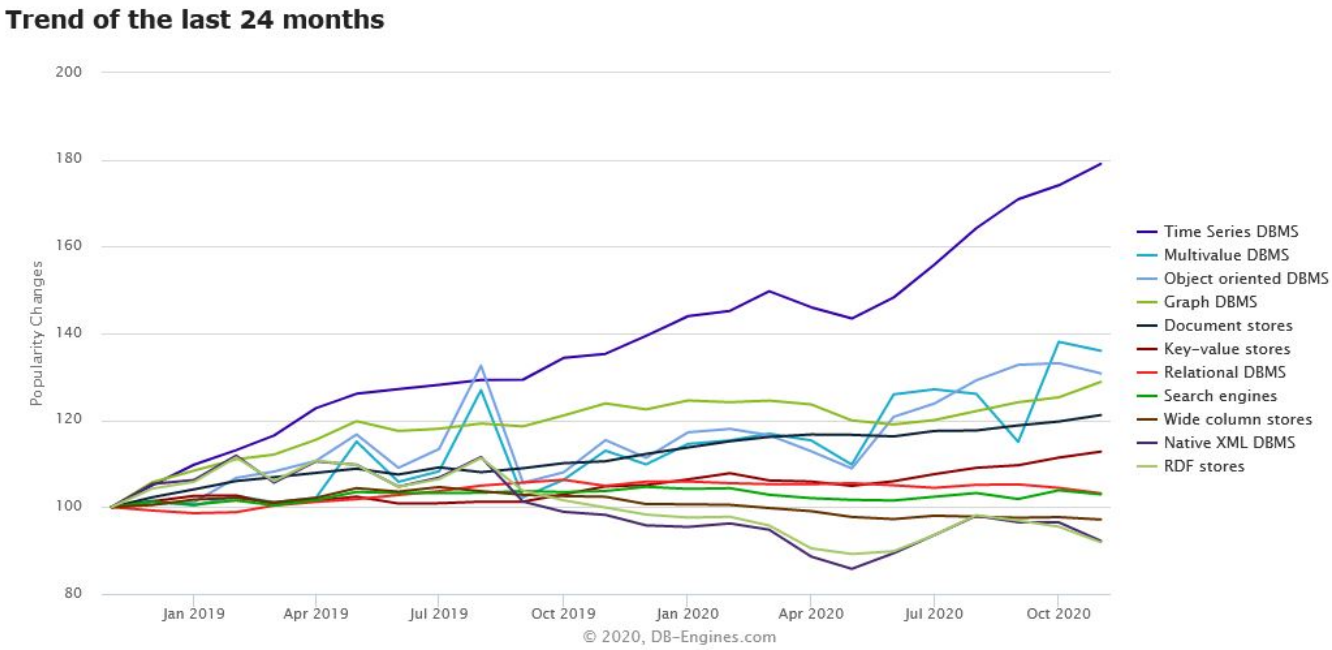


Figura 8.1: Evolución de la tendencia de bases de datos en los últimos 24 meses. Extraído el 2 de noviembre de 2020 [22].

id	imei	wearable	tipo_sensor	fecha	datos0	datos1	datos2	tarea	sesion	muestra	usuario
18	358729083318503	Microsoft Band 2	ACC	2019-09-23 05:45:30.742	0.962891	0.011963	0.258057	1	1	1	cavipa
19	358729083318503	Microsoft Band 2	ACC	2019-09-23 05:45:30.950	0.954834	0.014648	0.231201	1	1	1	cavipa
20	358729083318503	Microsoft Band 2	ACC	2019-09-23 05:45:31.000	0.956787	0.040283	0.167969	1	1	1	cavipa
21	358729083318503	Microsoft Band 2	ACC	2019-09-23 05:45:31.045	0.922852	-0.070557	0.338135	1	1	1	cavipa
22	358729083318503	Microsoft Band 2	GYR	2019-09-23 05:45:31.259	-13.597561	3.993902	-8.871951	1	1	1	cavipa

Figura 8.2: Ejemplo de dato a almacenar en la base de datos.

- International Mobile Equipment Identity (IMEI) del teléfono móvil utilizado para la adquisición de los datos.
- Wearable para identificar el dispositivo ubicado en la muñeca (reloj o pulsera) del que se están adquiriendo datos.
- Tipo de sensor al que pertenece el dato.
- Las coordenadas X, Y, Z del sensor indicado (campos *dato0*, *dato1* y *dato2* en 8.2).
- Número de la tarea, la sesión y la muestra para distinguir entre las diferentes actividades realizadas por el usuario. Actualmente,

estos campos se utilizan para distinguir entre las diferentes tomas de datos del mismo usuario, pero en una adquisición continua no sería necesario.

- Identificador del usuario, campo *usuario* en 8.2.
- Podrían almacenarse otros parámetros adiciones como el origen (lugar/ubicación de donde proviene la métrica).

TSDB es un sistema de software que está optimizado para manejar datos de series de tiempo, pensadas para la ingesta de gran cantidad de datos. Se caracteriza por los siguientes aspectos [3]:

- **Concurrencia:** soporta altos picos de consultas y escrituras de manera simultánea.
- **Consultas optimizadas:** las consultas en este tipo de bases de datos están optimizadas para realizar análisis de datos, llegando a poder realizar consultas de varios tera-bytes de datos en cuestión de segundos.
- **Orientadas al almacenamiento:** son capaces de almacenar grandes cantidades de datos que pueden llegar hasta peta-bytes de información.
- **Alta disponibilidad:** proporcionando mecanismos para que la información esté siempre accesible.
- **Arquitectura distribuida:** la razón principal por la que deben ser distribuidas es por sus altos requisitos de escritura y almacenamiento. Permite asegurar la alta capacidad de ingesta y concurrencia de consultas, facilitando así servicios altamente escalables y de alta disponibilidad.

Como ventajas del uso de bases de datos de series temporales respecto a las relacionales, se presentan dos aspectos:

- **Escalabilidad:** presentan gran capacidad de adaptación al aumento de los datos. Las bases de datos relacionales no soportan el almacenamiento de una gran cantidad de datos, por ejemplo, millones de datos recogidos por un sensor y estos multiplicados por los distintos sensores disponibles, como podrían ser estos datos. La explicación se

basa en que las bases de datos relacionales usan el paradigma ACID (Atomicidad, Consistencia, Aislamiento y Durabilidad) mientras que las no relaciones, por lo general, siguen el paradigma BASE (Disponibilidad básica, Estado suavizado, Consistencia eventual). Este paradigma permite mayor ingesta de datos y fácil operatividad con ellos, mientras que los basados en el sistema ACID necesitan más garantías y pautas para asegurar los datos, haciendo así que sean menos eficientes para tareas que necesitan alta escalabilidad.

- **Rendimiento:** Interesa realizar búsquedas por marcas temporales (fecha, hora) que es precisamente para lo que están optimizadas las bases de datos de series temporales.

Por otro lado, HDFS es el sistema de ficheros distribuido de Hadoop que se diseña e implementa para satisfacer las necesidades de almacenar grandes cantidades de datos, para posteriormente ejecutar sobre ellos aplicaciones con una carga intensiva de procesado de datos. Se caracteriza por los siguientes aspectos [86]:

- **Arquitectura distribuida:** como su propio nombre indica. Tiene capacidad para almacenar los archivos en un clúster de varias máquinas. La cantidad de datos que se pueden almacenar se pueden escalar con facilidad, añadiendo nuevos nodos al clúster para aumentar la capacidad de almacenamiento.
- **Redundancia:** proporciona redundancia, almacenando los ficheros varias veces y en varios equipos distintos, para evitar que si uno de ellos falla, los datos se pierdan.
- **Detección de fallos y recuperación:** cuenta con mecanismos para una rápida y automática detección de fallos y recuperación.
- **Velocidad de transferencia de datos:** reduce el tráfico de red y aumenta el rendimiento.
- **Facilidad de procesamiento** a través de sistemas heterogéneos de hardware y sistema operativo.

Por todo lo expuesto y dado que se trata de datos temporales, se propone la utilización de bases de datos timeseries (TSDB) aprovechando la concurrencia y el procesamiento más rápido, al disponer de consultas temporales optimizadas. No obstante, utilizarlo como almacenamiento

intermedio y crear un proceso de extracción de los datos a través de un JDBC que los lleve desde timeseries a HDFS cada cantidad fija de tiempo, por ejemplo, 6 meses. El objetivo de utilizar HDFS es tener un histórico de los datos, con redundancia, mecanismos de detección de fallos y recuperación. Otra de las razones por las que crear este sistema y no utilizar únicamente HDFS es porque éste no se comporta especialmente bien cuando se actualizan los datos con frecuencia.

8.2. Sistema real, en streaming

Este segundo escenario busca explotar los datos a medida que el usuario va andando. Es un escenario más realista, en el que podrían estar millones de usuarios generando datos que necesitan ser almacenados y explotados en tiempo real y para ello, se necesita una estructura distribuida eficiente. Se proponen los siguientes pasos:

1. Seguir utilizando una aplicación Android de adquisición de los datos de los dispositivos (wearables) pero, en lugar de almacenarlos en el móvil y tener que sincronizarlos, enviarlos a la base de datos, directamente, en tiempo real, a través de una conexión de tipo HTTP.

2. Disponer de un servicio Kafka de recogida continua de los datos.

3. Proceso batch de Spark que lee el servicio Kafka y almacena los datos en una base de datos timeseries. El proceso de almacenamiento sería el mismo que el propuesto en el escenario anterior. Base de datos timeseries y utilizar HDFS como histórico.

4. Explotación de los datos almacenados en la base de datos timeseries a través de consultas al uso (SQL) y desde el lenguaje de programación que se elija.

Para la explotación de los datos, actualmente, se utiliza el lenguaje de programación R a través de la lectura de ficheros de texto plano almacenados en local. Aquí se propone la explotación de los datos directamente desde la base de datos timeseries y como lenguaje de programación, se proponen los siguientes [47, 42]:

- **Spark:** Apache Spark es un sistema de computación distribuido de código abierto basado en Hadoop, pensado para el análisis y

procesamiento de datos en los campos del Big Data y el Machine Learning. Al ejecutarse sobre Apache Spark, Spark Streaming permite la utilización de potentes aplicaciones interactivas para el procesamiento en tiempo real de flujos de datos, aprovechando las facilidades de uso de Spark y su tolerancia a fallos al ser un sistema distribuido. Sus ventajas son la velocidad, ya que su diseño se ha enfocado en optimizar el rendimiento en el procesamiento de datos a gran escala, aprovechando conceptos como el procesamiento en memoria y otras optimizaciones, su facilidad de uso y su motor unificado, al ir empaquetado con bibliotecas de nivel superior, que incluyen soporte para consultas SQL, transmisión de datos, aprendizaje automático y procesamiento de gráficos. Estas bibliotecas estándar aumentan la productividad del desarrollador y se pueden combinar sin problemas para crear flujos de trabajo complejos.

- **Python:** es uno de los lenguajes más populares y calificados con más futuro por la cantidad de ventajas que aporta: una sintaxis muy sencilla, requiere una curva de aprendizaje muy rápida con una comunidad de grandes programadores que comparten código y librerías. Tiene infinidad de aplicaciones tanto para desarrollos cloud, aplicaciones de escritorio, scripting... y en los últimos años ha crecido mucho en el mundo Big Data y Data Science, siendo un lenguaje de alto rendimiento donde aplicar algoritmos de manera ágil. Su punto débil es que tiene menor frecuencia de actualización de sus librerías y funcionalidades en comparación con Scala (Spark) o Java. Hay una versión de Apache Spark que se programa con la API de Python, "PySpark".

- **Go:** permite agilizar la transacción en la nube. A pesar de ser un lenguaje muy nuevo (2009), al estar respaldado por Google, es uno de los lenguajes que más rápido se está expandiendo. Tiene buen rendimiento, un manejo de memoria muy eficiente junto con su punto fuerte: la concurrencia, ahorrando costes e inversión en plataformas cloud, memoria y CPUs virtuales.

Entre estas opciones, la más adecuada sería Spark, ya que es el más escalable de los tres.

Otra decisión importante es donde almacenar el clasificador del sistema de reconocimiento. En local, en el dispositivo móvil de cada usuario, lo que supone un problema de seguridad porque el código podría ser accesible para el usuario o en un servidor central. Se propone hacerlo en un servidor

central, de manera que se envíen y procesen los datos en tiempo real. Esto también supone un problema de seguridad ya que la transferencia de datos e información se realiza por la red, pero para ello se propone la securización a través de los siguientes protocolos de red que dificulten el acceso a los datos y garanticen la integridad y privacidad de la información [14].

- **SSL/TLS** integrado al protocolo HTTP formando, por lo tanto, HTTPS, protocolo HTTP bajo una capa intermedia de seguridad que permite encriptar los datos del cliente hasta el punto final de la comunicación.
- Utilización de un **proxy:** para una conexión HTTPS, el proxy tiene la capacidad de realizar la conexión segura con la dirección remota que se está solicitando, y presentar un certificado propio (generalmente auto firmado), teniendo así acceso a todo dato de la comunicación y, por lo tanto, aceptando o bloqueando la solicitud.

Siempre que se trabaja con datos, es útil su monitorización para detectar problemas de manera rápida. Como se explicará en el próximo capítulo, una buena visualización (report) puede ayudar al éxito de un proyecto. En definitiva, es más fácil y rápido ver un dashboard que un log de información. Esto tiene que ir acompañado de una herramienta de Business Intelligence. Se analizan las siguientes [63, 50, 6, 90].

- **Qlik:** dispone de Qlik (Attunity) que admite data streaming con Apache Kafka.
- **Power BI:** cuando se trabaja con datos en tiempo real, solo los almacena en una caché temporal que expira rápidamente, pudiendo visualizarlos utilizando iconos que están optimizados para mostrarlos rápidamente, en tiempo real.
- **Arcadia Data:** permite visualizar streaming data en plataformas para análisis en tiempo real como Apache Kafka (plus Confluent KSQL), Apache Kudu y Apache Solr.

Eligiendo Power BI, ya que permite una mejor conexión a los datos procedentes de bases de datos no relacionales (NoSQL) o timeseries y funciona bien con plataformas en la nube y potentes servicios como Azure Stream Analytics. Qlik compensa la falta de conectores a través de

compatibilidad con todos los servicios de Google y proporcionando una amplia gama de fuentes de datos externas en Qlik Data Market, pero no es relevante para los objetivos de este proyecto.

El esquema de la arquitectura propuesta se muestra en la figura 8.3.

Figura 8.3: Esquema de la arquitectura propuesta.

Capítulo 9

Business Intelligence

Incrementar el tamaño experimental, con más datos y usuarios conlleva más vías de trabajo futuro. Además de implementar una arquitectura escalable, como se ha explicado en el capítulo anterior, es necesario un sistema de monitorización y administración de los resultados, que se va a explicar en el presente capítulo.

El término Business Intelligence (BI) incluye el conjunto de metodologías, aplicaciones y tecnologías que permiten reunir, depurar y transformar datos de los sistemas transaccionales e información desestructurada en información estructurada, para su explotación directa (reporting, análisis OLTP/OLAP, alertas...) o para su análisis y conversión en conocimiento, dando así soporte a la toma de decisiones sobre el negocio [1]. Es decir, es la habilidad para transformar los datos en información, y la información en conocimiento, de forma que se pueda optimizar el proceso de toma de decisiones en los negocios, pudiendo tomar decisiones de manera más rápida.

En el presente proyecto de reconocimiento biométrico interesa tener disponible la información de los intentos de acceso al sistema de los usuarios (intentos de autenticación aceptados y rechazados). La idea es visualizar las medidas acumuladas y que se vayan actualizando en tiempo real, generando avisos si existe una acumulación de intentos de acceso rechazados, que pueda detectar un posible ataque al sistema o un error en el funcionamiento del sistema de reconocimiento. Por otro lado, sería interesante gestionar un sistema de administración que muestre como ha funcionado el sistema de reconocimiento en un día concreto.

Como resultado final se desea construir cuadros de mandos compuestos por una combinación de representaciones gráficas interactivas, paneles

y testigos que permitan el seguimiento de los indicadores que se van a explicar a lo largo de este capítulo.

9.1. Requisitos

Los requisitos son aquellas cosas que se van a solicitar al sistema, es decir, los datos que se van a analizar y cómo se van a evaluar a partir de la información disponible. En este caso, se tiene un único requisito, un sistema de monitorización en tiempo real que permita detectar problemas de manera rápida y eficiente, en la que se quieren identificar los siguientes datos.

- Número de intentos de acceso exitosos al sistema (reconocimiento aceptado).
- Número de intentos de acceso fallidos al sistema (reconocimiento rechazado).

O bien, gestionar un sistema de administración que muestre como ha funcionado el sistema de reconocimiento en un periodo de tiempo, donde además de los datos anteriores, se podría buscar respuesta a los siguientes datos.

- Porcentaje de reconocimiento correcto (intentos de acceso bien reconocidos)
- Porcentaje de reconocimiento incorrecto (intentos de acceso mal reconocidos).

Pero en un sistema real no se conocen estos datos. Sería necesario hacerlo a posteriori, a través de reclamaciones de los usuarios (para el caso de falsas aceptaciones) o permitiendo una alternativa de acceso al sistema, en caso de rechazo (falsos rechazos). Un ejemplo de alternativa de acceso sería una serie de preguntas personales previamente guardadas por el usuario. Consiguiendo conocer, de esta manera, las falsas aceptaciones (reconocimiento incorrecto). Y suponiendo que, salvo reclamaciones, el resto de los intentos han sido correctos. De esta manera, se consigue una idea aproximada del funcionamiento del sistema.

Los datos pueden ser evaluados por la siguiente información:

- **Usuario** que intenta acceder al sistema de reconocimiento.
- **Tipo de dispositivo:** modelo de dispositivo con el que se ha realizado el intento de acceso.
- **Tipo de sensor** del que se están adquiriendo los datos.
- **Fecha:** día, mes, año, hora, minuto y segundo del intento de acceso.
- **Lugar:** ciudad desde donde se ha realizado el acceso.

No obstante, no se tiene información real disponible. Se está planteando como trabajo futuro.

9.2. Indicadores

En base al requisito planteado, se tienen los siguientes indicadores, que representan los valores clave que se pretenden alcanzar con el proceso de negocio.

Número de intentos de acceso aceptados al sistema

Se considera que un intento de acceso es aceptado cuando un usuario intenta realizar un acceso y el sistema de reconocimiento biométrico verifica que el usuario es quien dice ser y, por tanto, permite su acceso. Se explica en la tabla 9.1

Número de intentos de acceso aceptados al sistema (AccesoAceptado)	
Definición	Indicador que calcula el número total de intentos de acceso aceptados al sistema.
Cálculo	AccesoAceptado = contar número de intentos de acceso aceptados
Fuente de datos	NA

Tabla 9.1: Indicador BI - Número de intentos de acceso aceptados al sistema

Número de intentos de acceso rechazados al sistema

Se considera que un intento de acceso es rechazado cuando un usuario intenta realizar un acceso y el sistema de reconocimiento biométrico falla, indicando que el usuario no es quien dice ser y, por tanto, no permite su acceso. Se explica en la tabla 9.2

Número de intentos de acceso rechazados al sistema (AccesoRechazado)	
Definición	Indicador que calcula el número total de accesos rechazados al sistema.
Cálculo	AccesoRechazado = contar número de intentos de acceso rechazados
Fuente de datos	NA

Tabla 9.2: Indicador BI - Número de intentos de acceso rechazados al sistema

Porcentaje de reconocimiento correcto

Tanto por ciento de intentos de acceso que han sido reconocidos de forma correcta. Se pueden producir de dos formas.

- El sistema verifica que el usuario es quien dice ser y es correcto, se trata de un usuario auténtico (verdadero positivo).
- El sistema rechaza la conexión del usuario y es correcto, se trata de un impostor (negativo verdadero).

Se explica en la tabla 9.3

Porcentaje de reconocimiento correcto (ReconocCorrecto)	
Definición	Indicador que relaciona el número de intentos de acceso al sistema respecto a aquellos que son reconocidos de manera correcta.
Cálculo	ReconocCorrecto = $\frac{\#Accesos\ bien\ reconocidos}{\#total\ de\ accesos} \cdot 100$
Excepciones	El número de intentos de acceso correctos es un valor aproximado que se conoce a posteriori, suponiendo que un negativo verdadero fallará en un reconocimiento alternativo mediante preguntas.
Fuente de datos	NA

Tabla 9.3: Indicador BI - Porcentaje de reconocimiento correcto

Porcentaje de reconocimiento incorrecto

Tanto por ciento de intentos de acceso que no se han reconocido de manera correcta. Se pueden producir de dos formas.

- El sistema verifica que el usuario es quien dice ser y no es correcto, se trata de un usuario impostor (falso positivo).

- El sistema rechaza la conexión del usuario y no es correcto, se trata de un usuario auténtico mal reconocido al que se le impide el acceso (falso negativo).

Se explica en la tabla 9.4

Porcentaje de reconocimiento incorrecto (ReconocIncorrecto)	
Definición	Indicador que relaciona el número de intentos de acceso al sistema respecto a aquellos que son reconocidos de manera incorrecta.
Cálculo	ReconocIncorrecto = $\frac{\#Accesos\ mal\ reconocidos}{\#total\ de\ accesos} \cdot 100$
Excepciones	El número de intentos de acceso mal reconocidos es un valor aproximado que se conoce a posteriori, suponiendo que un falso negativo será aquel que consiga acceso en un reconocimiento alternativo mediante preguntas. Y un falso positivo será reclamado al sistema por el usuario auténtico.
Fuente de datos	NA

Tabla 9.4: Indicador BI - Porcentaje de reconocimiento incorrecto

9.3. Identificación de hechos y dimensiones

Los hechos representan *¿qué queremos medir?* y las dimensiones *¿cómo lo queremos medir?*. De manera que en torno al proceso de negocio Sistema de Administración, se va a tener una única tabla de Hechos que se llamará "Administración" y englobará las siguientes variables.

- **Score:** se almacena a modo informativo para ir ajustando el umbral del sistema de reconocimiento con el objetivo de mejorar los resultados.
- **Resultado del intento (ResultadoIntento):** aceptado o rechazado.
- **Identidad correcta (IdentidadCorrecta):** si, no o nulo, que equivale a que no se conoce. Indicará si el usuario ha realizado un intento de acceso alternativo mediante preguntas y ha acertado, después de haber sido rechazado por el sistema de reconocimiento (IdentidadCorrecta=no) o ha fallado (IdentidadCorrecta=si).

Las dimensiones son las mencionadas en los requisitos.

- **Usuario:** contiene atributos sobre la identidad del usuario.

  - Identificador único del usuario (IdUsuario).
  - Nombre del usuario.
  - Apellidos del usuario.
  - Sexo: Hombre o Mujer.
  - Edad del usuario.
  - Mano de uso habitual del dispositivo, a lo que se hace referencia como mano dominante.

- **Tipo de dispositivo:** contiene atributos para la identificación del dispositivo de reconocimiento utilizado por el usuario, el cual llevará ubicado en su muñeca.

  - Identificador del dispositivo (IdDispositivo).
  - Nombre completo del dispositivo.

- **Tipo de sensor:** contiene atributos para la identificación del sensor usado en el renocimiento.

  - Identificador del sensor (IdSensor).
  - Nombre completo del sensor.

- **Fecha:** contiene atributos para conocer el momento temporal en el que se ha realizado el intento de acceso.

  - Día.
  - Mes.
  - Año.
  - Hora.
  - Minuto.
  - Segundo.

- **Lugar:** contiene atributos para conocer la ubicación desde la cual se ha intentado realizar el acceso al sistema.

  - Identificador del lugar desde donde se ha realizado el acceso (IdLugar).
  - Ciudad desde la que se ha realizado el acceso.
  - País desde el cual se ha realizado el acceso.

9.4. Bus de Dimensiones

Una vez definidos los hechos y las dimensiones se han relacionado qué dimensiones aplican a cada uno de los hechos, conocido en BI como bus de dimensiones. Se muestra en la tabla 9.5, donde se puede ver que todas las dimensiones están relacionadas con cada uno de los hechos.

¿Qué? \¿Cómo?	Usuario	Dispositivo	Sensor	Fecha	Lugar
# Conexiones exitosas	X	X	X	X	X
# Conexiones fallidas	X	X	X	X	X
% reconocimiento correcto	X	X	X	X	X
% reconocimiento incorrecto	X	X	X	X	X

Tabla 9.5: BI - Bus de dimensiones

9.5. Esquema del Almacén de Datos

Se va a crear un único DataMart para el único proceso de negocio que se tiene: el sistema de gestión de los accesos al sistema de reconocimiento (Administración). Se van a distinguir los hechos con el prefijo "Fact" y las dimensiones con "Dim". Su esquema se muestra en la figura 9.1.

9.6. Boceto del Cuadro de Mandos

Se plantea la implementación de dos cuadros de mandos. Un **sistema de monitorización**, cuyo boceto se muestra en la figura 9.2. Se propone una actualización periódica de, por ejemplo, cada 5 minutos y muestra la información acumulada de las últimas 24 horas. No obstante, estos parámetros se pueden modificar con los selectores de la parte inferior. Se han puesto en esa zona para que ocupen poco espacio y no molesten en la visualización, ya que no se van a usar de manera habitual. Este cuadro de mando tiene como objetivo informar de manera rápida de cualquier problema en el sistema de reconocimiento: detectar ataques, funcionamiento anómalo, etc. Las visualizaciones elegidas de arriba hacia abajo han sido las siguientes:

- Un velocímetro que marca el porcentaje de intentos de acceso rechazados (va a ser uno de nuestros principales KPIs, key performance indicator) que ha habido en el sistema en las últimas 24 horas. Aparece el número en la parte central, pero también hay que fijarse en

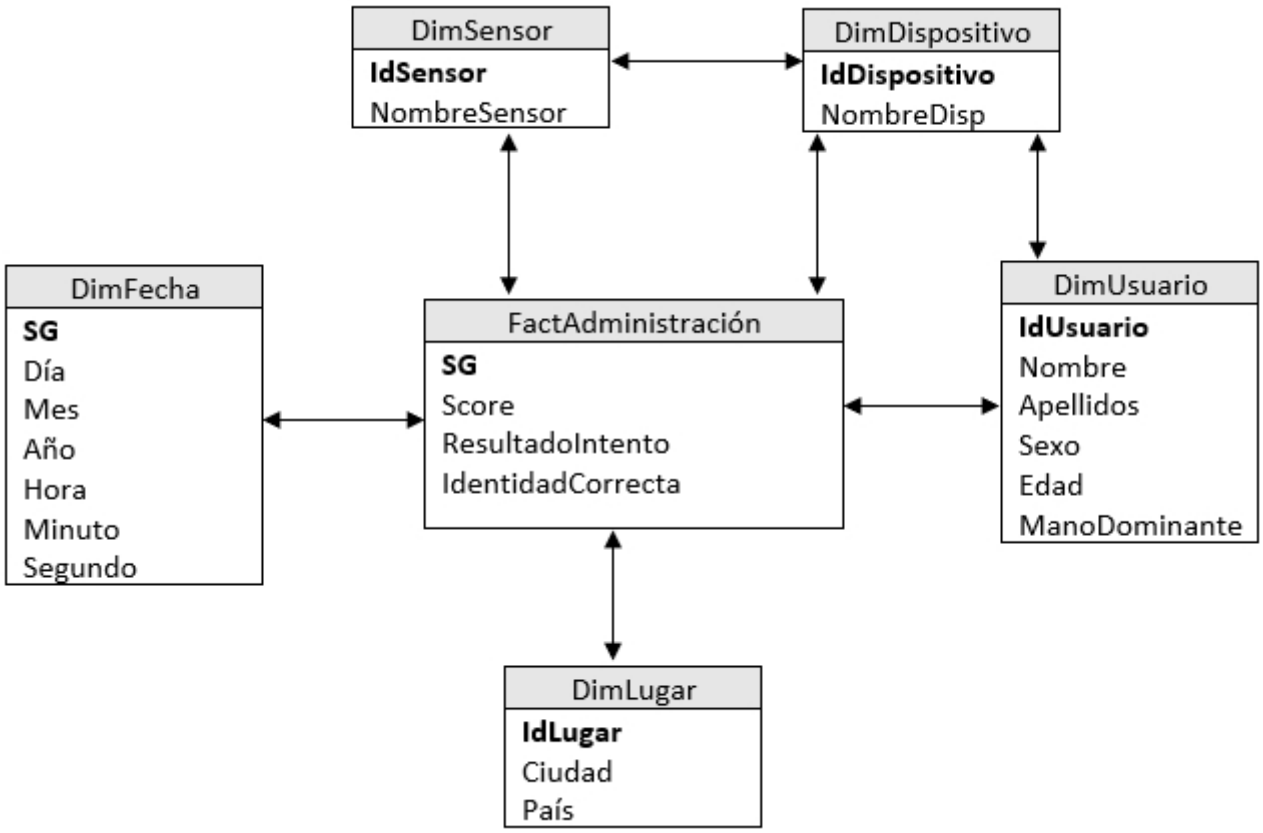


Figura 9.1: Esquema del almacén de datos - Business Intelligence

la zona hacia la que apunta la aguja. Un valor entre 0 y 10 %, color verde, es señal de un funcionamiento normal, en el que la mayor parte de los usuarios intentan autenticarse y consiguen acceso y unos pocos no, bien porque no los reconoce correctamente o porque son impostores. Un porcentaje entre el 10 y el 30 %, color amarillo, sería indicativo de alerta, una alta cantidad de intentos están siendo rechazados y un valor por encima del 30 %, color rojo, sería indicativo de problemas y de que hay que tomar medidas.

- Dos etiquetas (labels) que muestran el número total de intentos de acceso aceptados y rechazados en el sistema (lo otros dos KPIs de este cuadro de mandos), así como una comparativa (en porcentaje) de si esos números han aumentado o disminuido con respecto a las 24 horas anteriores a las mostradas. Esto permite hacer un seguimiento del sistema.

- Un mapa mundial que muestra el porcentaje de rechazo en cada uno de los países. En caso de que el velocímetro indicará alerta o problemas permitiría ver rápidamente en qué países existen esos problemas y actuar en consecuencia, por ejemplo, restringiéndoles el acceso al sistema. El color del punto muestra el número de rechazos utilizando el verde para poca cantidad y el rojo para mucha. El tamaño del punto se utiliza para indicar la existencia de valores anómalos con respecto a las 24 horas anteriores. Se trata de un gráfico interactivo, donde al pasar el cursor por cada país se muestra el número de intentos de acceso aceptados, rechazados, el total y un ranking de la posición en la que se encuentra el país con respecto a problemas de rechazos, es decir, la primera posición será ocupada por el país con el mayor porcentaje de rechazo.
- Un segundo mapa mundial que muestra el número total de intentos de acceso al sistema en cada país. El tamaño del punto se utiliza para indicar una mayor o menor cantidad de intentos de acceso al sistema de reconocimiento. Es un gráfico interactivo, que muestra la misma información que el anterior, a diferencia de que ahora la primera posición del ranking es ocupada por el país que tiene un mayor número de intentos de accesos totales. El objetivo es detectar sobrecargas de acceso en lugares concretos.

La paleta de colores utilizada ha sido la siguiente:

- Verde para indicar porcentajes bajos de rechazo. También para mostrar en las labels un aumento en el número de intentos de reconocimiento aceptados en el sistema y una disminución en el número de intentos rechazados. Si el número de intentos aceptados disminuyera o los rechazados aumentarán con respecto a las 24 horas anteriores, las flechas y el porcentaje de los labels se mostrarían en color rojo.
- Amarillo para indicar alerta respecto al porcentaje de rechazo.
- Rojo para indicar problemas.
- Azul para mostrar el número de intentos totales.

Por otro lado, en la figura 9.3 se muestra el boceto para el **sistema de administración** del sistema de reconocimiento. Tiene una serie de selectores, en la parte derecha, para cada una de las dimensiones.

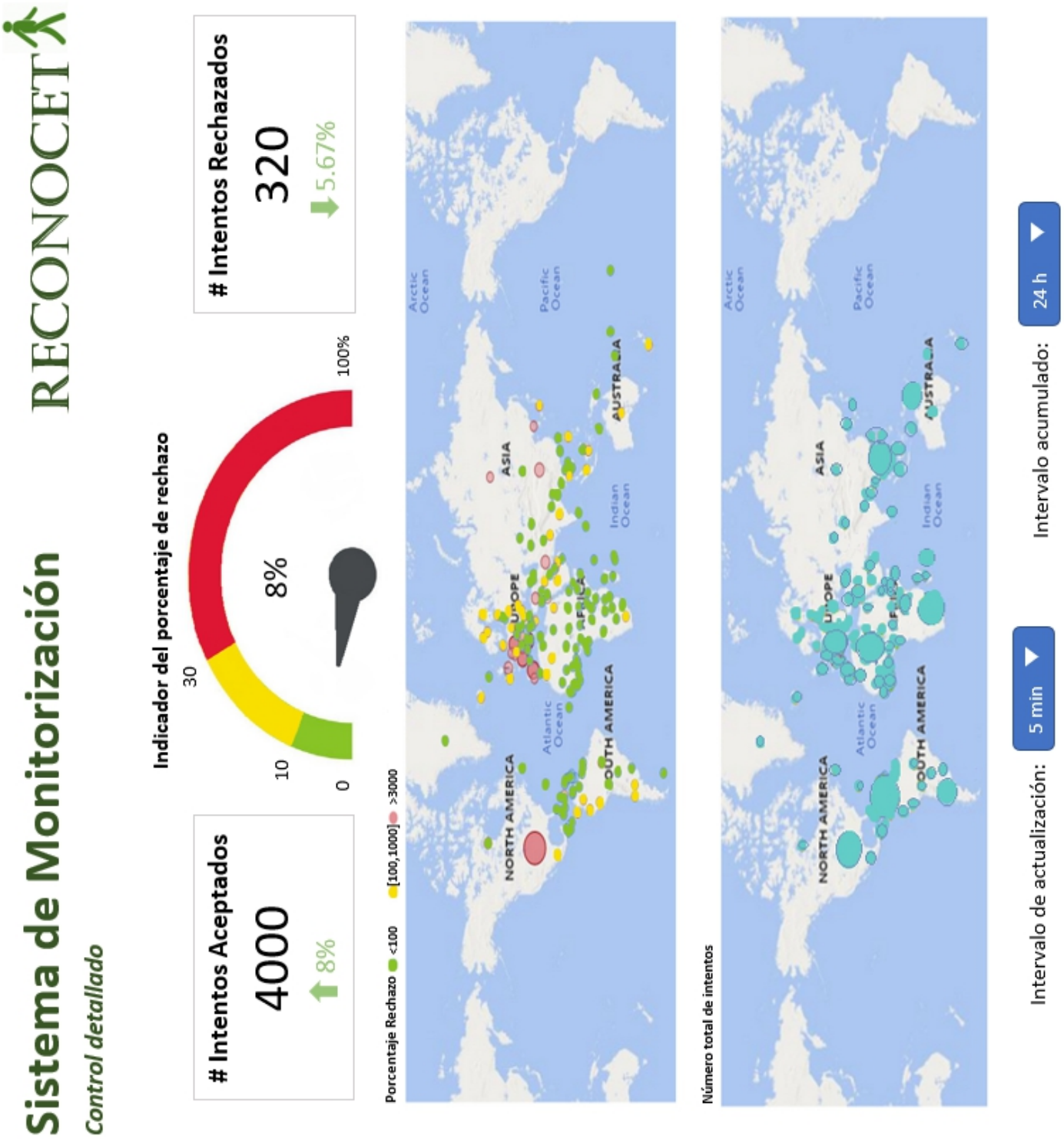


Figura 9.2: Boceto del cuadro de mando del Sistema de Monitorización - Business Intelligence

- Periodo para el que se muestran los datos, comprendido entre una fecha de inicio y una fecha de fin.
- Se puede ver la información de todos los usuarios (por defecto) o de algunos en concreto. Permitirá selección múltiple.
- Dispositivo para el que se muestran los resultados. Permitirá selección múltiple.
- Sensor con el que se muestran los resultados. Permitirá selección múltiple.
- Lugar para el que se muestran los resultados, Permitirá, si se desea, hacer selección múltiple.

Este cuadro de mando tiene como objetivo mostrar los resultados, o bien diarios o por periodos de tiempo, y poder hacer un seguimiento del funcionamiento del sistema de reconocimiento, pudiendo comparar unos días/periodos con otros y ver la evolución. Las visualizaciones elegidas, de arriba hacia abajo, han sido las siguientes:

- Gráfico de barras apiladas que muestra el conteo, en cada hora, del número de intentos de acceso que han sido aceptados (en color verde) o rechazados (en color rojo claro). Permite ver las horas de mayor uso del sistema, así como aquellas en las que se produce un mayor número de rechazos. Es un gráfico interactivo donde al pasar el cursor se muestran los valores numéricos. Si estuvieran seleccionados varios lugares, al pasar el cursor se verían los valores desagregados por países.
- Cuatro etiquetas (labels) que muestran el número total de intentos de acceso aceptados y rechazados en el sistema, así como el valor aproximado de los que se conoce que han sido bien o mal reconocidos. Igual que en el cuadro de mandos anterior, en la parte inferior de cada etiqueta se muestra un porcentaje con la comparativa de los resultados, respecto a un periodo anterior equivalente al que se está mostrando. Permite hacer un seguimiento del funcionamiento del sistema y ver cómo va evolucionando.
- Un gráfico circular que muestra de manera gráfica el porcentaje de reconocimiento correcto e incorrecto, según la información que se tenga disponible. Es un gráfico interactivo donde al pasar el cursor se muestran los valores numéricos, desagregados por ubicación.

La paleta de colores utilizada contiene el color verde para indicar intentos de acceso aceptados y reconocimiento correcto y el color rojo claro para indicar lo contrario: intentos de acceso rechazados y reconocimiento incorrecto.

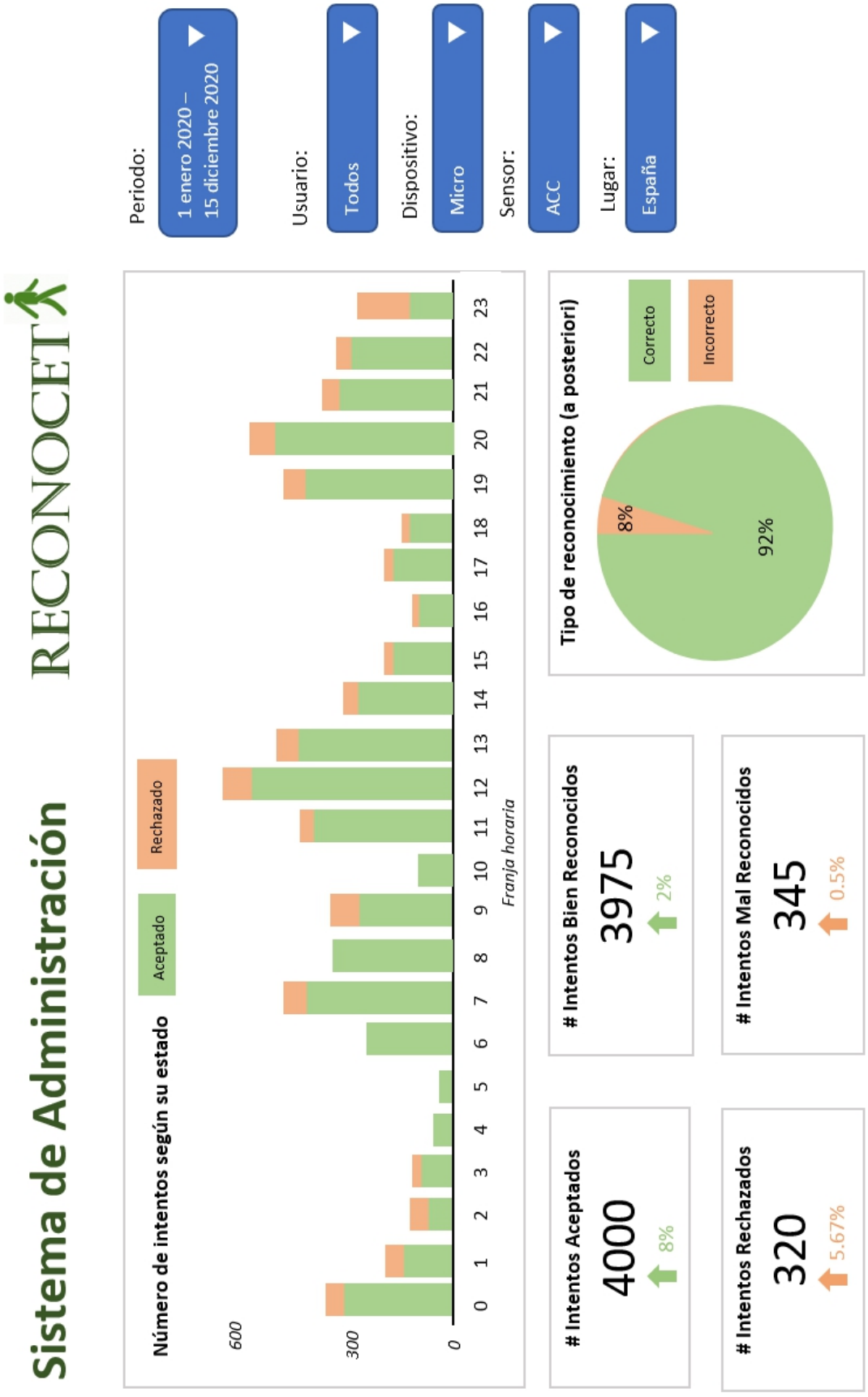


Figura 9.3: Boceto del cuadro de mando del Sistema de Administración - Business Intelligence

Capítulo 10

Ciberseguridad

La necesidad de seguridad se ha disparado con el auge de Internet y de los dispositivos electrónicos, las compras online, las transacciones bancarias vía web, etc. La biometría se erige como el futuro de los sistemas de seguridad y su desarrollo en los últimos años ha experimentado un crecimiento geométrico respecto a otras tecnologías de seguridad.

Algunos lugares donde se han empleado sistemas biométricos de seguridad son los siguientes [20]:

- La huella digilar se ha utilizado para las siguientes acciones:
  - Control de acceso a áreas en el Pentágono.
  - Acceso a computadoras de redes financieras en Italia.
  - Automated Banking Terminal en Australia.
  - Aduana e inmigración en Ámsterdam.
  - Control de acceso Expo'92 Sevilla.
- Mientras que el reconocimiento de la mano se ha utilizado en las siguientes acciones.
  - San Francisco Intl Airport para el control de acceso de las operaciones.
  - Lotus para los visitantes fuera de áreas reservadas.
  - University of Georgia para el control de alimentos consumidos.
  - Carcel en Jessup.

- Aeropuerto Kenneddy y Newark para la inspección automática del pasaporte y el control de personas que se registran como pasajeros frecuentes.
- Cámara de Diputados y Senado en Colombia para evitar el fraude en las votaciones.

La biometría abordada en este trabajo es el reconocimiento de la forma de andar de una persona, que se graba y se somete a un proceso analítico que genera una plantilla biométrica única derivada de dicho comportamiento. Aparece reconocida como tecnología biométrica de comportamiento en la guía de tecnologías biométricas aplicadas a la ciberseguridad, publicada por el Incibe [36], donde se afirma que esta tecnología todavía está en desarrollo y no ha alcanzado aún los niveles de rendimiento necesarios para ser implantada como tecnología biométrica.

Los smartphones y wearables ofrecen cada vez más funcionalidades y nos sirven para acceder a un creciente número de servicios, lo que les convierte en una potencial fuente de amenazas. Los dispositivos biométricos ofrecen un excelente nivel de seguridad en el proceso de autenticación, muy superior a la ofrecida por los códigos PIN (Personal Identification Number), sin embargo, la mayor parte de estas tecnologías tienen una vulnerabilidad consistente en la aceptación de muestras biométricas falsas como fotografías, dedos de goma, etc. Se trata de una vulnerabilidad muy importante puesto que permiten una autenticación falsa. Estos problemas pueden ser resueltos incluyendo módulos de detección de vida en el proceso de verificación para asegurarse de que la muestra presentada pertenece a una persona viva y la detección de vida es diferente en cada tecnología biométrica: a través de la cámara del dispositivo o mediante una serie de preguntas personales. Existen aplicaciones que se están utilizando de manera complementaria a la biometría para aumentar la seguridad.

- **Combinación de biometría con NFS:** con la tecnología NFC (Near Field Communication) integrada en los smartphones se pueden realizar pagos y existen aplicaciones que combinan esta tecnología con la biometría para comprobar la identidad del usuario. Un uso es en entornos hospitalarios para administrar medicamentos. Con NFC, en una pulsera o similar, se determina qué medicamentos administrar y con biometría se identifica al paciente para evitar errores.
- **Match on card:** Es una combinación de la biometría y de tarjetas inteligentes para proporcionar una autenticación de doble factor

(algo que tienes y algo que eres). En el chip de la tarjeta inteligente se almacena de forma segura el patrón biométrico (generalmente la huella dactilar).

El Incibe [36] reconoce que la implantación de tecnologías biométricas conlleva un conjunto de ventajas tanto para entidades públicas y privadas como para los usuarios finales.

- **Aumento de seguridad en el control de accesos:** en su uso para la autenticación de empleados, garantizando que la persona es quien dice ser, es decir, que los rasgos biométricos se encuentran exclusivamente ligados a su legítimo usuario. Mediante el robo de credenciales o tarjetas identificativas, un individuo puede acceder a zonas restringidas o realizar operaciones no permitidas, inculpando a terceros. Asimismo, es posible que estas credenciales se compartan voluntariamente entre empleados. A través de la implementación de sistemas biométricos, se aumenta la seguridad reduciendo la probabilidad de que alguien no autorizado acceda a zonas o a aplicaciones restringidas.

- **Mejora de la imagen corporativa de las organizaciones:** La implantación de tecnologías biométricas contribuye a que una empresa sea más eficiente, más segura y reduzca el fraude interno. Produciéndose una importante mejora en la opinión general sobre la compañía, asociando a la entidad con la innovación, la inversión en investigación y desarrollo y la apuesta por tecnología puntera.

- **Posibilidad de tramitaciones remotas:** Es posible emplear técnicas biométricas como forma de verificación en operaciones no presenciales de forma altamente fiable, pudiendo superar a las firmas electrónicas actuales. De esta forma se pueden reducir los traslados y trámites innecesarios e inconvenientes para el usuario final.

- **Aumento de la privacidad:** Utilizando técnicas biométricas se incrementa la seguridad en la transmisión de los datos de carácter personal de los clientes al cifrarlos utilizando una clave única y personal del propio cliente. Es la consecuencia de la notable dificultad que supone la falsificación de rasgos biométricos con el objetivo de acceder a la información personal de un usuario final o cliente.

En una comparación directa entre las tecnologías biométricas y las técnicas de identificación y autenticación tradicionales, destacan los siguientes beneficios e inconvenientes sobre el uso de biometría.

- **Necesidad de secreto:** las contraseñas han de ocultarse y las tarjetas no deben de estar al alcance de terceros, mientras que la biometría no requiere de estas medidas de protección que son exclusivamente dependientes del usuario.
- **Posibilidad de robo:** las tarjetas y contraseñas pueden ser robadas. Sin embargo, robar un rasgo biométrico es extremadamente más complejo.
- **Posibilidad de pérdida:** las contraseñas son fácilmente olvidables y las tarjetas se pueden perder. Los rasgos biométricos permanecen invariables salvo en contadas excepciones y siempre están con el sujeto a quien identifican.
- **Registro inicial y posibilidad de regeneración:** la facilidad con la que se puede enviar una contraseña o tarjeta nueva contrasta con la complejidad que supone el registro en un sistema biométrico, ya que requiere de la presencia física del individuo en esta fase. Hay que añadir que los rasgos biométricos son por definición limitados, mientras que la generación de contraseñas es ilimitada, lo cual es una ventaja. Sin embargo, la biometría propuesta en este trabajo reduce este problema.
- **Proceso de comparación:** la comparación de dos contraseñas es un proceso sencillo. Sin embargo, comparar dos rasgos biométricos requiere de mayor capacidad computacional.
- **Comodidad del usuario:** el usuario ha de memorizar una o múltiples contraseñas y, en el caso de que use una tarjeta, ha de llevarse siempre consigo. Utilizando tecnología biométrica no se necesita realizar estos esfuerzos.
- **Vulnerabilidad ante el espionaje:** una discreta vigilancia de nuestra actividad podría servir para obtener nuestra contraseña o robar nuestra tarjeta. Ese método es más dificil ante los sistemas biométricos.
- **Vulnerabilidad a un ataque por fuerza bruta:** las contraseñas tienen una longitud de varios caracteres. Por su parte, una muestra biométrica emplea cientos de bytes, lo que complica mucho los ataques por fuerza bruta.

- **Medidas de prevención:** los ataques contra sistemas protegidos por contraseña o tarjeta se producen desde hace años, y las medidas de prevención contra ellos ya se encuentran maduras. Por el contrario, los ataques a los sistemas biométricos son un área en la que estas medidas de prevención aún se están generando en estos momentos.
- **Autenticación de usuarios "reales":** la autenticación de usuarios mediante contraseña o tarjeta y su efectividad, dependen absolutamente de la voluntad del usuario a la hora de hacerlas personales e intransferibles. La biometría está altamente relacionada con el propio usuario pues no puede ser prestada ni compartida.
- **Coste de implantación:** en el momento de la implantación, el hecho de instaurar un sistema de contraseñas tiene un coste bajo, mientras que en el caso de un sistema basado en muestras biométricas es más costoso.
- **Coste de mantenimiento:** el coste de mantenimiento de un sistema biométrico, una vez está implantado con éxito, es menor al de un sistema de contraseña o tarjeta ya que no conlleva gastos de gestión asociados a la pérdida u olvido de credenciales.

Sin embargo, con el reconocimiento de la forma de andar que aquí se propone, aunque tiene inconvenientes relacionados con la existencia de factores externos (condiciones meteorológicas, ropa, calzado...) e internos (estado físico, mental, una enfermedad...), se trata de un sistema discreto con fácil registro inicial y posibilidad de regeneración, difícil de robar o falsificar. Aunque evidentemente, requiere de más investigación, mejorando sus requisitos computacionales y su coste de implantación para ser capaces de satisfacer cada vez mayores demandas de carga de trabajo.

Algunas de las amenazas y vulnerabilidades reconocidas por el Incibe son específicas y otras compartidas con las demás tecnologías y técnicas de identificación [36].

- **Pérdida o robo de información biométrica:** el robo de información es especialmente sensible en el caso de la biometría al tratarse de información exclusiva y extremadamente ligada al individuo, por lo que el robo de la misma supondría un incidente de seguridad grave.
- **Suplantación de identidad:** Se trata del uso de información biométrica robada o falsificada con el propósito de acceder a espacios o

aplicaciones restringidas, falsificar el control de presencia, enmascarar o suplantar una personalidad, etc. Es de especial gravedad cuando se utiliza para cometer un crimen ya que su repudio resulta complicado.

- **Sabotaje:** Pueden darse ataques al sensor de forma consciente para tratar de impedir su funcionamiento. Frecuentemente, la causa de estos ataques es una expresión del desacuerdo o descontento con la implantación de biometría precisamente debido a la alta fiabilidad que ofrece el sistema a la hora evitar conductas fraudulentas y accesos no autorizados.

- **Incumplimiento de la normativa de protección de datos personales:** Los rasgos biométricos se consideran datos de carácter personal a todos los efectos legales por lo que su tratamiento se encuentra sometido al cumplimiento de las distintas exigencias de carácter jurídico, técnico, físico y organizativo previstas, principalmente por la Ley de Protección de Datos de Carácter Personal (LOPD) y por su normativa de desarrollo (RDLOPD). Los datos biométricos se han considerado, con carácter general "de nivel básico", siendo equiparables a una simple dirección o un número de teléfono, pero siempre es aconsejable tratar estos datos con la máxima cautela y protección posible, ya que en muchos casos el usuario final los percibe como de una alta sensibilidad, precisamente por tratarse de rasgos intrínsecamente ligados a su persona. La LOPD establece una serie de obligaciones y principios de obligado cumplimiento para todas aquellas entidades o empresas, tanto del sector público como privado, que traten datos de carácter personal para el desarrollo de su actividad. Un tratamiento inadecuado puede derivar en riesgos para la privacidad de los datos personales almacenados además de los derivados de la infracción e incumplimiento de la normativa vigente.

- **Idoneidad de la implantación:** Existe el riesgo de creer erróneamente que un sistema biométrico garantiza la seguridad total y que es la solución a cualquier problema de seguridad. La implantación de tecnologías biométricas supone un alto coste económico y una implicación de personal específicamente dedicado a ello durante la fase inicial. Por esta razón es necesario realizar un análisis previo para evaluar la verdadera necesidad, escenario adecuado de implantación y el beneficio logrado frente al coste incurrido, ya que es posible que ésta no sea la solución más adecuada para todos los casos.

- **Calidad de la tecnología:** Si la calidad de la tecnología implantada no alcanza los niveles recomendables, podría acarrear graves brechas de seguridad así como un deterioro notable de la percepción de las tecnologías debido a su mal funcionamiento. Los elementos que se deben tener en cuenta al respecto son: la calidad del sensor, la eficiencia del algoritmo de comparación, la encriptación del almacenamiento de muestras obtenidas y la interoperabilidad con otros sistemas.

- **Incidencias con el sistema:** Como todo sistema electrónico e informático, los sistemas de autenticación biométricos son susceptibles de fallos eléctricos, caída de las líneas de comunicación, del propio sistema o de los sistemas de soporte (por ejemplo suministro eléctrico o sistema de comunicaciones), ataques informáticos, etc.

- **Indisponibilidad del sensor:** Si el acceso o la autenticación se realizan exclusivamente mediante biometría, es decir, no existe un método alternativo como puede ser el uso de contraseñas o tarjetas personales, el fallo o ausencia del dispositivo de adquisición de muestras supone la imposibilidad de autenticación o acceso. Un ejemplo de esta situación es un empleado que tenga que acceder de forma urgente al correo electrónico desde fuera de la oficina mediante huella dactilar pero no disponga de sensor en su ordenador o en nuestro caso, el acelerómetro deje de funcionar en el wearable.

- **Variación involuntaria en los rasgos biométricos:** Los cambios en los rasgos biométricos, como variaciones de la voz, vello facial, el peinado, cambios de peso también suceden de forma natural. En estos casos, el usuario no tiene intención de engañar al sistema. No obstante, estos cambios pueden dificultar el proceso de autenticación y generar, incluso, una percepción negativa para el usuario.

- **Experiencia de uso negativa (usabilidad):** Un mal uso involuntario del sensor realizado por una persona sin los conocimientos adecuados puede tener como consecuencia la desconfianza del usuario en la tecnología y el aumento de la tasa de error de la misma.

- **Falta de aceptación cultural:** Esta amenaza aparece en determinados grupos demográficos cuyas normas sociales o religiosas no favorecen la toma de muestras en determinadas técnicas. Por ejemplo, en algunas culturas el reconocimiento facial no es válido ya que no todos los ciudadanos llevan el rostro al descubierto; en otras la

lectura de la huella dactilar se considera una práctica antihigiénica, etc.

A todo lo anterior, se unen vulnerabilidades relacionadas con una posible calidad baja de los dispositivos de captura, una ubicación o colocación inadecuada del dispositivo de captura, la falta de conocimientos técnicos del usuario, la escasa concienciación en materia de seguridad, la percepción de ausencia de privacidad por parte de los usuarios, etc.

De todas formas, con el objetivo de reducir los riesgos asociados al empleo de biometría y hacer una adecuada gestión de los mismos, han de aplicarse una serie de controles mitigantes y buenas prácticas de seguridad. El Incibe propone los siguientes [36]:

- **Reforzar la seguridad del sistema:** La seguridad es fundamental en todos los elementos de un sistema biométrico. Es por ello que se debe garantizar la privacidad y evitar accesos no autorizados a la base de datos en que se guardan los registros biométricos. En la sección 8.2 se propone la utilización de un proxy en conexiones de tipo HTTPS.

- **Almacenamiento de muestras:** En el proceso de registro previo al uso de tecnologías biométricas han de almacenarse las muestras aportadas por los usuarios. Así, existe la posibilidad de almacenar una parte de la muestra o multitud de referencias en lugar de la muestra íntegra. Esto se hace para prevenir su utilización fraudulenta en caso de pérdida o robo. Un ejemplo es el almacenamiento de minucias de huellas dactilares, en lugar de la huella completa. De esta forma resulta imposible la recreación de la huella a partir de una minucia, en caso de que esta sea robada.

- **Autenticación de doble factor:** Con el objetivo de evitar el fraude se recomienda el uso de dos factores en el proceso de autenticación. Para ello se puede utilizar biometría bimodal (dos técnicas biométricas diferentes, por ejemplo huella dactilar e iris) o combinar la biometría con el uso de contraseña y/o tarjetas de identificación.

- **Realizar una buena adaptación:** No todas las empresas son iguales, por ello la adaptación a las circunstancias de cada caso es esencial para evitar futuros problemas. Por ejemplo, si una empresa va a incorporar un control de accesos en base a la huella dactilar y cuenta con empleados que realizan trabajos manuales o utilizan productos

abrasivos, puede ser aconsejable registrar las huellas de la mano que menos utilicen (izquierda en el caso de los diestros y derecha en el caso de los zurdos) y de los dedos que menos se utilicen (generalmente anular y meñique). Con esta simple adaptación, se podrán evitar en buena medida futuros problemas relacionados con cortes o deterioros en la huella.

- **Adquisición de tecnología de calidad:** La obtención de muestras adecuadas y la realización de comparativas fiables es importante para evitar falsos positivos, falsos negativos y altas tasas de error, y esto depende en gran medida de la calidad y fiabilidad de los sistemas utilizados. Una elección adecuada previene el fraude y la reticencia de los usuarios.

- **Formación de los usuarios:** Un factor clave en el éxito de las tecnologías biométricas es que sus usuarios las utilicen correctamente. Para la consecución de este objetivo se puede ofrecer una fase inicial de formación que, por norma general, no será excesivamente larga y en la que se informe de lo que son las tecnologías biométricas y se aporten unas breves instrucciones y recomendaciones sobre su uso. En este sentido, los acuerdos con los representantes de los trabajadores previos a la implantación de las tecnologías favorecen este proceso.

- **Cumplimiento normativo:** A la hora de implantar este tipo de tecnologías biométricas aplicadas a la seguridad, resulta de vital importancia el evaluar previamente las ventajas e inconvenientes posibles que, desde un punto de vista jurídico, puede suponer la implantación de dicho sistema en relación con la vida privada de las personas afectadas, así como tener en cuenta posibles sistemas o soluciones alternativas que puedan suponer una menor intrusión contra los derechos de los interesados. Teniendo en cuenta que los datos biométricos que, en su caso, pudiesen ser sometidos a tratamiento a través de dichos sistemas deberían ser siempre adecuados, pertinentes y no excesivos en comparación con la finalidad del proceso (por ejemplo: control horario, control de accesos, control de presencia, etc.).

Por otro lado, a nivel de la Unión Europea (EU), producidas por el Grupo de Expertos de Alto Nivel sobre Inteligencia Artificial (AI HLEG), también existen guías de pautas éticas para hacer el uso de la Inteligencia Artificial más responsable, intentando evitar problemas. En ella se reconoce

el enorme impacto positivo que la Inteligencia Artificial (AI) tiene a nivel mundial, tanto comercial como socialmente, siendo una tecnología tanto transformadora como disruptiva que ha ido evolucionando en los últimos años produciendo enormes cantidades de datos digitales, creando una importante innovación científica y de ingeniería. Aseguran que la AI continuará impactando a la sociedad y a los ciudadanos de una manera que aún no podemos imaginar. Por ello, consideran importante que se preste la debida atención a garantizar un entendimiento y un compromiso para construir una AI digna de confianza y han redactado las directrices para que esto sea así, asegurando el propósito ético. Y aunque, afirman que la AI puede provocar daños no intencionados, han desarrollado un marco para implementar la AI confiable, ofreciendo una orientación concreta para su logro, proponiendo métodos técnicos y no técnicos de ayuda para su realización e implementación [56]. Pudiendo ser la AI aplicado a la biometría.

Por tanto, el uso de sistemas biométricos, en especial, el reconocimiento a través de la forma de andar, se trata de un mercado controvertido donde no todo es positivo y que aún requiere de mucho desarrollo e investigación, antes de ser implantado. Y aunque tenga inconvenientes, un estudio de la empresa PWC sobre el uso de "wearables" en el trabajo, indica que alrededor del 40 % de los empleados aseguran no confiar en que la empresa no use los datos biométricos obtenidos con otros fines. Sin embargo, en ese mismo informe la generación de los Millennials no demuestra reparo en utilizar los dispositivos a cambio de beneficios laborales y un mejor ambiente de trabajo [62].

Capítulo 11

Conclusiones y Líneas de trabajo futuras

11.1. Conclusiones

Tras el trabajo expuesto se puede concluir que se han cumplido todos los objetivos inicialmente planteados.

Se han adquirido nuevos datos, de 24 usuarios, analizado su señal y comprobado sus similitudes con los datos que ya se tenían disponibles, en una base de datos previa. Llegando a la conclusión de que son factibles de poder ser usados en biometría y construyendo un sistema de limpieza automática de la señal, capaz de corregir anomalías y eliminar el ruido.

Con el objetivo de buscar las mejores opciones en las que seguir trabajando, se ha hecho un análisis de estabilidad de la señal, alternando las muestras de los datos utilizadas para crear la plantilla de cada usuario (entrenamiento) y comprobando, a nivel de usuario, la componente de los datos en la que se consiguen los mejores resultados, teniendo disponibles los tres ejes tridimensionales (X, Y, Z) o su fusión, a través del módulo. Esto ha demostrado un comportamiento caótico indicativo de que la construcción de un sistema de predicción capaz de detectar el comportamiento del usuario no iba a generar resultados mejores a los ya disponibles. Por lo que se tomó la decisión de mejorar el sistema de reconocimiento que ya se tenía y que utilizaba el clasificador 1-vecino más próximo, elegido por su simplicidad, no habiendo probado hasta ese momento, ningún clasificador más.

La prueba de diferentes clasificadores ha llevado a encontrar uno que genera resultados prometedores. Son las máquinas de vectores soporte,

SVM. Una vez elegido el clasificador se han realizado distintas opciones de preprocesamiento que habían quedado en duda: el efecto de interpolar los datos para conseguir una frecuencia de muestreo fija, normalizar la amplitud para llevarlos a una escala común, muestrear los datos de entrenamiento para solucionar el problema del desbalanceo de las clases o barajear los datos para evitar sobreajuste. Los resultados obtenidos han sido claros destacando el kernel radial por ser bueno y estable y el sigmoide por ser muy bueno pero muy inestable, por lo que se ha decidido utilizar el primero interpolando los datos, normalizando su amplitud, sobremuestreandolos para eliminar el problema del desbalanceo de las clases y barajándolos para evitar el sobreajuste. Con todo ello, y utilizando la componente módulo en todos los usuarios, se han obtenido tasas de error inferiores al 1 % con la extracción de características tanto en el dominio del tiempo como en el dominio de la frecuencia.

Con resultados ya muy buenos, se ha probado a seguir mejorándolos, utilizando la técnica de fusionar scores. Esto ha llevado a conseguir tasas de error del 0 % en todos los usuarios, utilizando cualquier componente y manteniéndose los mismos resultados en ambas bases de datos disponibles y con los dos dispositivos, lo cual verifica la idoneidad del sistema construido.

Todos los resultados anteriores se han obtenido, suponiendo que el usuario utilizaba la mano en la que habitualmente llevaría situado su propio reloj, que es el escenario más cómodo y realista. Pero se ha analizado el efecto de utilizar la mano opuesta llegando a la conclusión, tras los buenos resultados obtenidos, de que es indiferente la mano en que se lleve ubicado el dispositivo porque el movimiento de las manos es simétrico y, por tanto, si en un momento concreto de la adquisición, el usuario se cambia el dispositivo de mano o si continuamente utiliza su mano no dominante, los resultados se mantienen.

Dados los buenos resultados obtenidos, se ha realizado un análisis de las arquitecturas escalables necesarias en estos datos y enfocadas a los escenarios futuros que se plantean. Pensando, también, en futuras implementaciones del sistema, se ha realizado una propuesta de monitorización y visualización de los datos que ayude a detectar problemas y tomar decisiones de la manera más rápida posible. Por último, también se han analizado las amenazas, vulnerabilidades, beneficios e inconvenientes existentes en estos datos biométricos, junto con una serie de controles mitigantes y buenas prácticas que el Incibe recomienda aplicar.

Se puede concluir que los resultados obtenidos son muy prometedores y muestran que el reconocimiento de la forma de andar a través de dispositivos ponibles puede ser una alternativa muy interesante a los sistemas biométricos actualmente disponibles. Además, se puede decir que se han cerrado cuestiones que habían quedado abiertas en los trabajos previos realizados, de los cuales éste es continuación, habiendo llegado a construir un sistema más eficiente y simple al propuesto en [57], por obtener la mejor tasa de error, 0 %, y utilizar un tamaño de ventana fijo, en vez de una fusión de tamaños entre 6 y 10, que multiplica por 5 el número de ejecuciones del clasificador.

En cuanto a las conclusiones personales, este trabajo me ha permitido poner en práctica los conocimientos adquiridos en diversas asignaturas de mis estudios, tanto de Grado como de Máster, aprender nuevos conceptos, el diseño experimental y la manera de trabajar en biometría. Pudiendo vivir la experiencia de un problema actual y novedoso que no tiene una solución única, sino una infinidad de posibilidades sobre las que he tenido que tomar muchas decisiones, así como plantear una arquitectura escalable, una monitorización apropiada y las consecuencias que el trabajo tendría en cuanto a ciberseguridad.

Por otro lado, ha sido una gran experiencia realizar mis estudios de manera no presencial y compaginándolo con el trabajo en empresa. Esto me ha ayudado a organizarme mejor, sacando cada semana las cosas que quedaban por hacer, priorizándolas y realizándolas de manera progresiva. Me ha permitido continuar trabajando en un grupo de investigación y entender mejor cómo funciona el proceso de la investigación científica desde dentro. Habiendo conseguido, con todo ello, una experiencia muy productiva y enriquecedora, tanto de manera personal como académica.

11.2. Líneas de trabajo futuro

En este proyecto se ha conseguido construir un sistema de reconocimiento biométrico eficiente y prometedor, así como reforzar unas bases sólidas sobre está biometría. Las tasas de error conseguidas son mínimas, habiendo hecho la prueba sobre un total de 32 usuarios. No obstante, para obtener resultados más significativos se necesitan datos tanto de más usuarios como una adquisición continúa de los ya existentes. Por lo que, como próximos pasos, se necesita profundizar en la parte de arquitectura para implementar un sistema eficiente de almacenamiento y sincronización de los datos que permita una adquisición no supervisada

en la que el usuario tenga los dispositivos durante varios días y de manera continua, capture sus datos, realizando una variedad de actividades, no solo caminar. Esto planteará otro reto: seguir reconociendo al usuario mientras realiza una mayor cantidad de actividades.

Una vez conseguido lo anterior, se podrá plantear un escenario real, en streaming, en el que se permita a cientos de usuarios generar datos, almacenarlos y explotarlos en tiempo real. Junto a una buena monitorización de dichos datos que ayude a tomar decisiones, detectar problemas y mejorar el sistema.

Pero sin duda, a la vista de los resultados obtenidos, el presente TFM planta una semilla muy interesante y positiva para seguir trabajando en esta prometedora biometría.

Acrónimos y abreviaturas

ACC Acelerómetro

AI Inteligencia Artificial

AI HLEG Grupo de Expertos de Alto Nivel sobre Inteligencia Artificial

ANN Artificial Neural Networks

API Application Programming Interface

APK Android Application Package

AUC Área bajo la curva

BD Base de datos

BI Business Intelligence

C Coste

CSV Comma-Separated Values

DCT Transformada Discreta del Coseno

DF Dominio de la Frecuencia

DT Dominio del Tiempo

DTree Decision Trees

DTW Dynamic Time Warping

EER tasa de equierror

EMG Electromyography

EU Unión Europea

FC Fusión de Características

Fm Frecuencia de muestreo

FNR tasa de falsos negativos

FPR tasa de falsos positivos

FS Fusión de Scores

FT transformada de Fourier

FTT Transformada Rápida de Fourier

GMM Gaussian Mixture Model

GPS Sistema de Posicionamiento Global

GYR Giroscopio

H precisión

HDFS Hadoop Distributed File System

HTTP Hypertext Transfer Protocol

Hz Hercios

HMM Hidden Markov Model

I identificación

ICA Independent Component Analysis

ID Identificador

IMEI International Mobile Equipment Identity

IMU Unidad de Medida Inercial

IOT Internet of Things

JDBC Java Database Connectivity

K Kernel

KNN K-Nearest Neighbors

LDA Linear Discriminant Analysis

LOPD Ley de Protección de Datos de Carácter Personal

Mi Muestra i

MICRO Microsoft

ML Machine Learning

MLP Perceptrón multicapa

MM3 Filtro de la Media Móvil de orden 3

MOTO Motorola

MonoMono Monosesión- Monomuestra

MOMO Monosesión-Monomuestra

MultiMono Multisesión- Monomuestra

MUMO Multisesión- Monomuestra

MultiMulti Multisesión- Multimuestra

NB Naive Bayes

NFC Near Field Communication

NoSQL Not only SQL

PCA Principal Component Analysis

PIN Personal Identification Number

PPG sensores fotopletismográficos

PWC PriceWaterhouseCoopers

RBF Función de Base Radial

RDLOPD Reglamento General de Protección de Datos de Carácter Personal

RF Random Forest

ROC Receiver Operating Characteristic

SQL Structured Query Language

SOAP Simple Object Access Protocol

SDK Software Development Kit

Si Sesión i

SI Sistema Internacional

SSL Secure Sockets Layer

SVM Support Vector Machines

T Periodo

TDF Transformada de Fourier Discreta

TFG Trabajo Fin de Grado

TFM Trabajo Fin de Máster

TLS Transport Layer Security

TNR Tasa de negativos verdaderos

TPR Tasa de verdaderos positivos

TSDB Timeseries Database

TV Tamaño variable

UI Interfaz de Usuario

UVa Universidad de Valladolid

V verificación

WMA Weighted Moving Average

Índice alfabético

Apéndices

Apéndice A

Manual de uso de la visualización de datos crudos

Para la visualización de los datos crudos se ha construido una aplicación web interactiva. Se trata de un fichero R Markdown cuyo pseudocódigo se puede ver en la figura A.1 y que tras ejecutarlo, abre una aplicación en el navegador, bajo la IP localhost (127.0.0.1), tal y como se puede ver en la figura A.2.

Este manual pretende explicar tanto los pasos a seguir para llegar a ejecutar la aplicación de visualización de los datos crudos adquiridos en el trabajo como las posibles operaciones que se pueden realizar con la misma.

A.1. Manual de ejecución

Es requisito imprescindible tener instalado el software R en el ordenador que se quiere ejecutar la aplicación. Se van a explicar los pasos a seguir con un Sistema Operativo Windows. No obstante, también funcionaría en Linux, ya que a alto nivel se necesita el software R y un navegador web.

En Windows (64 bits):

- Rgui (distribuciones binarias precompiladas): `https://cran.r-project.org/bin/windows/base/R-4.0.2-win.exe`
- RStudio (entorno de desarrollo integrado para el lenguaje de programación R): `https://download1.rstudio.org/desktop/windows/RStudio-1.3.959.exe`

Entrada:
Identificador del usuario, usuario (desplegable con opciones)
Sesión de los datos, sesion (desplegable con opciones)
Muestra de los datos, muestra (desplegable con opciones)
Dispositivo empleado, dispositivo (desplegable con opciones)
Sensor capturado, sensor (desplegable con opciones)
Coordenada(s) a mostrar, coordenada (checks buttons)

Salida:
Aplicación web interactiva que permite la visualización de los datos bajo cualquier casuística de entrada

Funcionamiento:
Fija como directorio de trabajo la ruta donde se encuentra el fichero RMarkdown que se está ejecutando
Toma como ruta de los datos de entrada (ficheros en formato CSV), una carpeta llamada originales, a partir de la ruta anterior
Se genera un data.frame con los datos de todos los usuarios comprobando la existencia de valores negativos y corrigiéndolos, situación que se comentará más adelante, tabla_completa
Convierte la tabla de los datos a tipo tibble, data.frames perezosos y hoscos. Porque hacen menos (por ejemplo, no cambian los nombres o los tipos de las variables y no hacen coincidencias parciales) y se quejan más (por ejemplo, cuando una variable no existe). Esto los hace más rápidos y los obliga a enfrentar los problemas antes, data_ponibles
Se define una función con la interfaz de usuario (UI) de la aplicación, ui
Se define la función principal, que creará el subconjunto de los datos de interés a mostrar, a partir de data_ponibles, en función de los desplegables y checks buttons seleccionados en la UI como entrada y generará el gráfico, server
Crea el objeto Shiny a partir de las funciones ui y server

Figura A.1: Funcionamiento de la aplicación web interactiva de visualización de los datos crudos

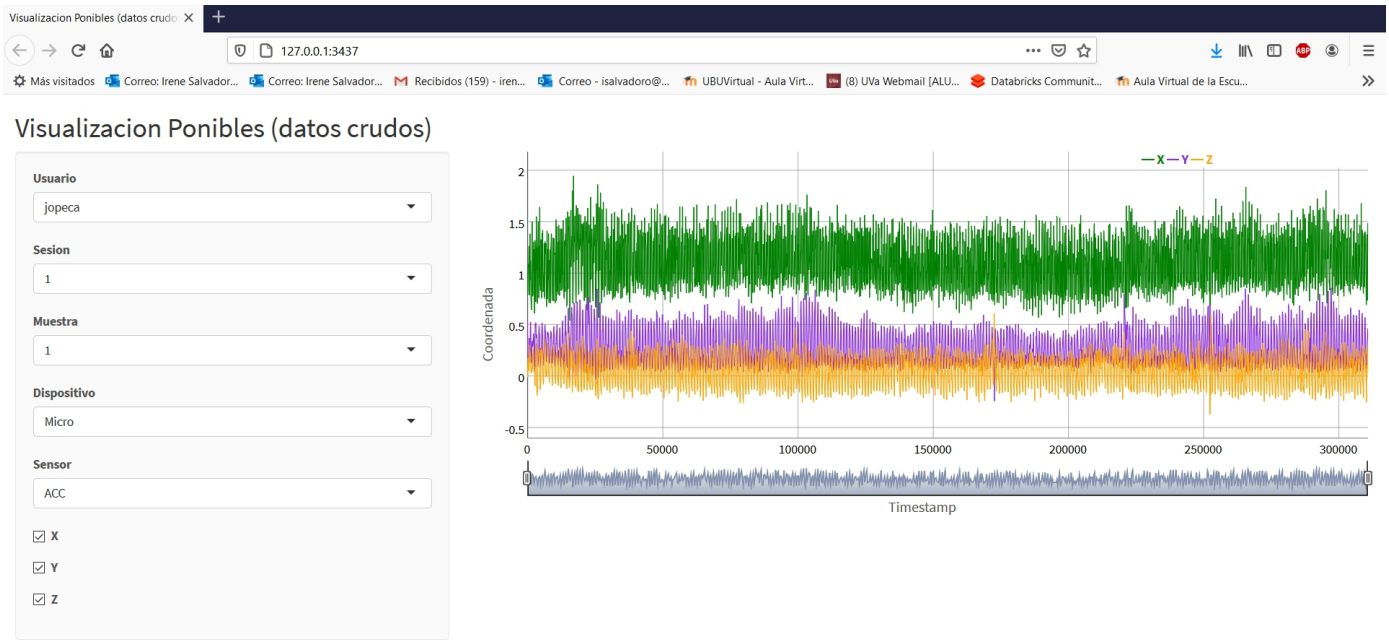


Figura A.2: Aplicación web interactiva de visualización de los datos crudos

Siendo necesario que la versión Java de nuestro ordenador y R tengan la misma arquitectura.

Como partida, se va a disponer de un fichero zip que va a contener un fichero RMarkdown (VisualizacionInicial.Rmd) y una carpeta con nombre "originales", la cual va a contener una nueva carpeta por cada usuario y dentro de cada una de ellas, todos los ficheros CSV de interés. Cada carpeta de usuario va a tener un total de 24 ficheros CSV, correspondientes a (2 sesiones) x (2 dispositivos) x (2 sensores) x (3 muestras).

Para conseguir la aplicación web interactiva, es suficiente con abrir el fichero con extensión Rmd, eligiendo RStudio como aplicación. Una vez abierto, es necesario indicar el directorio de trabajo donde se encuentra el fichero Rmd y la carpeta "originales" con los datos. Se hace a través de la variable path_file (línea 25 en la figura A.3) y seleccionando "Run/Restart R and Run All Chunks", tal y como se puede ver en la misma figura A.3.

Por defecto, siempre que se ejecuta la aplicación, se instalan todos los paquetes necesarios para su ejecución. Sin embargo, es suficiente con instalarlos la primera vez. No pasa nada por instalar varias veces un mismo paquete, pero invierte une pequeña cantidad de tiempo. Si ya están instalados los paquetes, se pueden dejar de instalar en siguientes

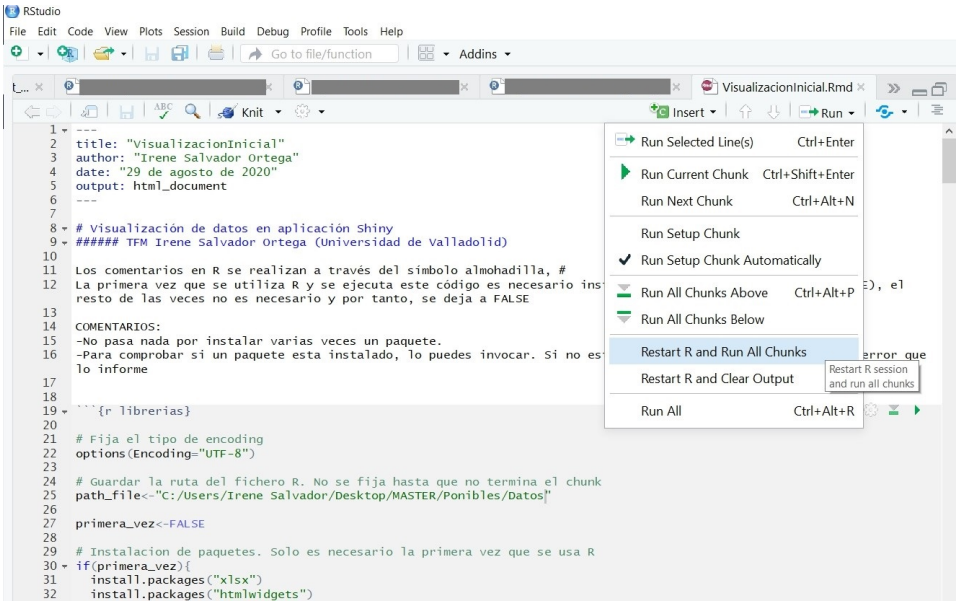


Figura A.3: Ejecutar en RStudio la aplicación de visualización de los datos crudos completa

ejecuciones desactivando el booleano de la línea 27 del código que se ve en la figura A.3 (primera_vez<-FALSE).

Si se cierra la aplicación Shiny, pero se ha quedado guardada la sesión de R, es decir, en la parte derecha de RStudio "Environment" se ven los objetos de datos cargados, no hace falta volver a invocar a la aplicación haciendo Run de todo el código, podemos ir al final del fichero Rmd, al último bloque de código y presionar "Run Current Chunk". Esto la abrirá de manera inmediata. Esta situación se puede ver en la figura A.4.

A.2. Manual de usuario

Inicialmente, cuando se abre la aplicación se tienen unas opciones por defecto. Como se puede ver en la figura A.5, aparece un usuario ya seleccionado, con sus datos en la sesión 1, muestra 1 con el dispositivo Micro y el sensor ACC. Aparecen las 3 coordenadas seleccionadas, no obstante se puede eliminar la selección de alguna de ellas, como se puede ver en la figura A.6 o de todas ellas, y el gráfico aparecerá vacío, figura A.7.

Se tienen disponibles:

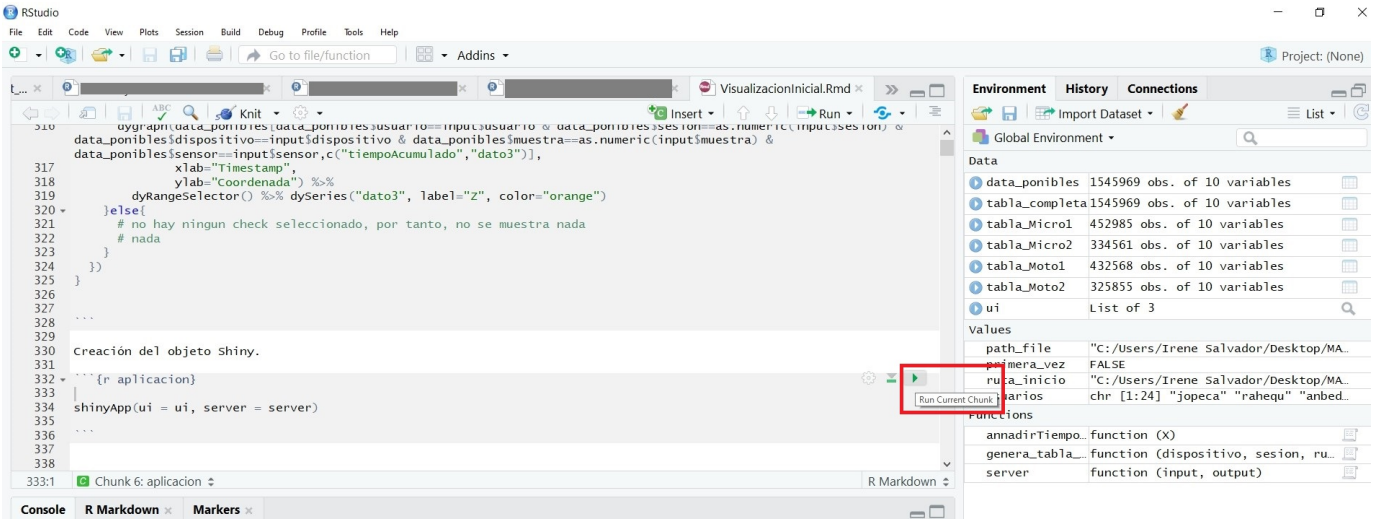


Figura A.4: Ejecutar en RStudio la aplicación de visualización de los datos crudos por 2ª vez

- Un total de 24 usuarios. Se puede seleccionar el que se desee a través de su identificador, en el desplegable “Usuario”.
- Dos sesiones en las que se han recogido datos, 1 o 2. Se puede cambiar de una a otra a través del desplegable “Sesion”.
- Tres muestras de datos en cada sesión, 1, 2 o 3, a seleccionar con el desplegable “Muestra”.
- Dos dispositivos de datos disponibles, la pulsera de Microsoft o el reloj de Motorola. Se puede elegir uno u otro, Micro o Moto, a través del desplegable “Dispositivo”.
- Dos sensores con los que se han adquirido datos, el acelerómetro o el giroscopio. Se puede elegir uno u otro, ACC o GYR, a través del desplegable “Sensor”.

Es de interés, en cada usuario, la visualización de 24 gráficos. Para cada sesión, habiendo un total de 2:

- En el dispositivo Micro y el sensor Acelerómetro, los gráficos de sus 3 muestras.
- En el dispositivo Micro y el sensor Giroscopio, los gráficos de sus 3 muestras.

- En el dispositivo Moto y el sensor Acelerómetro, los gráficos de sus 3 muestras.
- En el dispositivo Moto y el sensor Giroscopio, los gráficos de sus 3 muestras.

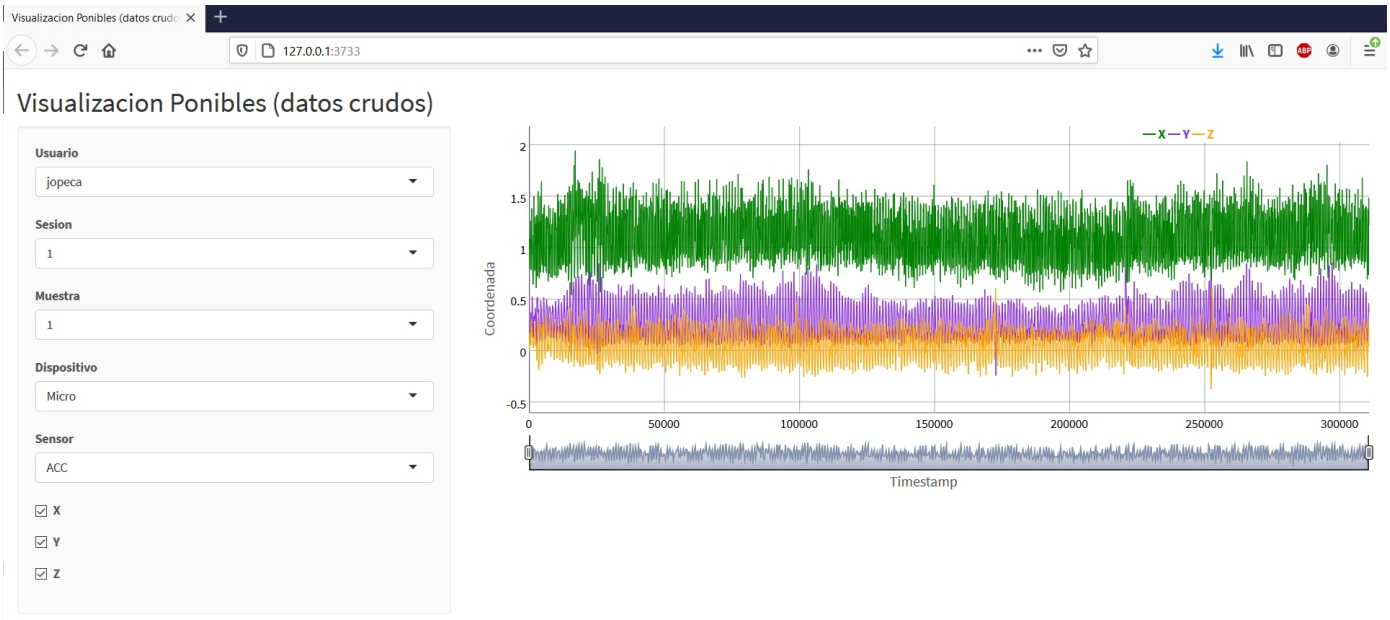


Figura A.5: Opciones por defecto de la aplicación de visualización de los datos crudos

En cada gráfico es posible seleccionar una zona de interés y verla en detalle, así como conocer las coordenadas exactas de los ejes X, Y, Z al pasar el cursor por encima de un punto. Se muestra un ejemplo de estas dos acciones en la figura A.8, el gráfico esta recortado entre 155000 y 177000 milisegundos y las coordenadas del punto seleccionado en torno a 158000 milisegundos, marcado con un círculo en el gráfico, se pueden ver en la parte superior derecha del mismo.

A.3. Ventajas e inconvenientes

Existen una serie de ventajas:

- Una vez se ha generado la aplicación, se puede abrir en el navegador y permite la visualización rápida y cómoda de los datos de cualquier usuario, con la posibilidad de cambiar de uno a otro.

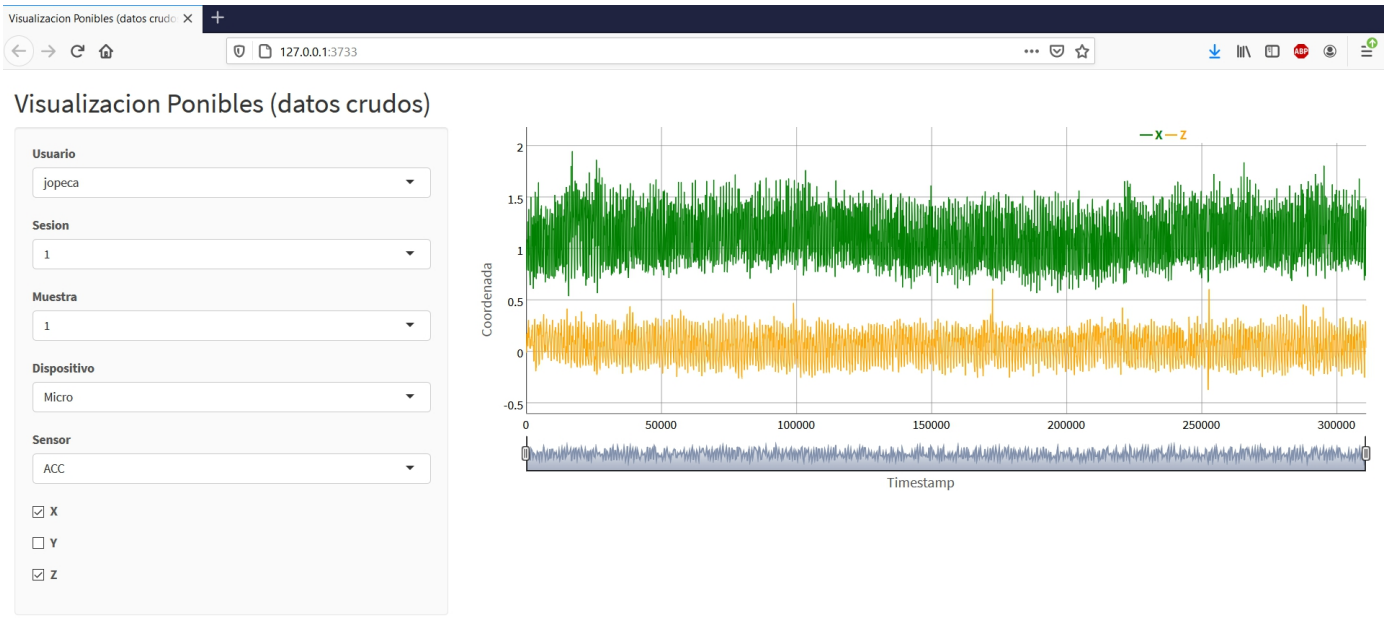


Figura A.6: Selección de coordenadas en la aplicación de visualización de los datos crudos

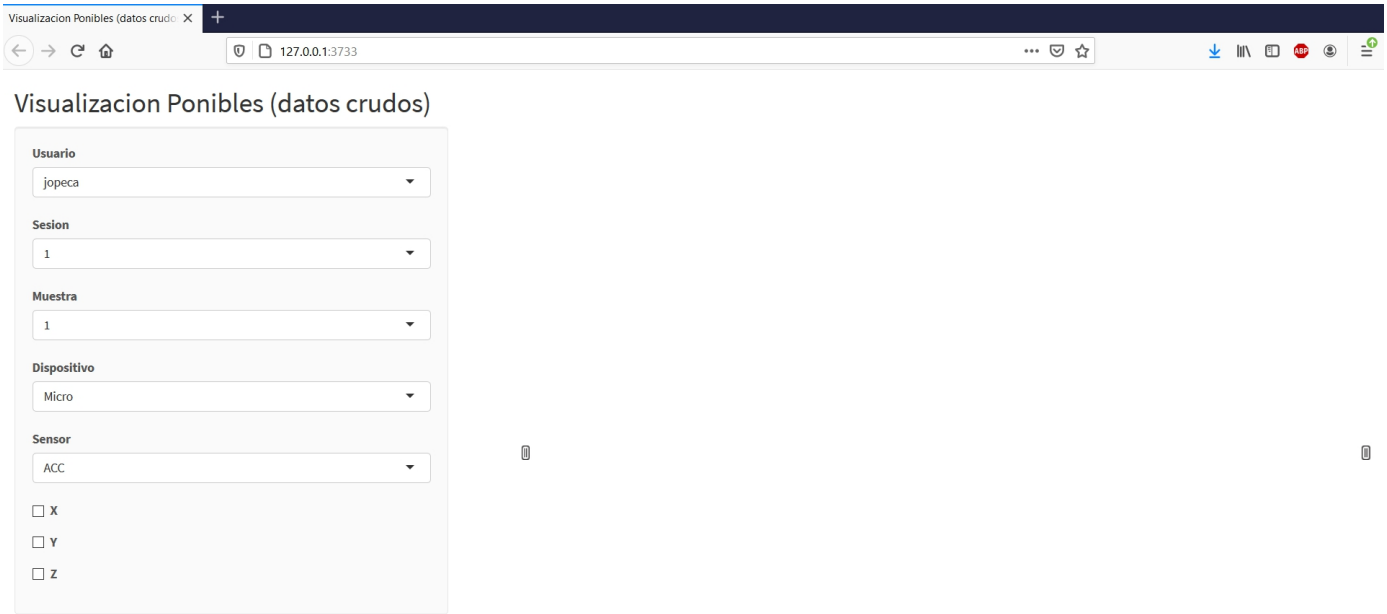


Figura A.7: Eliminar las coordenadas en la aplicación de visualización de los datos crudos

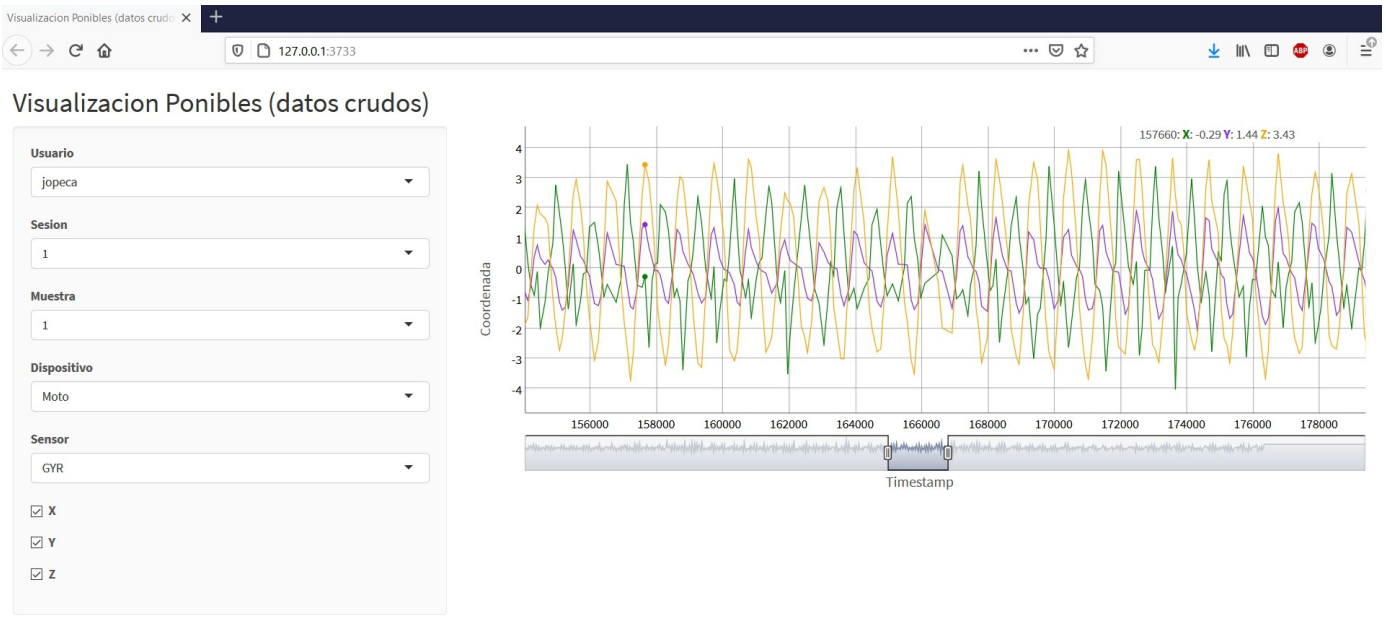


Figura A.8: Posibles acciones sobre el gráfico de la visualización de los datos crudos

- Dentro del gráfico se pueden ir viendo los valores exactos de los ejes X/Y/Z o recortar una zona de interés concreta más pequeña para verla en detalle. Este recorte se eliminará al cambiar las opciones de los desplegables o recargar la página del navegador.
- Adaptable. Es posible aumentar la cantidad de usuarios o datos, respetando la estructura de carpetas y con pequeñas modificaciones de código.

Como inconvenientes:

- Depende del software R, el cual hay que tener instalado en el ordenador que se quiere ejecutar. Para evitar la falta de portabilidad, sería posible compartir la aplicación como una página web, por ejemplo, a través de *shinyapps.io* [15], un servicio de alojamiento de RStudio para este tipo de aplicaciones. Es fácil de usar, seguro y escalable, pero de manera gratuita, sólo se pueden subir 5 aplicaciones, las cuales pueden estar un máximo de 25 horas activas.
- Depende de ficheros CSV locales con los datos de cada usuario. Una alternativa sería obtenerlos directamente de la base de datos *phpmyadmin*. No se ha hecho porque no es el objetivo de este trabajo.

Bibliografía

[1] ¿qué es business intelligence? https://www.sinnexus.com/business_intelligence/. [Internet; acceso 17-noviembre-2020].

[2] Heikki Ailisto. Identifying people from gait pattern with accelerometers. 2005.

[3] Petar Georgiev Aleksandrov. Una propuesta basada en aprendizaje automático para la mejora de la predicción en tiempos de llegada. http://uvadoc.uva.es/handle/10324/41514, 2020. [Internet; acceso 2-noviembre-2020].

[4] Daniel González Alonso. *Estudio preliminar del uso de Wearables en reconocimiento biométrico de personas*. TFG de la Escuela de Ingeniería Informática de la Universidad de Valladolid, Curso 2016/2017.

[5] Johanna Orellana Alvear. Arboles de decision y random forest. https://bookdown.org/content/2031/ensambladores-random-forest-parte-i.html#random-forest, 16 de noviembre de 2018. [Internet; acceso 17-noviembre-2020].

[6] página oficial Arcadia Data. Why use streaming visualizations? https://www.arcadiadata.com/product/streaming-visualizations/. [Internet; acceso 17-noviembre-2020].

[7] Kaan Ozyazici Gulsen Ayluctarhan Onur Agbulut Ibrahim Zincir Aybuke Kececi, Armagan Yildirak. Implementation of machine learning algorithms for gait recognition. *Department of Computer Engineering, Yasar University, Agacli Yol, No. 35-37, Bornova, Izmir 35100, Turkey*, 2020.

[8] Ahmed Banafa. Diez tendencias del internet de las cosas en 2020. https://www.bbvaopenmind.com/tecnologia/mundo-digital/

diez-tendencias-del-internet-de-las-cosas-en-2020/, 16 de diciembre de 2019. [Internet; acceso 17-noviembre-2020].

[9] Oresti Banos. Window size impact in human activity recognition. 2014.

[10] Akram Bayat. Classifying human walking patterns using accelerometer data from smartphone. 2017.

[11] Joe Belfiore. Making windows 10 more personal and more secure with windows helloy. 2015.

[12] Thomas Bernecker. Activity recognition on 3d accelerometer data (technical report). s.f.

[13] Jorge Blasco. A survey of wearable biometric recognition systems. septiembre de 2016.

[14] Cassio Brodbeck. 26 jul proxy https: Entienda cómo funciona. https://ostec.blog/es/seguridad-perimetral/proxy-https-como-funciona, 26 de julio de 2020. [Internet; acceso 17-noviembre-2020].

[15] Shinyapps.io by RStudio. *servicio de alojamiento de RStudio para aplicaciones Shiny*. último acceso el 30 de septiembre de 2020.

[16] Luis Javier Mena Camaré. Aprendizaje automático a partir de conjuntos de datos no balanceados y su aplicación en el diagnóstico y pronóstico médico. https://inaoe.repositorioinstitucional.mx/jspui/bitstream/1009/533/1/MenaCaLJ.pdf, Septiembre 2008. [Internet; acceso 17-noviembre-2020].

[17] Pierluigi Casale. Personalization and user verification in wearable systems using biometric walking patterns. *Personal and Ubiquitous Computing*, 2012.

[18] Alexander J. Casson. Gyroscope vs. accelerometer measurements of motion from wrist ppg during physical exercise. 2016.

[19] Guglielmo Cola, Marco Avvenuti, Fabio Musso, and Alessio Vecchio. Gait-based authentication using a wrist-worn device. In *Proceedings of the 13th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, MOBIQUITOUS 2016, page 208–217, New York, NY, USA, 2016. Association for Computing Machinery.

[20] Álvaro Giz Bueno César Tolosa Borja. Sistemas biométricos. https://www.dsi.uclm.es/personal/MiguelFGraciani/mikicurri/Docencia/Bioinformatica/web_BIO/Documentacion/Trabajos/Biometria/Trabajo%20Biometria.pdf. [Internet; acceso 22-noviembre-2020].

[21] Ghina Dandachi. A novel identification/verification model using smartphone's sensors and user behavior. *2nd International Conference on Advances in Biomedical Engineering*, 2013.

[22] DB-Engines. Ranking. https://db-engines.com/en/ranking_categories. [Internet; acceso 2-noviembre-2020].

[23] Mohammad Derawi. Wireless chest-based ecg biometrics. *Springer*, 2015.

[24] Mohammad Omar Derawi. Accelerometer-based gait analysis, a survey. 2010.

[25] Weka Documentation. Class jrip. https://weka.sourceforge.io/doc.dev/weka/classifiers/rules/JRip.html. [Internet; acceso 17-noviembre-2020].

[26] George Doddington. Sheep, goats, lambs and wolves. a statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation. 1998.

[27] Miikka Ermes. Detection of daily activities and sports with wearable sensors in controlled and uncontrolled conditions. 2006.

[28] Rosa Fernández. Evolución de los envíos de dispositivos wearables a nivel mundial de 2014 a 2024. https://es.statista.com/estadisticas/702760/envios-mundiales-de-wearables-en-unidades/, 14 de septiembre de 2020. [Internet; acceso 17-noviembre-2020].

[29] Davrondzhon Gafurov. Biometric gait authentication using accelerometer sensor. 2006.

[30] Davrondzhon Gafurov and Einar Snekkenes. Gait recognition using wearable motion recording sensors. 2009.

[31] J. M. Galán. *Wearables: Análisis de dispositivos y recogida de datos en Android para estudios biométricos*. TFG de la Escuela de Ingeniería Informática de la Universidad de Valladolid, Curso 2015/2016.

[32] Zoubin Ghahramani. An introduction to hidden markov models and bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence*, 2001.

[33] Martin Reese Hestbek. Biometric gait recognition for mobile devices using wavelet transform and support vector machines. 2012.

[34] Chiung Ching Ho. An unobtrusive android person verification using accelerometer based gait i. 2010.

[35] Chiung Ching Ho. An unobtrusive android person verification using accelerometer based gait ii. *10th International Conference on Advances in Mobile Computing & Multimedia*, 2012.

[36] Instituto Nacional de Ciberseguridad Incibe. Tecnologías biométricas aplicadas a la ciberseguridad. `https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjhn5O6t-_sAhUJ5uAKHVMWCsEQFjACegQIBhAC&url=https%3A%2F%2Fwww.incibe.es%2Fprotege-tu-empresa%2Fguias%2Ftecnologias-biometricas-aplicadas-ciberseguridad-guia-aproximacion-e usg=AOvVaw2tpC1WG34ylfIstlA_Jlai`. [Internet; acceso 22-noviembre-2020].

[37] Lucas Introna and Helen Nissenbaum. Facial recognition technology a survey of policy and implementation issues. technical report. *The Department of Organisation, Work and Technology, Lancaster University.*, 2010.

[38] Anil K. Jain. Multibiometric systems. communications. *ACM*, 2004.

[39] A. H. Johnston and G. M. Weiss. Smartwatch-based biometric gait recognition. In *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–6, Sep. 2015.

[40] Samer K Al Kork. Biometric database for human gait recognition using wearable sensors and a smartphone. 2017.

[41] Hindra Kurniawan. Stress detection from speech and galvanic skin response signals. *IEEE 26th International Symposium on Computer-Based Medical Systems*, 2013.

[42] Unir (la Universidad en Internet). Apache spark en big data: qué es y para que se emplea. `https://www.unir.net/ingenieria/revista/apache-spark-big-data/`, 7 de mayo de 2020. [Internet; acceso 17-noviembre-2020].

[43] Elena Campo León. Introducción a las máquinas de vector soporte (svm) en aprendizaje supervisado. https://zaguan.unizar.es/record/59156/files/TAZ-TFG-2016-2057.pdf?version=1. [Internet; acceso 8-noviembre-2020].

[44] Hong Lu. Unobtrusive gait verification for mobile phones. *ACM*, 2014.

[45] Hong Lu, Jonathan Huang, Tanwistha Saha, and Lama Nachman. Unobtrusive gait verification for mobile phones. In *Proceedings of the 2014 ACM International Symposium on Wearable Computers*, ISWC '14, page 91–98, New York, NY, USA, 2014. Association for Computing Machinery.

[46] Davide Maltoni. Handbook of fingerprint recognition. *Springer*, 2009.

[47] Data Center Market. Cinco lenguajes de programación que deberían incorporar todas las empresas. https://www.datacentermarket.es/mercado/noticias/1120802032609/cinco-lenguajes-de-programacion-deberian-incorporar-todas-empresas.1.html, 21 de septiembre de 2020. [Internet; acceso 17-noviembre-2020].

[48] Maria De Marsico and Alessio Mecca. A survey on gait recognition via wearable sensors. *ACM Comput. Surv.*, 52(4), August 2019.

[49] Pallavi Meharia and Dharma P. Agrawal. Unobtrusive gait verification for mobile phones. *Journal of Information Privacy & Security*, 2015.

[50] página oficial Microsoft Power BI. Transmisión en tiempo real en power bi. https://docs.microsoft.com/es-es/power-bi/connect-data/service-real-time-streaming, 16 de julio de 2020. [Internet; acceso 17-noviembre-2020].

[51] Francisco Dionicio Pichardo Morales. Apuntes de desbalance de clases en clasificación inteligente. https://franciscopichardoblog.files.wordpress.com/2017/06/apuntes-de-desbalance1.pdf. [Internet; acceso 17-noviembre-2020].

[52] João Neto. Fourier transform: A r tutorial. http://www.di.fc.ul.pt/~jpn/r/fourier/fourier.html, marzo de 2013.

[53] Claudia Nickel. Using hidden markov models for accelerometer-based biometric gait recognition. *IEEE 7th International Colloquium on Signal Processing and Its Applications*, 2011.

[54] Claudia Nickel. Authentication of smartphone users based on the way they walk using k-nn algorithm. *8th International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 2012.

[55] Michael Fitzgerald Nowlan. Human identification via gait recognition using accelerometer gyro forces. 2009.

[56] European Commission's High-Level Expert Group on Artificial Intelligence. Draft ethics guidelines for trustworthy ai. `https://ec.europa.eu/digital-single-market/en/news/draft-ethics-guidelines-trustworthy-ai`, 18 de diciembre de 2018. [Internet; acceso 22-noviembre-2020].

[57] Irene Salvador Ortega. *Reconocimiento Biométrico mediante Dispositivos Ponibles (wearables)*. TFG de la Escuela de Ingeniería Informática y Facultad de Ciencias de la Universidad de Valladolid, Curso 2018/2019.

[58] Sergio J. Ortiz. ¿representarán realmente los smartwatches la evolución tecnológica del futuro? `https://www.ipadizate.es/2020/02/10/smartwatch-evolucion-futuro/`, 10 de febrero de 2020. [Internet; acceso 17-noviembre-2020].

[59] Pinterest. 50 lecciones de vida para vivir al máximo cada día. `https://www.pinterest.es/pin/418834834098510966/`. [Internet; acceso 17-noviembre-2020].

[60] Pixabay. Cómo realizar una clasificación de aprendizaje automático explicable sin ningún árbol. `https://sitiobigdata.com/2019/12/24/como-realizar-una-clasificacion-de-aprendizaje-automatico-explicable-s` 2019. [Internet; acceso 17-noviembre-2020].

[61] Salil Prabhakar. Biometric recognition: Security and privacy concerns. *IEEE Security & Privacyg*, 2003.

[62] página oficial PwC. Use of wearables in the workplace is halted by lack of trust – pwc research. `https://pwc.blogs.com/press_room/2016/06/use-of-wearables-in-the-workplace-is-halted-by-lack-of-trust-pwc-resea` `html`, 20 de junio de 2016. [Internet; acceso 22-noviembre-2020].

[63] página oficial Qlik. What is data streaming. https://www.qlik.com/us/data-streaming/what-is-data-streaming. [Internet; acceso 17-noviembre-2020].

[64] J. Ross Quinlan. C4.5: Programs for machine learning. *Elsevier*, 2014.

[65] RashaWahid. A gaussian mixture models approach to human heart signal verification using different feature extraction algorithms. *Computer Applications for Bio-technology, Multimedia, and Ubiquitous City*, 2012.

[66] página oficial RDocumentation. Support vector machines. https://www.rdocumentation.org/packages/e1071/versions/1.7-3/topics/svm. [Internet; acceso 8-noviembre-2020].

[67] Kenneth Revett. Behavioral biometrics: A remote access approach. 2008.

[68] Joaquín Amat Rodrigo. Máquinas de vector soporte (support vector machines, svms). https://rpubs.com/Joaquin_AR/267926, abril de 2017. [Internet; acceso 8-noviembre-2020].

[69] Joaquín Amat Rodrigo. Árboles de predicción: random forest, gradient boosting y c5.0. https://rpubs.com/Joaquin_AR/255596, febrero de 2017. [Internet; acceso 17-noviembre-2020].

[70] Liu Rong. A wearable acceleration sensor system for gait recognition. 2017.

[71] Sherif Said. Experimental investigation of human gait recognition database using wearable sensors. 2018.

[72] Sherif Said, Samer Al-kork, Vishnu Nair, Itta Gowthami, Taha Beyrouthy, Xavier Savatier, and M Fayek Abdrabbo. Experimental Investigation of Human Gait Recognition Database using Wearable Sensors. *Advances in Science, Technology and Engineering Systems Journal*, 3(4):201–210, 2018.

[73] R software. Package "rose". https://cran.r-project.org/web/packages/ROSE/ROSE.pdf, 19 de febrero de 2015. [Internet; acceso 17-noviembre-2020].

[74] Daisuke Sugimori. A study about identification of pedestrian by using 3-axis accelerometer. 2011.

[75] Bing Sun. Gait characteristic analysis and identification based on the iphone's accelerometer and gyrometer. 2014.

[76] F. Sun, C. Mao, X. Fan, and Y. Li. Accelerometer-based speed-adaptive gait authentication method for wearable iot devices. *IEEE Internet of Things Journal*, 6(1):820–830, Feb 2019.

[77] Toshiyo Tamura. Wearable photoplethysmographic sensors—past and present. *Electronics*, 2014.

[78] Hoang Minh Thang. Gait identification using accelerometer on mobile phone. 2012.

[79] Brynjulv Tveit. Analyzing behavioral biometrics of handwriting using myo gesture control armband, 2018.

[80] Adam S. Venable. Gender differences in skin and core body temperature during exercise in a hot, humid environment. *Internal Journal of Exercise Science: Conference Proceedings, Vol. 2. 9.*, 2013.

[81] S. Vhaduri and C. Poellabauer. Multi-modal biometric-based implicit authentication of wearable device users. *IEEE Transactions on Information Forensics and Security*, 14(12):3116–3125, Dec 2019.

[82] Changsheng Wan, Li Wang, and Vir V. Phoha. A survey on gait recognition. *ACM Comput. Surv.*, 51(5), August 2018.

[83] Wikipedia. Biometría — wikipedia, la enciclopedia libre. `https://es.wikipedia.org/wiki/Biometr%C3%ADa`, 2020. [Internet; descargado 17-noviembre-2020].

[84] Wikipedia. Curtosis. `https://es.wikipedia.org/wiki/Curtosis`, agosto de 2020. [Internet; acceso 17-noviembre-2020].

[85] Wikipedia. Asimetría estadística. `https://es.wikipedia.org/wiki/Asimetr%C3%ADa_estad%C3%ADstica`, mayo de 2020. [Internet; acceso 17-noviembre-2020].

[86] Wikipedia. Hadoop distributed file system. `https://es.wikipedia.org/wiki/Hadoop_Distributed_File_System`, noviembre de 2020. [Internet; acceso 17-noviembre-2020].

[87] Guannan Wu. A continuous identity authentication scheme based on physiological and behavioral characteristics. 2018.

[88] Guannan Wu, Jian Wang, Yongrong Zhang, and Shuai Jiang. Authentication scheme based on physiological and behavioral characteristics. *Sensors (Basel)*, 18(1), January 2018.

[89] Xello. Base de datos relacional, escuela de ingeniería informática de valladolid. `http://greidi.infor.uva.es/phpmyadmin/`. [Internet; acceso 17-noviembre-2020].

[90] Xello. Qlikview vs power bi: What's the difference? `https://xo.xello.com.au/blog/qlikview-vs-power-bi-whats-the-difference`, 11 de junio de 2020. [Internet; acceso 17-noviembre-2020].

[91] W. Xu, Y. Shen, Y. Zhang, N. Bergmann, and W. Hu. Gait-watch: A context-aware authentication system for smart watch based on gait recognition. In *2017 IEEE/ACM Second International Conference on Internet-of-Things Design and Implementation (IoTDI)*, pages 59–70, April 2017.

[92] Weitao Xu. Gait-watch: A context-aware authentication system for smart watch based on gait recognition. 2017.

[93] Roman V. Yampolskiy. Taxonomy of behavioral biometrics. *Behavioral Biometrics for Human Identification*, 2010.

[94] Fan Yang. Real-time human activity classification by accelerometer embedded wearable devices. 2017.

[95] Liu Yiyan. An hidden markov model based complex walking pattern recognition algorithm. 2016.

[96] Zhidong Zhao and Qinqin Shen. A human identification system based on heart sounds and gaussian mixture models. *4th International Conference on Biomedical Engineering and Informatics*, 2011.